BEATRIZ TOMAZELA TEODORO

Sistema de Reconhecimento Automático de Língua Brasileira de Sinais

São Paulo

2015

BEATRIZ TOMAZELA TEODORO

Sistema de Reconhecimento Automático de Língua Brasileira de Sinais

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de Concentração: Sistemas de Informação.

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 23 de outubro de 2015. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2015

CATALOGAÇÃO-NA-PUBLICAÇÃO
(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Teodoro, Beatriz Tomazela
Sistema de reconhecimento automático de Língua Brasileira de Sinais / Beatriz Tomazela Teodoro ; orientador, Luciano Antonio Digiampietri. – São Paulo, 2015
114 p. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo
Versão corrigida

1. Processamento de imagens. 2. Língua Brasileira de Sinais. 3. Sistemas de Informação. I. Digiampietri, Luciano Antonio, orient. II. Título

CDD 22.ed.– 621.367

Dissertação de autoria de Beatriz Tomazela Teodoro, sob o título *"Sistema de Reconhecimento Automático de Língua Brasileira de Sinais"*, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Sistemas de Informação, aprovada em 23 de outubro de 2015 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri
Presidente
Universidade de São Paulo

Prof. Dr. Moacir Antonelli Ponti
Universidade de São Paulo

Profa. Dra. Fátima de Lourdes dos Santos Nunes Marques
Universidade de São Paulo

Dedico este passo importante da minha vida a todos aqueles que acreditaram no meu potencial e estiveram sempre ao meu lado, me dando força para superar os obstáculos, erguer a cabeça e seguir em frente. Também dedico este trabalho a toda a comunidade surda, que enfrenta um grande obstáculo em seu dia a dia, a dificuldade em se comunicar com a população ouvinte.

Agradecimentos

Agradeço primeiramente a minha família e aos meus amigos, que mesmo sem entender direito o assunto deste trabalho, estiveram sempre ao meu lado durante todo o seu decorrer, por todo o apoio, paciência, conselhos, broncas e compreensão.

Agradeço ao meu orientador, Prof. Dr. Luciano Antonio Digiampietri, que além de ter me guiado neste trabalho, me orientou também durante a graduação, por estar sempre presente, prestando toda a orientação e esclarecimentos necessários, além de toda paciência e clareza ao transmitir seus ensinamentos, me fazendo aprender e muito.

Agradeço à professora de LIBRAS Maria Carolina Casati Digiampietri que auxiliou na construção do banco de imagens, bem como na validação do sistema.

Agradeço também aos demais professores da graduação e da pós-graduação da EACH-USP, pela grande contribuição na minha formação, em especial àqueles que valorizaram e incentivaram o meu trabalho.

Agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro.

Por fim, agradeço a todos os voluntários que participaram deste trabalho, sem os quais este trabalho não seria possível.

"A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê."

(Arthur Schopenhauer)

Resumo

TEODORO, Beatriz Tomazela. **Sistema de Reconhecimento Automático de Língua Brasileira de Sinais**. 2015. 114 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015.

O reconhecimento de língua de sinais é uma importante área de pesquisa que tem como objetivo atenuar os obstáculos impostos no dia a dia das pessoas surdas e/ou com deficiência auditiva e aumentar a integração destas pessoas na sociedade majoritariamente ouvinte em que vivemos. Baseado nisso, esta dissertação de mestrado propõe o desenvolvimento de um sistema de informação para o reconhecimento automático de Língua Brasileira de Sinais (LIBRAS), que tem como objetivo simplificar a comunicação entre surdos conversando em LIBRAS e ouvintes que não conheçam esta língua de sinais. O reconhecimento é realizado por meio do processamento de sequências de imagens digitais (vídeos) de pessoas se comunicando em LIBRAS, sem o uso de luvas coloridas e/ou luvas de dados e sensores ou a exigência de gravações de alta qualidade em laboratórios com ambientes controlados, focando em sinais que utilizam apenas as mãos. Dada a grande dificuldade de criação de um sistema com este propósito, foi utilizada uma abordagem para o seu desenvolvimento por meio da divisão em etapas. Considera-se que todas as etapas do sistema proposto são contribuições para trabalhos futuros da área de reconhecimento de sinais, além de poderem contribuir para outros tipos de trabalhos que envolvam processamento de imagens, segmentação de pele humana, rastreamento de objetos, entre outros. Para atingir o objetivo proposto foram desenvolvidas uma ferramenta para segmentar sequências de imagens relacionadas à LIBRAS e uma ferramenta para identificar sinais dinâmicos nas sequências de imagens relacionadas à LIBRAS e traduzi-los para o português. Além disso, também foi construído um banco de imagens de 30 palavras básicas escolhidas por uma especialista em LIBRAS, sem a utilização de luvas coloridas, laboratórios com ambientes controlados e/ou imposição de exigências na vestimenta dos indivíduos que executaram os sinais. O segmentador implementado e utilizado neste trabalho atingiu uma taxa média de acurácia de 99,02% e um índice *overlap* de 0,61, a partir de um conjunto de 180 *frames* pré-processados extraídos de 18 vídeos gravados para a construção do banco de imagens. O algoritmo foi capaz de segmentar pouco mais de 70% das amostras. Quanto à acurácia para o reconhecimento das palavras, o sistema proposto atingiu 100% de acerto para reconhecer as 422 amostras de palavras do banco de imagens construído, as quais foram segmentadas a partir da combinação da técnica de distância de edição e um esquema de votação com um classificador binário para realizar o reconhecimento, atingindo assim, o objetivo proposto neste trabalho com êxito.

Palavras-chaves: Reconhecimento de língua de sinais. Processamento de imagens. Segmentação de pele humana. Língua Brasileira de Sinais. LIBRAS.

Abstract

TEODORO, Beatriz Tomazela. **Automatic Recognition System of Brazilian Sign Language**. 2015. 114 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2015.

The recognition of sign language is an important research area that aims to mitigate the obstacles in the daily lives of people who are deaf and/or hard of hearing and increase their integration in the majority hearing society in which we live. Based on this, this dissertation proposes the development of an information system for automatic recognition of Brazilian Sign Language (BSL), which aims to simplify the communication between deaf talking in BSL and listeners who do not know this sign language. The recognition is accomplished through the processing of digital image sequences (videos) of people communicating in BSL without the use of colored gloves and/or data gloves and sensors or the requirement of high quality recordings in laboratories with controlled environments focusing on signals using only the hands. Given the great difficulty of setting up a system for this purpose, an approach divided in several stages was used. It considers that all stages of the proposed system are contributions for future works of sign recognition area, and can contribute to other types of works involving image processing, human skin segmentation, object tracking, among others. To achieve this purpose we developed a tool to segment sequences of images related to BSL and a tool for identifying dynamic signals in the sequences of images related to the BSL and translate them into portuguese. Moreover, it was also built an image bank of 30 basic words chosen by a BSL expert without the use of colored gloves, laboratory-controlled environments and/or making of the dress of individuals who performed the signs. The segmentation algorithm implemented and used in this study had a average accuracy rate of 99.02% and an overlap of 0.61, from a set of 180 preprocessed frames extracted from 18 videos recorded for the construction of database. The segmentation algorithm was able to target more than 70% of the samples. Regarding the accuracy for recognizing words, the proposed system reached 100% accuracy to recognize the 422 samples from the database constructed (the ones that were segmented), using a combination of the edit distance technique and a voting scheme with a binary classifier to carry out the recognition, thus reaching the purpose proposed in this work successfully.

Keywords: Sign language recognition. Image processing. Human skin segmentation. Brazilian Sign Language. BSL.

Lista de figuras

Lista de tabelas

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

A comunidade surda enfrenta um grande obstáculo em seu dia a dia, que é a dificuldade de se comunicar com a sociedade predominantemente ouvinte. Tarefas comuns e simples para a maioria da população, como fazer compras, realizar uma operação bancária e ir a uma consulta médica, podem ser um grande desafio para os deficientes auditivos.

Segundo o censo realizado em 2010 pelo Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2010), cerca de 9,7 milhões de brasileiros possuem deficiência auditiva, o que representa 5,1% da população brasileira. Deste total, a deficiência auditiva severa foi declarada por cerca de 2 milhões de pessoas (344,2 mil são surdas e 1,7 milhões têm grande dificuldade para ouvir). Para a maioria destas pessoas, a língua principal utilizada para comunicação é a Língua Brasileira de Sinais (LIBRAS) e não a língua portuguesa.

A LIBRAS, assim como as demais línguas de sinais, não corresponde a simples transcrições das línguas faladas, ou então, mímicas e gestos soltos utilizados por surdos e deficientes auditivos para se comunicarem, mas sim, uma língua com estruturas gramaticais próprias. Desta forma, compreende-se a dificuldade de comunicação existente entre os deficientes auditivos e a sociedade ouvinte.

Nos últimos anos o empenho em facilitar a comunicação entre surdos e/ou deficientes auditivos com pessoas que não conhecem uma língua gestual tem aumentado, mas ainda há poucos ambientes acessíveis para eles. A inclusão dos deficientes auditivos na sociedade tem enfrentado a falta de conhecimento dos ouvintes sobre línguas de sinais. Esta falta de conhecimento torna extremamente difícil a comunicação entre surdos e ouvintes. Além disso, o reconhecimento e a tradução de línguas de sinais por computadores são áreas bastante complexas, com estudos ainda recentes e abertos à pesquisa (NETO; OQUENDO, 2013).

Nas línguas de sinais, tanto na brasileira como em outras (ALBRES, 2010; QUADROS; KARNOPP, 2004; WILCOX; WILCOX, 2005), existem sinais estáticos para indicar algumas palavras e letras, porém muitos dos sinais são dinâmicos (envolvem não só as configurações das mãos, mas também os movimentos). Desta forma, para o processamento automático de uma conversa em língua de sinais é necessário o processamento de um vídeo (conjunto sequencial de imagens) e não somente das imagens individualmente.

O processamento de um vídeo pode ser visto como o processamento de uma sequência de imagens em que se pode tirar proveito da ordem na qual essas imagens se encontram, com a finalidade de descrever ou reconhecer certas características do vídeo.

O foco desse trabalho é o reconhecimento de LIBRAS, com a finalidade de simplificar a comunicação entre deficientes auditivos conversando em LIBRAS e ouvintes que não conheçam esta língua. O reconhecimento é realizado por meio do processamento de vídeos digitais de pessoas expressando palavras em LIBRAS e a partir da combinação de um conjunto de técnicas de processamento de imagens com a técnica de distância de edição e um esquema de votação com um classificador binário. Os vídeos utilizados foram gravados sem a utilização de luvas de dados e/ou coloridas, laboratórios com ambientes controlados e/ou imposição de exigências na vestimenta dos indivíduos que executam os sinais.

A principal motivação deste trabalho é a ideia de facilitar a comunicação utilizando uma abordagem pouco intrusiva. A partir da revisão sistemática realizada e apresentada no capítulo 3, foi possível observar que poucos trabalhos reconhecem sinais dinâmicos em vídeos gravados sem a utilização de luvas e/ou ambientes controlados. A maioria dos trabalhos trata apenas o reconhecimento de sinais estáticos. Além disso, foi identificado um baixo número de trabalhos que tratam de reconhecimento de sinais em LIBRAS, em relação às outras línguas de sinais.

Esta dissertação de mestrado estendeu trabalhos prévios (DIGIAMPIETRI et al., 2012; TEODORO, 2012) sobre reconhecimento de LIBRAS, por meio da especificação, implementação e testes de uma nova ferramenta, que tem como objetivo reconhecer e traduzir sinais dinâmicos sem a utilização de luvas coloridas e ambientes controlados.

Com o propósito de melhor apresentar o estudo realizado neste trabalho, o restante desta introdução apresentará os objetivos determinados, a metodologia adotada para alcançar esses objetivos, e por fim, a organização do trabalho.

1.1 Objetivos

Esta dissertação de mestrado tem como objetivo desenvolver um sistema de informação para analisar e reconhecer sinais dinâmicos em sequências de imagens (vídeos) relacionadas à LIBRAS e traduzir automaticamente as palavras expressas nas sequências dessas imagens para a língua portuguesa, utilizando apenas uma abordagem visual, sem o uso de luvas de dados e/ou coloridas ou a exigência de gravações de alta qualidade

em laboratórios com ambientes controlados, focando em sinais que utilizam apenas as mãos (sem considerar, por exemplo, as expressões faciais). Desta forma, este trabalho pretende proporcionar uma maior inclusão dos surdos e/ou deficientes auditivo na sociedade, facilitando a comunicação destes com pessoas que não têm o conhecimento de LIBRAS.

Este trabalho possui os seguintes objetivos específicos, derivados do objetivo geral desta dissertação:

- estudar, especificar e implementar uma ferramenta para segmentar sequências de imagens relacionadas à LIBRAS (separar a região de pele humana do restante das imagens das sequências), capturadas sem a utilização de luvas coloridas, marcadores e/ou ambientes controlados.
- estudar, especificar e implementar uma ferramenta para a identificar sinais dinâmicos em sequências de imagens relacionadas à LIBRAS e traduzi-los para o português.
- organizar um novo banco de imagens de sinais utilizados na comunicação em LIBRAS, sem a utilização de luvas coloridas, laboratórios com ambientes controlados e/ou exigências na vestimenta dos indivíduos que executam os sinais.

Vale ressaltar que, devido aos desafios encontrados no reconhecimento de línguas de sinais, não é objetivo desta dissertação produzir um sistema que seja uma solução definitiva para o reconhecimento automático de LIBRAS, mas sim, uma ferramenta extensível para auxiliar neste processo.

1.2 Metodologia

A metodologia deste trabalho consistiu, primeiramente, da revisão sistemática da literatura sobre reconhecimento automático de línguas de sinais, que, diferente da revisão tradicional da literatura, é uma metodologia rigorosa que utiliza métodos previamente definidos e explícitos para identificar, selecionar e avaliar criticamente pesquisas relevantes sobre um tema em questão (TORRE-UGARTE-GUANILO; TAKAHASHI; BERTOLOZZI, 2011). Com base nesta revisão, foram escolhidas as técnicas de processamento de imagens e de reconhecimento de sinais que serviram de base para a especificação e implementação do sistema para o reconhecimento automático de LIBRAS proposto.

As técnicas consideradas apropriadas para este trabalho foram combinadas ou estendidas de forma a se construir o sistema. Essas técnicas foram então implementadas,

testadas e validadas comparando os resultados obtidos com os disponíveis na literatura e com a ajuda de uma especialista do domínio de LIBRAS.

Em paralelo à etapa de implementação da ferramenta para o reconhecimento automático de LIBRAS, também foi construído um banco de imagens para ser utilizado nos testes e validação do sistema, sem a utilização de luvas coloridas e/ou laboratórios com ambientes controlados, o qual será apresentado com mais detalhes na seção 4.1.

O desenvolvimento do sistema foi feito na plataforma Eclipse[1], utilizando a linguagem de programação Java[2]. Esta escolha foi feita para facilitar a integração dos novos métodos implementados neste projeto com as ferramentas já desenvolvidas anteriormente (DIGIAMPIETRI et al., 2012; TEODORO, 2012).

Também foi utilizada a biblioteca de funções OpenCV[3] (*Open Source Computer Vision*) versão 2.4.8 na implementação de algumas técnicas de processamento de imagens. Este pacote é gratuito e mantido por diversos pesquisadores, universidades e usuários interessados na área de processamento de imagens.

Para avaliar os resultados dos testes realizados, foi utilizada a técnica de validação cruzada (*cross-validation*) e foram calculadas as medidas padrão de acurácia, sensibilidade e especificidade. Para avaliar os métodos de segmentação implementados, também foi utilizado o índice *overlap*.

1.3 Organização do trabalho

Este texto está dividido em sete capítulos, além desta introdução. O capítulo 2 fornece fundamentos básicos para o entendimento e acompanhamento do trabalho. O capítulo 3 contém a revisão sistemática dos trabalhos correlatos. O capítulo 4 descreve a especificação e desenvolvimento do sistema proposto. O capítulo 5 expõe e discute os experimentos realizados, enquanto que o capítulo 6 as conclusões, destacando as principais contribuições e possibilidades de trabalhos futuros. Por fim, são incluídos três apêndices e um anexo na seguinte ordem: o apêndice A contém exemplos de sequências de imagens para cada palavra do banco de imagens construído; no apêndice B é apresentado o Termo de Consentimento Livre e Esclarecido (TCLE) assinado pelos voluntários que participaram da construção do banco de imagens; o apêndice C apresenta os resultados finais obtidos

[1] http://www.eclipse.org
[2] http://www.java.com
[3] http://docs.opencv.org/2.4.8/index.html

no reconhecimento das palavras; e o anexo A contém o parecer constatando a aprovação desta pesquisa pelo Comitê de Ética em Pesquisa Envolvendo Seres Humanos da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP).

2 Conceitos fundamentais

Neste capítulo serão apresentadas definições de alguns conceitos e ferramentas utilizados na especificação e implementação do sistema proposto, a fim de proporcionar subsídios para um melhor entendimento e acompanhamento deste trabalho.

2.1 Língua Brasileira de Sinais (LIBRAS)

As línguas de sinais (línguas gestuais) são línguas naturais utilizadas como forma de comunicação entre deficientes auditivos. Elas não são simplesmente mímicas e gestos soltos utilizados pelos surdos para facilitar a comunicação; são línguas com estruturas gramaticais próprias. Assim como a língua falada, é composta por sua própria gramática, semântica, pragmática, sintaxe e outros elementos que preenchem os requisitos básicos para ser considerada um instrumento linguístico eficiente (CHAVEIRO et al., 2009).

Cada país possui a sua própria língua de sinais que, assim como as demais línguas, é influenciada pela cultura nacional e também possui expressões que diferem de região para região (os regionalismos).

A Língua Brasileira de Sinais (LIBRAS) é a língua oficial da comunidade surda brasileira e pode ser aprendida por qualquer pessoa que tenha interesse. A LIBRAS obteve reconhecimento oficial do governo brasileiro como meio legal de comunicação e expressão da comunidade surda no Brasil pela Lei nº10.436, de 24 de abril de 2002, regulamentada três anos mais tarde em 22 de dezembro de 2005, pelo Decreto nº5.626.

A LIBRAS tem origem na Língua Francesa de Sinais (LSF) e foi sendo modificada com o tempo por meio da influência da cultura nacional. Essa influência linguística da LSF sobre a LIBRAS teve início com a vinda do professor surdo francês Hernest Huet ao Brasil em 1855, a convite de Dom Pedro II, para fundar a primeira escola para surdos brasileiros, o Instituto Imperial de Surdos-Mudos, hoje conhecido como Instituto Nacional de Educação de Surdos (INES)[1], localizado na capital do Rio de Janeiro (GOLDFELD, 1997).

Os sinais são formados a partir da combinação de cinco parâmetros: (a) configuração de mão; (b) ponto de articulação; (c) movimento (STOKOE, 2005); (d) orientação/direção; e (e) expressões não manuais (KLIMA; BELLUGI, 1979).

[1] http://www.ines.gov.br/

A configuração de mão é a forma assumida pela mão durante a execução de um determinado sinal, podendo ser a mesma durante toda a execução do sinal ou variar para outra configuração, ou seja, existem sinais que são formados por mais de uma configuração de mão. Não há um consenso sobre quantas e quais configurações de mão são mais adequadas para a descrição das línguas de sinais. Há estudos de LIBRAS que apresentam de 46 a 73 configurações (FERREIRA et al., 2011). A Figura 1 apresenta as 73 configurações presentes no Dicionário Digital do INES[2] (LIRA; SOUZA, 2008).

Figura 1 – 73 configurações de mão extraídas do Dicionário Digital do INES

Fonte: (LIRA; SOUZA, 2008)

O ponto de articulação representa a região onde o sinal é realizado, podendo ser em pontos do corpo ou no espaço (STOKOE, 2005).

O movimento é definido como a ação das mãos no espaço em torno do enunciador. Os sinais podem ter ou não movimento, mas a grande maioria deles possui.

A orientação é a direção para a qual a palma da mão aponta durante a execução de um sinal, podendo ser para cima ou para baixo, para frente ou para trás e para os lados (direita e esquerda) (KLIMA; BELLUGI, 1979).

Por fim, as expressões não manuais (faciais e corporais) compreendem os movimentos da face, olhos, cabeça e tronco durante a execução do sinal, intensificando a característica do sinal, sendo assim um parâmetro importante para a identificação e entendimento real do sinal (SOARES, 2014).

[2] http://www.acessobrasil.org.br/libras/

2.2 Métodos de aquisição de sinais

Atualmente, existem duas abordagens principais na área de reconhecimento de língua de sinais para aquisição de sinais: a abordagem visual, em que os dados são obtidos por uma câmera de vídeo, e a abordagem baseada em dispositivos eletromecânicos, como luvas de dados e sensores.

A primeira abordagem é mais adequada para ser aplicada no dia a dia, sendo mais conveniente para o usuário, mas exige um pré-processamento mais delicado. Na segunda abordagem existem desconfortos e limitações para o usuário utilizar uma luva de dados e/ou sensores, além de ser geralmente de custo mais elevado, porém esta abordagem facilita o reconhecimento, evitando problemas enfrentados na primeira abordagem, como o de segmentação e rastreamento das mãos. A fim de diminuir a complexidade do pré-processamento das imagens obtidas por meio da abordagem visual, muitos trabalhos utilizam ambientes controlados e/ou luvas coloridas na gravação dos vídeos (HIENZ; GROBEL; OFFNER, 1996; GROBEL; ASSAN, 1997), que facilitam o reconhecimento dos sinais.

A Figura 2 apresenta um exemplo de luva de dados bastante utilizada em trabalhos de reconhecimento de sinais, a CyberGlove II, uma luva de dados com 22 sensores, conectada via *bluetooth*, capaz de captar os movimentos da mão e dos dedos (YAZADI, 2009). Já a Figura 3 apresenta um exemplo de luva multicolorida utilizada em trabalhos prévios (TEODORO, 2012; DIGIAMPIETRI et al., 2012; TEODORO; DIGIAMPIETRI, 2013), com o objetivo de facilitar a identificação das mãos e consequentemente o reconhecimento dos sinais.

2.3 Técnicas de processamento de imagens

Nesta seção serão definidos alguns conceitos básicos sobre as técnicas de processamento de imagens utilizadas.

2.3.1 Modelos de representação de cores

Os modelos ou espaço de cores são métodos utilizados para definir cores, especificando-as em um formato padronizado para atender a diferentes dispositivos gráficos ou aplicações que requerem a manipulação de cores.

Figura 2 – Exemplo de luva de dados utilizada em trabalhos de reconhecimento de sinais.

Fonte: Disponível em: www.cyberglovesystems.com/products/cyberglove-II/photos-video. Acesso em: 24 de junho de 2015.

Figura 3 – Exemplo de luva colorida utilizada em trabalhos de reconhecimento de sinais.

Fonte: Elaborada pela autora.

Em linhas gerais, pode-se dizer que um modelo de cores é uma representação multidimensional (tipicamente tridimensional) na qual cada cor é representada por um ponto no sistema de coordenadas multidimensional. Os modelos mais utilizados para representação de cores são: *RGB* (*red*, *green*, *blue*), *CMY* (*cyan*, magenta, *yellow*), *CMYK* (variante do modelo *CMY*, onde *K* denota *black*), *YCbCr* (padrão normalizado pela recomendação ITU-R BT.601 e utilizado em várias técnicas de compressão de vídeo), *YIQ* (padrão NTSC de TV em cores) e *HSI* (*hue*, *saturation*, *intensity*), às vezes também denominado *HSV* (*hue*, *saturation*, *value*) (FILHO; NETO, 1999).

Os modelos de cores mais frequentemente utilizados para processamento de imagens são o *RGB*, o *YIQ* e o *HSI* (GONZALEZ; WOODS, 2000).

O modelo *RGB* define a cor utilizando a combinação dos componentes vermelho (R), verde (G) e azul (B). Esses componentes são as quantidades de luz vermelha, verde e azul que uma cor *RGB* possui. Quando os valores de todos os três componentes são iguais, o resultado é um tom de cinza. Considerando que os valores dos componentes são codificados em 8 bits, variando de 0 a 255, quando o valor de todos os componentes é igual a 0, o resultado é preto puro e quando o valor de todos os componentes é definido como 255, o resultado é branco puro. É o modelo mais utilizado por câmeras e monitores de vídeo (GONZALEZ; WOODS, 2000).

O modelo *YIQ* é utilizado para transmissão de sinal de televisão em cores. O componente Y corresponde à luminância e fornece todas as informações de vídeo necessárias para um aparelho de TV monocromático, enquanto que os componentes I (matiz) e Q (saturação) juntos codificam as informações de crominância. A conversão de *RGB* para *YIQ* pode ser obtida pela Equação 1 (GONZALEZ; WOODS, 2000).

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

O modelo *HSI* representa uma cor em termos de matiz (H), saturação (S) e intensidade (I), da forma como o ser humano as percebe. Sua utilização é mais intensa em sistemas de visão artificial fortemente baseados no modelo de percepção de cor pelo ser humano. As equações de conversão de *RGB* para *HSI* são consideravelmente mais complexas do que para os modelos anteriores e podem ser encontradas descritas detalhadamente em (GONZALEZ; WOODS, 2000), assim como os demais modelos de cores.

2.3.2 Equalização de histograma

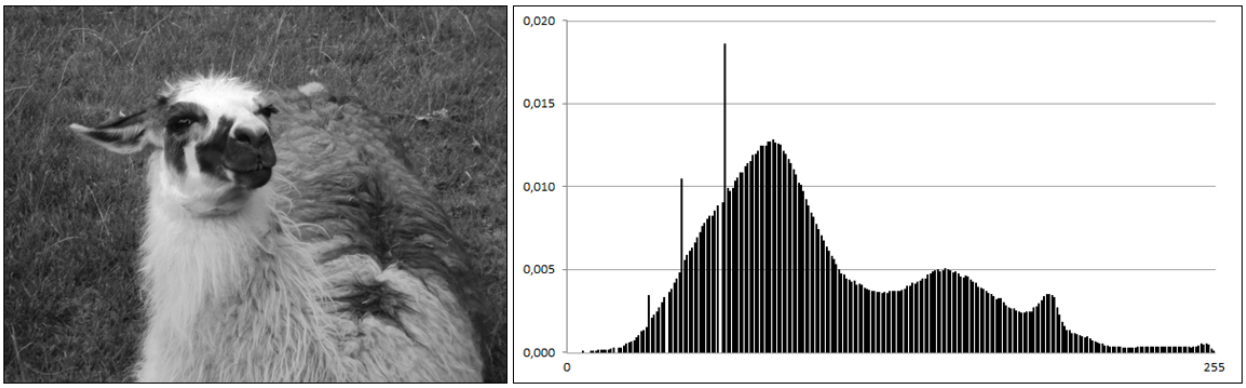
O histograma de uma imagem pode ser visto como uma função que fornece a frequência de cada cor na imagem, normalmente representado por um gráfico de barras, conforme exemplificado na Figura 4.

Dada uma imagem digital com níveis de cinza no intervalo $[0, L-1]$, o seu histograma pode ser calculado pela Equação discreta 2.

$$p(r_k) = \frac{n_k}{n} \tag{2}$$

Sendo r_k o k-ésimo nível de cinza, n_k o número de pixels cujo nível de cinza corresponde a k, n o número total de pixels na imagem e $k = 0, 1, ..., L-1$ (GONZALEZ; WOODS, 2000).

Figura 4 – Exemplo de histograma para imagem com 256 níveis de cinza.



Fonte: Elaborada pela autora.

A partir do histograma de uma imagem é possível observar a sua qualidade quanto ao nível de contraste (baixo ou alto contraste) e se a imagem é predominantemente clara ou escura. Também é possível manipular e modificar o histograma de uma imagem a partir de algumas técnicas de processamento de imagens, a fim de melhorar sua qualidade; a equalização de histograma é uma delas.

A técnica de equalização de histograma, também conhecida como "linearização de histograma", é aplicada com a finalidade de obter um histograma uniforme, a partir do espalhamento da distribuição dos níveis de cinza dos pixels em uma imagem. Para tanto, utiliza-se uma função auxiliar, denominada função de transformação. A forma mais usual de se equalizar um histograma é utilizar a função de distribuição acumulada (*cdf* -

cumulative distribution function) da distribuição de probabilidades original, que pode ser expressa pela Equação 3 (FILHO; NETO, 1999):

$$s_k = T(r_k) = \sum_{j=0}^{k} \frac{n_j}{n} = \sum_{j=0}^{k} p_r(r_j) \tag{3}$$

onde:

$0 \leq r_k \leq 1$

$k = 0, 1, ..., L-1$

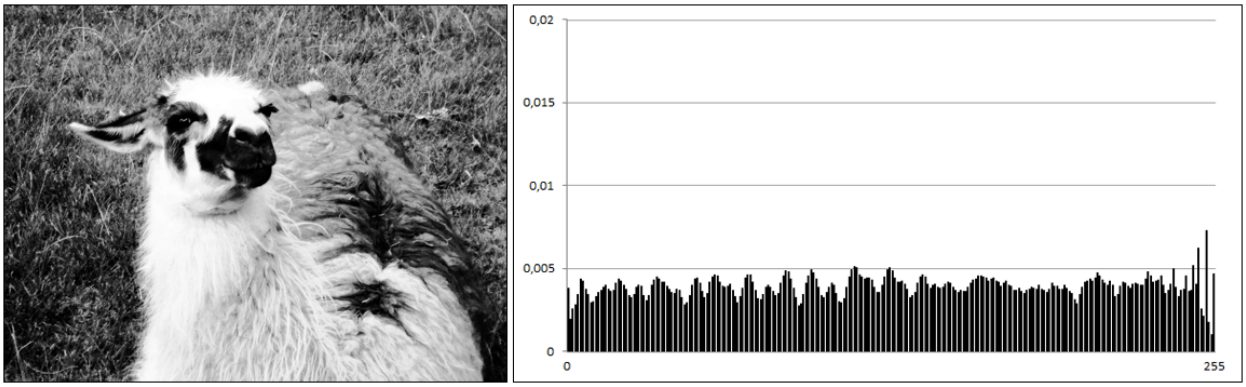
Primeiramente é calculado o histograma da imagem e um outro histograma, que é o histograma ideal ou esperado para a imagem. Ambos os histogramas são colocados na forma de histograma acumulado. Em seguida, os valores quantitativos de cada tom de cinza presentes no histograma acumulado original são comparados com os do histograma acumulado ideal, gerando um terceiro histograma uniforme, que é o histograma equalizado da imagem, e que possui os níveis de cinza igualmente distribuídos dentro do intervalo $[0, 1]$, gerando uma imagem com maior contraste.

Além de ser feito de maneira global, em que a transformação é executada usando todos os pixels da imagem, a equalização de histograma também pode ser realizada localmente, aplicando-se a equalização de histograma na vizinhança de cada pixel da imagem. Desta forma, somente o valor do pixel central da vizinhança é modificado. O centro da região é então movido para o pixel adjacente e o procedimento é repetido. Essa técnica é chamada de equalização local de histograma e é útil para realçar detalhes de áreas pequenas.

As Figuras 5 e 6 apresentam a imagem da Figura 4 após a aplicação das técnicas de equalização global e local, respectivamente, junto com seus histogramas.

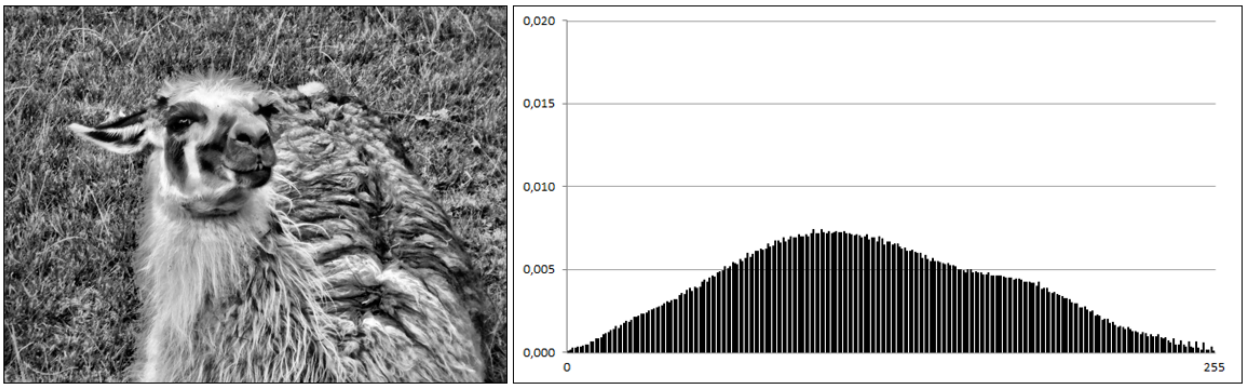
Em imagens coloridas também é possível aplicar o conceito de histograma. Neste caso, a imagem é decomposta de alguma maneira (por exemplo, em seus componentes R, G e B ou luminância) e para cada componente é calculado o histograma correspondente. A Figura 7 apresenta um exemplo de resultado obtido pela aplicação da técnica de equalização local em uma imagem colorida. Neste caso, foram utilizados apenas os níveis de luminosidade da imagem para realizar a equalização.

Figura 5 – Exemplo de resultado gerado pela aplicação da equalização global.



Fonte: Elaborada pela autora.

Figura 6 – Exemplo de resultado gerado pela aplicação da equalização local.



Fonte: Elaborada pela autora.

Figura 7 – Exemplo de resultado gerado pela aplicação da equalização local em uma imagem colorida: (a) imagem original; (b) imagem com histograma equalizado.

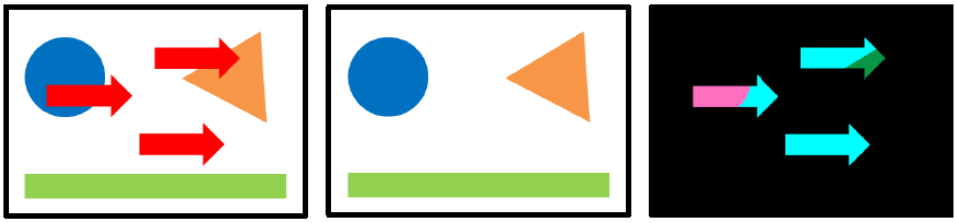
Fonte: Elaborada pela autora.

2.3.3 Subtração de fundo em sequências de imagens

A subtração de fundo (*background subtraction*) é uma das principais etapas de pré-processamento, bastante utilizada em aplicações baseadas na abordagem visual para separar objetos ou partes de objetos dinâmicos (em movimento) do restante da imagem. Desta forma, elementos da cena que não se desejam analisar podem ser removidos, tornando o processamento mais rápido ou, muitas vezes, mais preciso.

Uma das técnicas mais usadas de subtração de fundo em vídeos consiste em comparar cada *frame* do vídeo com um *frame* de referência. Este *frame* deve conter apenas os elementos estacionários (parados) da cena, ou seja, o "fundo" da imagem. Quando um pixel do *frame* analisado é muito diferente do pixel correspondente no *frame* de referência, considera-se que esse pixel pertence a um objeto em movimento (BRITTO, 2011). A Figura 8 apresenta um exemplo de subtração de fundo aplicada a uma imagem em que as setas representam os objetos em movimento e as formas geométricas os objetos estacionários.

Figura 8 – Representação da aplicação da técnica de subtração de fundo: (a) *frame* do vídeo; (b) imagem de fundo; (c) resultado da subtração de fundo.



Fonte: (BRITTO, 2011)

2.3.4 Fechamento

O fechamento é uma técnica de processamento de imagens utilizada para reparar imagens, obtida a partir do encadeamento do filtro morfológico de dilatação, seguido pelo de erosão. Essa técnica é utilizada para suavizar contornos, unir quebras estreitas e golfos longos e delgados e remover os pixels ruidosos do interior do objeto. Ela preenche vazios mas mantém a forma e o tamanho do objeto inalterados (GONZALEZ; WOODS, 2000).

Sendo A e B conjuntos de Z^2 e $\emptyset$ o conjunto vazio, a dilatação de A por B, denotada por $A \oplus B$, é definida pela Equação 4.

$$A \oplus B = \{x \mid (\widehat{B})_x \cap A \neq \emptyset\} \tag{4}$$

O processo de dilatação começa na obtenção da reflexão de B em torno de sua origem, seguido da translação dessa reflexão por x. Já o processo de erosão de A por B, é o conjunto de todos os pontos x tais que B, quando transladado por x, fique contido em A. Desta forma, a erosão de uma imagem A por um elemento estruturante B, denotada por $A \ominus B$, é definida pela Equação 5. Segundo Facon (1996), o elemento estruturante é um conjunto definido e conhecido (forma e tamanho), o qual é comparado, a partir de uma transformação, ao conjunto desconhecido da imagem.

$$A \ominus B = \{x \mid (B)_x \subseteq A\} \tag{5}$$

O fechamento do conjunto A pelo elemento estruturante B é a dilatação de A por B seguida da erosão do resultado pelo mesmo elemento estruturante B, denotado por $A \bullet B$ e definido pela Equação 6.

$$A \bullet B = (A \oplus B) \ominus B \tag{6}$$

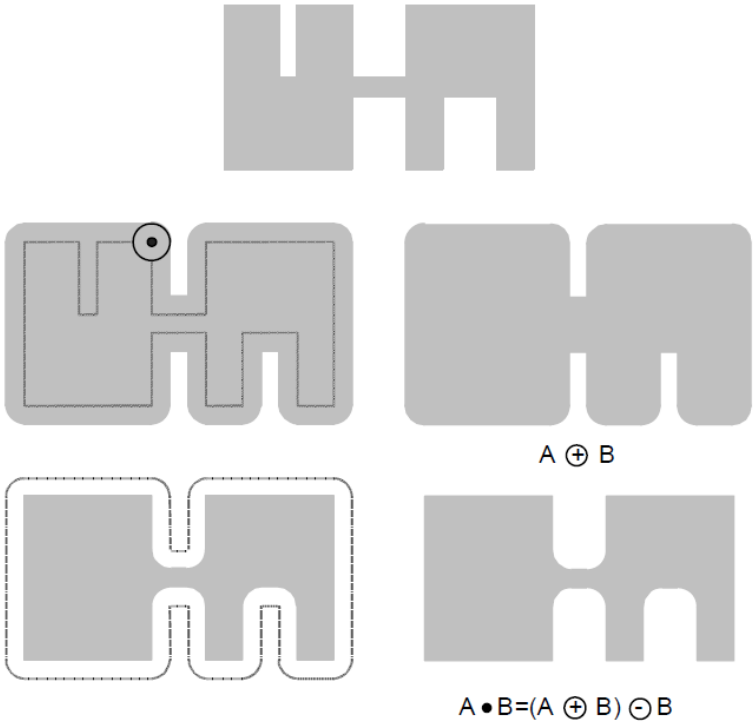
A figura 9 mostra um exemplo de operação de fechamento utilizando um elemento estruturante circular.

2.3.5 Filtro da mediana

O filtro da mediana é um filtro espacial passa-baixa, também conhecido como filtro de suavização. A sua função principal é de forçar pontos com intensidades distintas a assemelharem-se a seus vizinhos. Este filtro é utilizado com o objetivo de alcançar a redução de ruído em vez de borrar, preservando bordas e detalhes finos da imagem.

Nesta técnica, cada pixel da imagem final é substituído pelo nível de cinza mediano dos pixels situados em sua vizinhança. O nível mediano m de um conjunto de n elementos é tal que metade dos n elementos do conjunto são menores que m e a outra metade é constituída de elementos maiores que m. Quando n é ímpar, o nível mediano é o próprio elemento central do conjunto ordenado. Nos casos em que n é par, o nível mediano é

Figura 9 – Exemplo de operação de fechamento.



Fonte: (FILHO; NETO, 1999)

calculado pela média aritmética dos dois elementos mais próximos do centro (GONZALEZ; WOODS, 2000).

A figura 10 apresenta um exemplo de resultado gerado pela aplicação da filtragem por mediana com máscara 3x3 na imagem da figura 4 contaminada por ruído.

Figura 10 – Exemplo da aplicação do filtro da mediana: (a) imagem contaminada por ruído; (b) resultado da filtragem pelo filtro da mediana com máscara 3x3.

Fonte: Elaborada pela autora.

2.3.6 Segmentação

A tarefa básica da etapa de segmentação é a de dividir uma imagem digital em múltiplas regiões (conjunto de pixels) ou objetos de interesse que a compõem, eliminando os ruídos e artefatos da imagem. Esta tarefa, apesar de simples de descrever, é uma das mais difíceis de implementar e é fundamental para determinar o eventual sucesso ou fracasso do processo de análise da imagem (FILHO; NETO, 1999).

Apesar do ser humano ser capaz de identificar regiões com as mesmas características ou os objetos presentes em uma imagem, para se realizar a mesma tarefa com um computador deve-se implementar algoritmos que analisem a distribuição dos pixels ou as características de cada um deles.

As técnicas de segmentação de imagens podem ser classificadas nas seguintes principais categorias: detecção de descontinuidades, limiarização, orientada a regiões e métodos híbridos (GONZALEZ; WOODS, 2000). Descrições completas das técnicas podem ser encontradas no livro de Gonzalez e Woods (2000). A escolha de uma técnica de segmentação em relação à outra é feita baseando-se principalmente nas características do problema em questão.

2.4 Reconhecimento de padrões

Reconhecimento de padrões é uma área de pesquisa na qual o principal objetivo é a classificação de objetos (padrões) em um número de categorias ou classes a partir da observação de suas características (THEODORIDIS; KOUTROUMBAS, 2006). Dependendo da aplicação, esses objetos podem ser qualquer tipo de informação que possa ser mensurada de alguma maneira e classificada posteriormente (imagens e áudios, por exemplo).

No caso do reconhecimento de língua de sinais que utilizam a abordagem visual, as características extraídas das sequências de imagens (vídeos) dos indivíduos executando um sinal em língua de sinais são os objetos, enquanto que as classes são as letras ou palavras referentes em língua portuguesa.

Um projeto de reconhecimento de padrões é formado normalmente por três etapas:

- Extração de características dos objetos a classificar (ou a descrever);
- Seleção das características mais discriminativas;

- Construção de um classificador (ou descritor).

O reconhecimento de padrões pode ser realizado de duas maneiras: supervisionada ou não supervisionada. A abordagem supervisionada é baseada em um conjunto de dados previamente rotulado com classes pré-definidas (conjunto de treinamento), construindo um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados. Já na abordagem não supervisionada, as classes são desconhecidas, sendo os dados agrupados com base na similaridade entre os padrões de treinamento não rotulados.

2.5 Alinhamento de sequências

O alinhamento de sequências é uma técnica bastante utilizada na área de bioinformática para medir a similaridade entre duas ou mais sequências biológicas, podendo ser global ou local. Em geral, as sequências biológicas são polímeros representados por uma cadeia de caracteres ("strings") e a comparação entre elas é feita comparando apenas suas respectivas letras. Apesar da aparente simplicidade do processo, a análise de similaridade das sequências é uma tarefa consideravelmente complexa.

Durante o alinhamento, uma sequência é organizada na linha e a outra na coluna da matriz. As técnicas de alinhamento buscarão então a melhor correspondência para as sequências.

A ideia central dos algoritmos de alinhamento é minimizar as diferenças entre as sequências, buscando um alinhamento ótimo. Geralmente, a similaridade entre as sequências envolvidas é expressa pelo termo identidade, que quantifica a porcentagem de caracteres idênticos entre duas sequências (JUNQUEIRA; BRAUN; VERLI, 2014).

2.5.1 Alinhamento global

Os algoritmos que utilizam a técnica de alinhamento global buscam comparar duas sequências em toda a sua extensão. Quando necessário, estes algoritmos permitem a inserção de espaçamentos (*gaps*) entre os caracteres para que, ao final, todas as sequências tenham o mesmo comprimento. É apropriado para comparar sequências cujas semelhanças sejam esperadas em toda a sua extensão.

O principal algoritmo envolvido no processamento de alinhamentos globais é o algoritmo desenvolvido por Needleman e Wunsch durante a década de 1970, que foi o primeiro algoritmo a utilizar o método de programação dinâmica para a comparação de sequências biológicas.

2.5.2 Alinhamento local

Os algoritmos de alinhamento local procuram locais de semelhança entre as sequências sem ter de considerar todo o comprimento destas, ou seja, busca somente alinhamento de regiões de alta similaridade, não importando as sequências adjacentes a estas regiões. É muito útil para fazer pesquisas em base de dados e em situações em que não existe qualquer conhecimento sobre a semelhança entre as sequências a comparar. Nestes algoritmos, o alinhamento termina no final das regiões de alta similaridade e substitui as regiões excluídas por buracos (*gaps*) no resultado final.

O algoritmo desenvolvido por Smith e Waterman (1981) ganhou maior destaque e atualmente é o principal algoritmo utilizado por programas para realização de alinhamentos locais. Este algoritmo é semelhante ao algoritmo de cálculo da distância de edição de *Levenshtein*, que será apresentado posteriormente, com a exceção de pequenas alterações. Ele alinha duas substrings e não duas strings como o de *Levenshtein*.

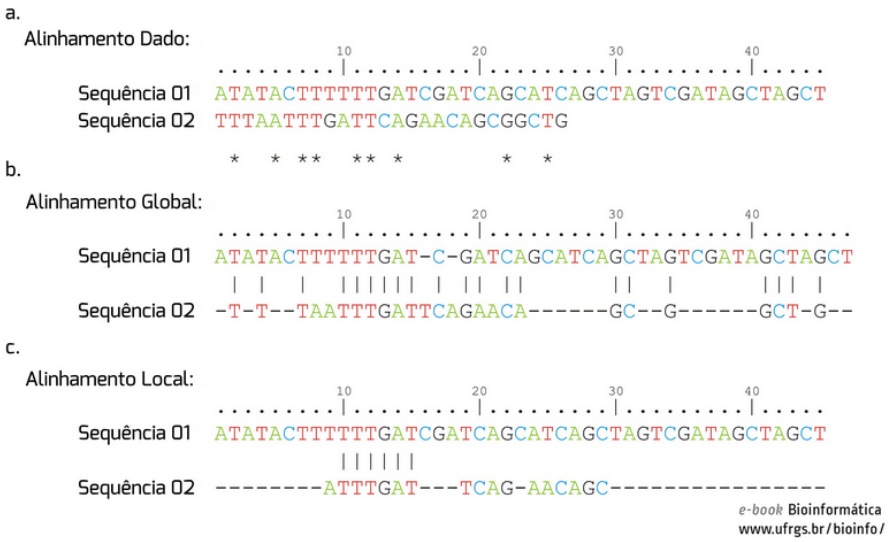
A figura 11 apresenta as diferenças entre alinhamento local e global. Uma descrição mais completa sobre alinhamento de sequências pode ser encontrado em (JUNQUEIRA; BRAUN; VERLI, 2014), assim como os principais métodos utilizados.

2.5.3 Distância de edição

A distância de edição ou distância de *Levenshtein* (LEVENSHTEIN, 1966) é uma das formas mais comuns e simples de medir a semelhança/diferença entre duas sequências de caracteres. Ela corresponde à transformação/edição de uma cadeia de caracteres em uma outra por meio de uma série de operações de edição, de forma individual, sobre cada um dos caracteres da cadeia.

Este algoritmo pode ser parafraseado como “o menor número de inserções, remoções e substituições para igualar duas strings” (NAVARRO, 2001).

Figura 11 – Diferenças entre alinhamento local e global. a) Duas sequências de nucleotídeos de tamanhos diversos são amostradas e alinhadas por algoritmos diferentes. b) No alinhamento local, a prioridade é encontrar as regiões altamente similares, independentemente do tamanho desta região. Neste caso, porções da sequência que não foram alinhadas com alta similaridade foram excluídas do resultado final. c) No alinhamento global, as duas sequências são alinhadas por completo, independentemente do número de lacunas que tenham que ser inseridas.



Fonte: (JUNQUEIRA; BRAUN; VERLI, 2014)

São definidos valores (*scores*) diferentes para cada possível operação: *match* (casamento, igualdade dos caracteres), *mismatch* (substituição), inserção e remoção. Todas as operações são avaliadas para se chegar à menor distância total (*score*), sendo possível definir diferentes penalidades de custo à inserção, deleção e substituição.

Uma matriz é montada a partir do tamanho das strings e os custos de cada operação são configurados. Ao final das comparações entre os caracteres das strings, a distância é dada pela última posição da matriz, sendo a distância zero a indicação de que as strings são idênticas (LEVENSHTEIN, 1966).

2.6 Medidas de avaliação

Nesta seção serão apresentadas as medidas utilizadas para avaliar os algoritmos de classificação e segmentação implementados neste trabalho. Sabendo-se que VP, VN, FP e FN são respectivamente, os números de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Para a segmentação, os verdadeiros positivos são os pixels

pertencentes a “pele” e os quais também foram classificados como “pele”, os verdadeiros negativos são os pixels que não são “pele” (neste caso, chamaremos de “fundo”) e os quais também foram classificados como “fundo”, os falsos positivos são os pixels que são “fundo” mas foram classificados como “pele” e os falsos negativos são os pixels que são “pele” mas foram classificados como “fundo”.

2.6.1 Acurácia

A acurácia mede a capacidade do algoritmo em determinar corretamente o que é verdadeiro entre todas as amostras, ou seja, o quão preciso é o algoritmo. Ela é definida pela Equação 7.

$$\text{Acurácia} = \frac{VP + FP}{(VP + FP + VN + FN)} \times 100 \tag{7}$$

2.6.2 Sensibilidade

A sensibilidade mede a capacidade do algoritmo em identificar corretamente o que é verdadeiro entre as amostras que são verdadeiras, ou seja, o quão sensível é o teste. Ela é definida pela Equação 8.

$$\text{Sensibilidade} = \frac{VP}{(VP + FN)} \times 100 \tag{8}$$

2.6.3 Especificidade

A especificidade mede a capacidade do algoritmo em identificar corretamente o que é falso entre as amostras que são falsas, ou seja, o quão específico é o teste. Ela é definida pela Equação 9.

$$\text{Especificidade} = \frac{VN}{(VN + FP)} \times 100 \tag{9}$$

2.6.4 Overlap

A medida *overlap* é uma das mais citadas na literatura para avaliar segmentação. Essa métrica consiste na área relativa da intersecção entre duas regiões consideradas (GRUSZAUSKAS et al., 2008). Considerando-se A_s, a área de uma região segmentada automaticamente, e A_m, a área considerada correta para o processo de segmentação, neste caso, segmentada manualmente por um indivíduo, a métrica *overlap* é definida pela Equação 10. O valor 0 indica o pior desempenho, ou seja, não há intersecção entre a área considerada correta e a área obtida automaticamente. O valor 1, por sua vez, indica uma segmentação perfeita.

$$\text{Overlap} = \frac{A_s \cap A_m}{A_s \cup A_m} \tag{10}$$

2.7 Validação cruzada

A validação cruzada é uma técnica utilizada para avaliar a capacidade de generalização de um modelo de classificação a partir de um conjunto de dados.

O conceito central dos métodos de validação cruzada é dividir o conjunto de dados em subconjuntos mutualmente exclusivos. Destes subconjuntos, alguns são utilizados para estimar os parâmetros do modelo (dados de treinamento) e o restante (dados de validação ou de teste) é aplicado na validação do modelo.

Existem diversos métodos de particionar os dados, sendo os dois mais utilizados: método *k-fold* e *leave-one-out*.

2.7.1 Método K-fold

O método de validação cruzada *k-fold* (*k-fold cross-validation*) é um dos mais utilizados. Este método consiste em dividir o conjunto de dados em k subconjuntos (*folds*) mutuamente exclusivos de mesmo tamanho. Destes, *k-1* subconjuntos são utilizados para o treinamento e um é utilizado para o teste. A acurácia do modelo pode ser então calculada a partir da Equação 7 apresentada anteriormente, assim como qualquer outra medida.

Este processo é repetido k vezes, alternando de forma que cada subconjunto seja utilizado uma vez como conjunto de teste.

Ao final das k iterações, as medidas de avaliação são calculadas pela média dos resultados obtidos em cada etapa, obtendo-se assim uma medida sobre a capacidade do modelo de representar o processo gerador dos dados, permitindo análises estatísticas (SANTOS et al., 2009).

2.7.2 Método leave-one-out

O método *leave-one-out* é um caso particular do método *k-fold*, em que o número de *folds* é igual ao número de instâncias *N*. Neste caso, são realizadas *N* iterações e, em cada iteração, *N-1* instâncias são utilizadas para o treinamento e o restante é utilizado para o teste, alternando de forma que cada instância seja utilizada uma vez para o teste.

Este método possui um alto custo computacional, desta forma, é indicado em casos com conjunto de dados consideravelmente pequeno.

2.8 Weka

Weka[3] (*Waikato Environment for Knowledge Analysis*) é um conjunto de algoritmos de aprendizado de máquina desenvolvido por um grupo de pesquisadores da Universidade de Waikato, Nova Zelândia. Ele é um software livre licenciado pela *GNU General Public License*, desta forma, se tem a liberdade de estudar, executar e alterar o respectivo código fonte para qualquer propósito (HALL et al., 2009).

Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamado a partir de seu próprio código Java. O Weka contém ferramentas para pré-processamento de dados, classificação, regressão, agrupamento (*clustering*), regras de associação e visualização. É também ideal para o desenvolvimento de novos sistemas de aprendizagem máquina.

[3] http://www.cs.waikato.ac.nz/ml/weka

2.9 OpenCV

OpenCV[4] (*Open Source Computer Vision*) é uma biblioteca de código aberto para desenvolvimento de algoritmos de visão computacional e aprendizado de máquina, licenciada pela licença BSD e amplamente utilizada em empresas, grupos de pesquisa e por órgãos governamentais. Possui interfaces C++, C, Python, Java e Matlab e suporta os sistemas operacionais Windows, Linux, Android e Mac OS (BRADSKI; KAEHLER, 2008).

A biblioteca possui módulos de processamento de imagens e vídeos, estrutura de dados, álgebra linear, interface gráfica para o usuário (GUI), controle de mouse e teclado, além de mais de 2.500 algoritmos, muitos dos quais são considerados estado da arte, tais como de segmentação, reconhecimento de faces, aprendizado de máquinas, filtragem de imagens, rastreamento de movimento, entre outros.

2.10 ImageJ

ImageJ[5] é um programa Java de domínio público de processamento e análise de imagens, desenvolvido pelos Institutos Nacionais de Saúde dos Estados Unidos (*National Institutes of Health* - NHI) amplamente utilizado nas ciências naturais graças à sua interface simples, a sua velocidade de processamento e capacidade de expansão. A partir do ImageJ é possível exibir, editar, analisar, processar, salvar e imprimir imagens. O seu código-fonte é totalmente acessível e o desenvolvedor é livre para adaptá-lo às suas necessidades se achar necessário.

[4] http://opencv.org
[5] http://imagej.nih.gov/ij/

3 Trabalhos correlatos

A fim de contextualizar o trabalho proposto em relação ao estado da arte da área de reconhecimento de língua de sinais, foi realizada uma revisão acerca dos trabalhos feitos sobre esta área. O objetivo da revisão foi identificar o estado da arte dos métodos, das técnicas de reconhecimento e os parâmetros relacionados à realização de sinais utilizados, focando nos trabalhos que utilizam a abordagem visual, isto é, quando os dados são obtidos por meio de câmera de vídeo, sem o uso de dispositivos eletromecânicos (luvas de dados e/ou sensores), identificando os principais desafios de se trabalhar com essa abordagem, que é a abordagem utilizada nesta dissertação. Adicionalmente, foram procurados trabalhos específicos sobre o processamento de Língua Brasileira de Sinais.

A revisão foi dividida em duas etapas. Em novembro de 2012 foi realizada uma revisão sistemática nas bibliotecas digitais da ACM e IEEE sobre trabalhos de reconhecimento de língua de sinais que utilizassem a abordagem visual. Em maio de 2014 esta revisão foi atualizada e adicionalmente foram procurados trabalhos na biblioteca BDBComp e em conferências específicas ligadas a área. A seguir são apresentados os resultados consolidados das duas etapas de revisão realizadas.

3.1 Método de revisão

Primeiramente, com o propósito de se familiarizar com os principais termos e conceitos relacionados ao campo de estudo objetivado, foi feita uma pesquisa exploratória sobre o assunto. A partir dessa pesquisa exploratória foram identificadas as seguintes palavras-chave relacionadas com o contexto: "*língua brasileira de sinais*", "*libras*", "*sign language*", "*recognition*", "*image*" e "*video*".

Para a seleção das fontes, foi tomado como critério aquelas que disponibilizam artigos na íntegra para acesso e que estejam relacionadas com as principais conferências e periódicos relacionados ao tema. Desta forma, três bibliotecas digitais foram selecionadas:

- ACM: ACM Digital Library[1].
- BDBComp: Biblioteca Digital Brasileira de Computação[2].

[1] dl.acm.org
[2] www.lbd.dcc.ufmg.br/bdbcomp

- IEEE: IEEExplore[3].

Além destas três bibliotecas, também foram buscados em todos os anais publicados e disponíveis para acesso na internet do Simpósio de Realidade Virtual e Aumentada (SVR), do Simpósio Brasileiro de Inteligência Artificial (SBIA) e do *SIBGRAPI Conference on Graphics, Patterns and Images*, sendo que este último, tem seus anais disponíveis na *SIBGRAPI Digital Library Archive*[4].

A partir das palavras-chave definidas, foram construídas as expressões de busca para cada biblioteca digital. Para a escolha destas expressões, foram testadas diversas *strings*, e foram escolhidas aquelas que apresentaram melhores e maiores resultados. A Tabela 1 apresenta as chaves e as opções de busca que foram utilizadas em cada fonte.

Tabela 1 – Chaves de busca utilizadas de acordo com a fonte de dados e condições utilizadas.

Fonte	**Chave de Busca**	**Condições de Filtragem**
ACM Digital Library	(“sign language”AND recognition) AND (image OR video)	Busca avançada, com varredura apenas no campo *abstract*
BDBComp	(“língua brasileira de sinais”) OR (libras)	Busca avançada, com varredura apenas no campo *título do trabalho*
IEEExplore	(“sign language” AND recognition) AND (image OR video)	Busca avançada, com o filtro “*metadata only*” ativo

As buscas foram realizadas nos dias quatro e cinco de novembro de 2012 nas bases ACM e IEEE, respectivamente. As mesmas buscas foram realizadas novamente no dia dois de maio de 2014, a fim de completá-las e atualizá-las. A busca na base BDBComp foi feita no dia cinco de maio de 2014 e as buscas nos anais dos simpósios SVR, SBIA e SIBGRAPI foram realizadas entre sete e nove de maio de 2014. Não foi definido um período para delimitar a busca, portanto, todos os artigos obtidos foram avaliados para seleção.

Os artigos foram selecionados com base na leitura do *abstract* de cada um dos resultados encontrados, levando em conta um critério de inclusão e quatro critérios de exclusão, sabendo-se que para ser aceito, o artigo precisa satisfazer o critério de inclusão e nenhum dos critérios de exclusão:

- Inclusão 1: trabalhos que discorram sobre reconhecimento de língua de sinais a partir de imagens e/ou vídeos digitais;
- Exclusão 1: trabalhos que utilizam dispositivos eletromecânicos (luvas de dados e/ou sensores) na aquisição de dados;

[3] ieeexplore.ieee.org
[4] sibgrapi.sid.inpe.br

- Exclusão 2: trabalhos que não estejam escritos na língua inglesa ou portuguesa;
- Exclusão 3: trabalhos que não estejam disponíveis na íntegra para acesso;
- Exclusão 4: trabalhos duplicados.

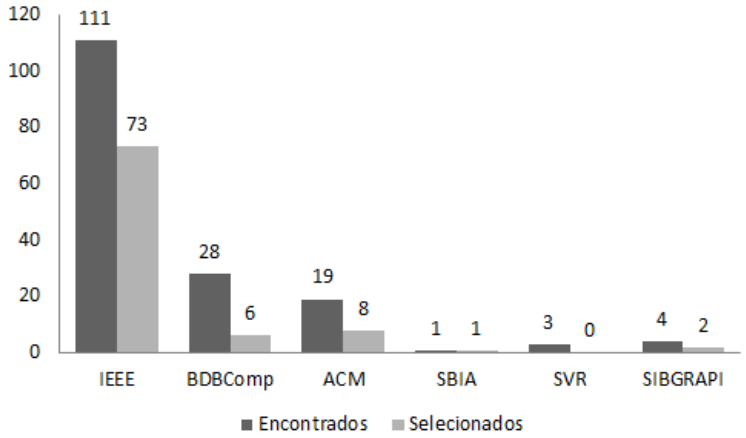
Os artigos selecionados foram lidos na íntegra e foi realizado um levantamento dos pontos considerados mais relevantes em cada um deles, identificando não só os métodos de aquisição de dados, os parâmetros relacionados à realização de sinais, as técnicas aplicadas no pré-processamento das imagens/vídeos e no processo de reconhecimento de língua de sinais, mas também os principais resultados obtidos. A partir destes dados foi feita a análise exposta adiante.

3.2 Resultados e discussão

A partir da pesquisa realizada nas bases de dados e nos anais dos simpósios apresentados na seção 3.1, obteve-se 166 artigos. Destes artigos, foram selecionados 127 que se enquadravam nos critérios de seleção estipulados. Estes artigos selecionados foram analisados e lidos na íntegra, passando pela fase de extração de dados. Nessa fase, mais 37 artigos também foram rejeitados por não se adequarem nos critérios de seleção, fato que não foi identificado apenas pela leitura do *abstract*, resultando 90 artigos.

A Figura 12 apresenta a distribuição dos artigos de acordo com as fontes de dados utilizadas.

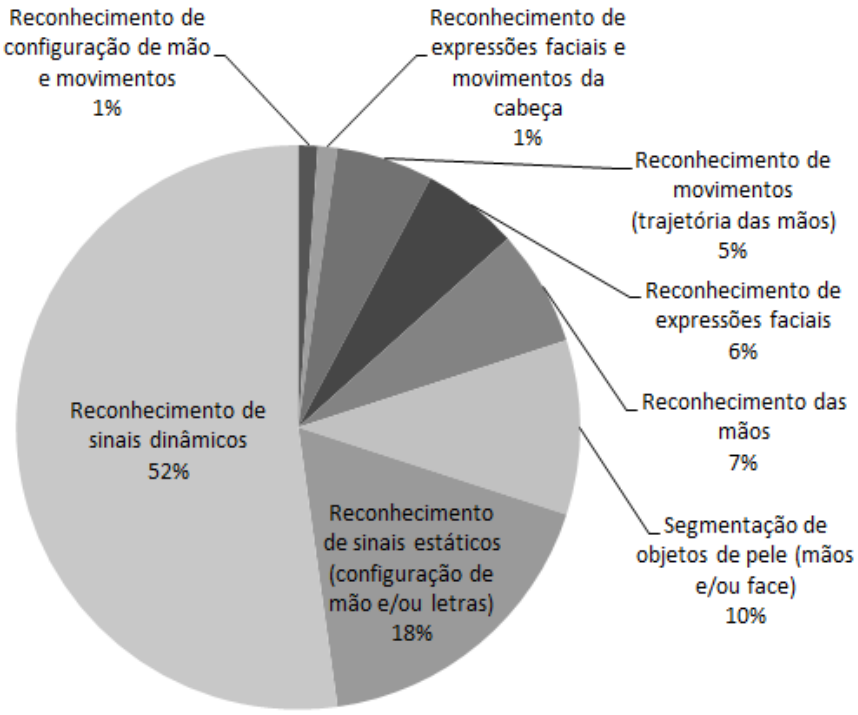
Figura 12 – Distribuição de artigos encontrados e selecionados por fonte de dados.



A distribuição dos artigos selecionados pelos seus principais objetivos está presente na Figura 13. Observa-se que a maioria dos trabalhos selecionados tem como objetivo

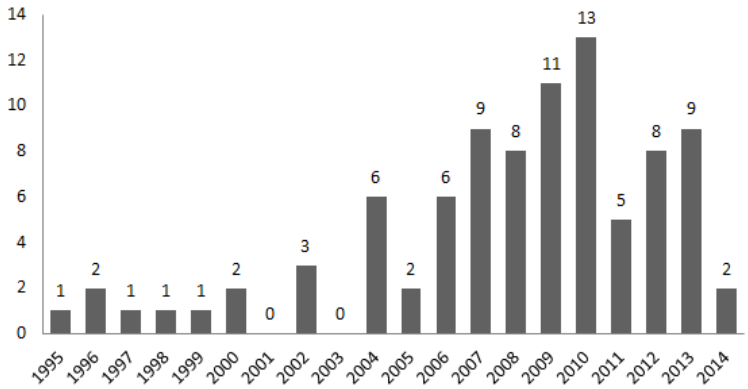
o reconhecimento de sinais dinâmicos, seguido pelo reconhecimento de sinais estáticos (configurações de mão e/ou letras), mas 30% apresentam apenas trabalhos que tratam sobre segmentação de imagens de língua de sinais ou reconhecimento de parâmetros relacionados à realização de sinais.

Figura 13 – Distribuição de artigos por principal objetivo.



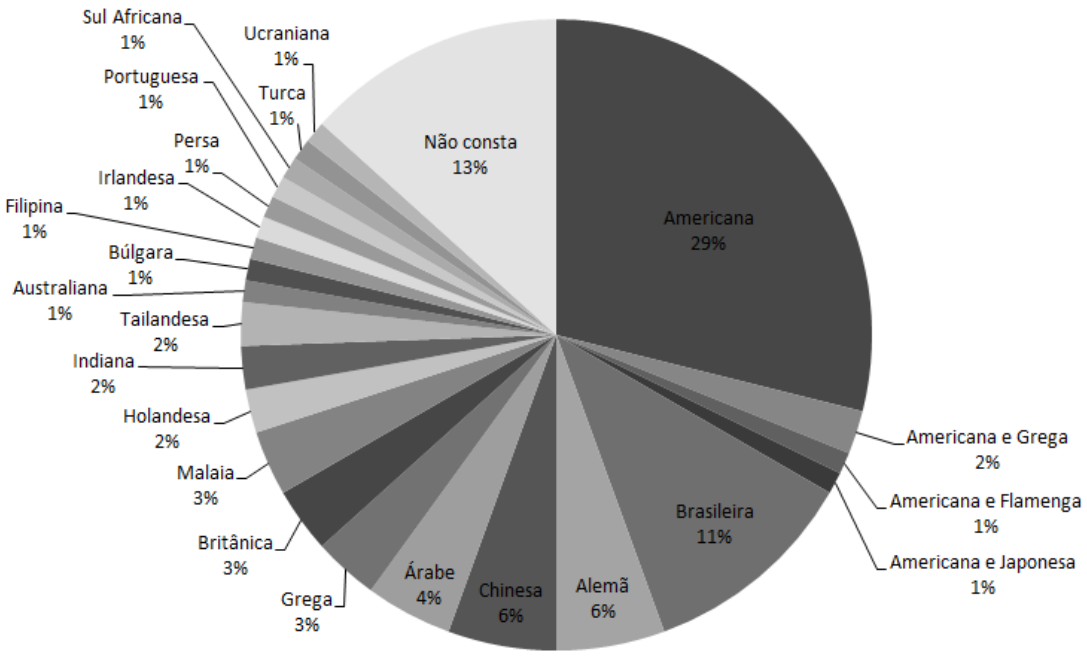
A distribuição dos artigos selecionados pelo ano de publicação é encontrado no gráfico da Figura 14. Nota-se que os artigos selecionados foram publicados durante os últimos 20 anos, e que a maioria corresponde a trabalhos mais recentes, publicados nos últimos 10 anos.

Figura 14 – Distribuição de artigos pelo ano de publicação.



Os artigos selecionados também foram classificados de acordo com a língua de sinais utilizada. Por meio desta classificação observou-se uma grande variedade de línguas de sinais, sendo a Língua de Sinais Americana a mais utilizada, enquanto que a Língua Brasileira de Sinais (LIBRAS) é a segunda língua de sinais mais utilizada. A distribuição dos artigos pelo tipo de língua de sinais é apresentada pelo gráfico da Figura 15.

Figura 15 – Distribuição de artigos por língua de sinais.



Em todos os trabalhos selecionados são utilizados vídeos gravados em estúdios por deficientes auditivos e/ou profissionais em língua de sinais. A maioria dos trabalhos selecionados não utiliza luvas coloridas para expressar os sinais na gravação dos vídeos (80%), o que exige um pré-processamento mais sofisticado das imagens.

A fim de facilitar o pré-processamento das imagens e consequentemente o reconhecimento dos sinais, alguns trabalhos utilizam vídeos gravados com algumas restrições. Em Hieu e Nitsuwat (2008), Goh e Holden (2006), Paulraj et al. (2008), Chanda, Auephanwiriyakul e Theera-Umpon (2012) é imposta a restrição de que os indivíduos que expressam os sinais nos vídeos devem utilizar camisas de manga longa de cores escuras. No trabalho de Hienz, Grobel e Offner (1996) é apresentada uma abordagem diferenciada. Além de utilizar luvas multicoloridas na gravação dos vídeos utilizados nos testes, também foram utilizados marcadores coloridos nos ombros e cotovelos, a fim de diminuir ainda

mais a complexidade do reconhecimento. Em Dimov, Marinov e Zlateva (2007) é utilizada uma restrição ainda mais intrusiva. Os indivíduos expressam os sinais em um ambiente controlado, em que eles ficam cobertos por um pano escuro e apenas a cabeça e as mãos aparecem.

A maioria dos trabalhos não especifica a base de imagens e/ou vídeos utilizada, ou então utilizam vídeos particulares, gravados especificamente para a pesquisa em questão. Apenas 18% dos trabalhos selecionados apresentam as bases de imagens e/ou vídeos utilizados, as quais são apresentadas na Tabela 2. Nota-se que a maioria das bases é de imagens e/ou vídeos de expressões faciais e que nenhuma delas é brasileira.

Tabela 2 – Bases de imagens encontradas nos trabalhos selecionados.

Base	**Site**
American Sign Language Lexicon Video Dataset (ASLLVD)	www.bu.edu/asllrp/lexicon/index.html
AR Face	www.ece.ohio-state.edu/~aleix/ARdatabase.html
Binghamton University 3D Facial Expression (BU-3DFE)	http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html
Cohn-Kanade	www.consortium.ri.cmu.edu/ckagree
ECHO	www.let.ru.nl/sign-lang/echo
Greek Sign Language Corpus (GSLC)	www.dictasign.eu
Japanese Female Facial Expression (JAFFE)	www.kasrl.org/jaffe.html
Multi-PIE	www.multipie.org
Purdue RVL-SLLL American Sign Language	www.engineering.purdue.edu/RVL/Database/ASL/asl-database-front.htm
SignStream	www.bu.edu/asllrp/SignStream/
XM2VTS	www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

Nas subseções seguintes serão apresentadas com mais detalhes algumas características relevantes observadas nos trabalhos selecionados. A subseção 3.2.1 contém as principais técnicas de pré-processamento de imagem utilizadas nos trabalhos selecionados. Na subseção 3.2.2 são apresentadas as principais características utilizadas pelos trabalhos selecionados, enquanto que na subseção 3.2.3 é discutido sobre as principais técnicas de reconhecimento de sinais utilizadas. Por fim, a subseção 3.2.4 descreve sobre os principais trabalhos que tratam de reconhecimento de LIBRAS.

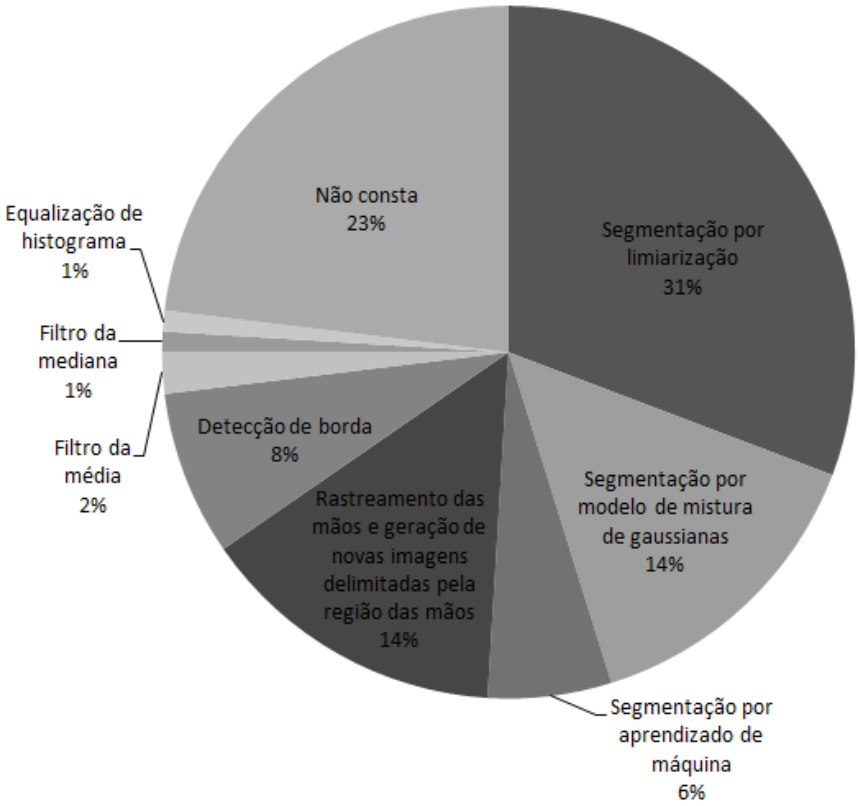
3.2.1 Técnicas de pré-processamento

As principais técnicas de pré-processamento de imagem utilizadas nos trabalhos selecionados são apresentadas no gráfico da Figura 16. Vários trabalhos não descrevem as técnicas de pré-processamento utilizadas, mas dos que descrevem, nota-se que a segmentação é a principal delas, presente na maioria dos trabalhos selecionados. A técnica de segmentação mais utilizada é a por limiarização ("*thresholding*") (RADHA; KRISHNAVENI, 2009; HABILI; LIM; MOINI, 2004; GOH; HOLDEN, 2006; WU; NAGAHASHI, 2013), seguida pela técnica que utiliza Modelo de Mistura de Gaussianas (RIBEIRO; GONZAGA, 2006; MUSHFIELDT; GHAZIASGAR; CONNAN, 2013; KIM; SHAKHNAROVICH; LIVESCU, 2013).

Alguns trabalhos optam ainda por utilizar algoritmos de aprendizado de máquina para realizar a segmentação. No trabalho de Han, Awad e Sutherland (2009) é utilizado *Support Vector Machine (SVM)* para segmentar a região de pele, classificando cada pixel das imagens como pele e não pele, com uma taxa de classificação média de 76,77% utilizando 240 *frames* da base de dados ECHO. Em Hieu e Nitsuwat (2008), *Two-Layer Neural Network (TLNN)* é utilizado para a construção de um modelo aproximado de pele, usando os componentes de cromaticidade Cb e Cr de 414.022 amostras de pixels, a fim de segmentar as imagens, atingindo uma taxa de acerto de 94%. Já em El-Jaber, Assaleh e Shanableh (2010) a segmentação é feita utilizando Mapa de Disparidade e o método *K-Means Clustering*, enquanto que em Gonçalves et al. (2012) a segmentação das imagens é feita utilizando *Multilayer Perceptron (MLP)* com o algoritmo de treinamento *backpropagation*, otimizado por meio do método *Levenberg-Marquardt*, mas as taxas de precisão desses dois últimos trabalhos não são apresentadas.

A técnica de detecção da região das mãos e geração de novas imagens delimitadas por essa região também é bastante utilizada nos trabalhos selecionados, que tem como objetivo facilitar o processo de reconhecimento (CHANDA; AUEPHANWIRIYAKUL; THEERA-UMPON, 2012; ROUSSOS et al., 2013). Os principais algoritmos utilizados para rastrear as mãos nas sequências de imagens são: *Mean-Shift* (KELLY et al., 2008; KELLY; MCDONALD; MARKHAM, 2009) e *Continuously Adaptive Mean-Shift (CamShift)* (MADANI; NAHVI, 2013; DIAS et al., 2004).

Figura 16 – Distribuição de artigos pelas principais técnicas de pré-processamento de imagem utilizadas.

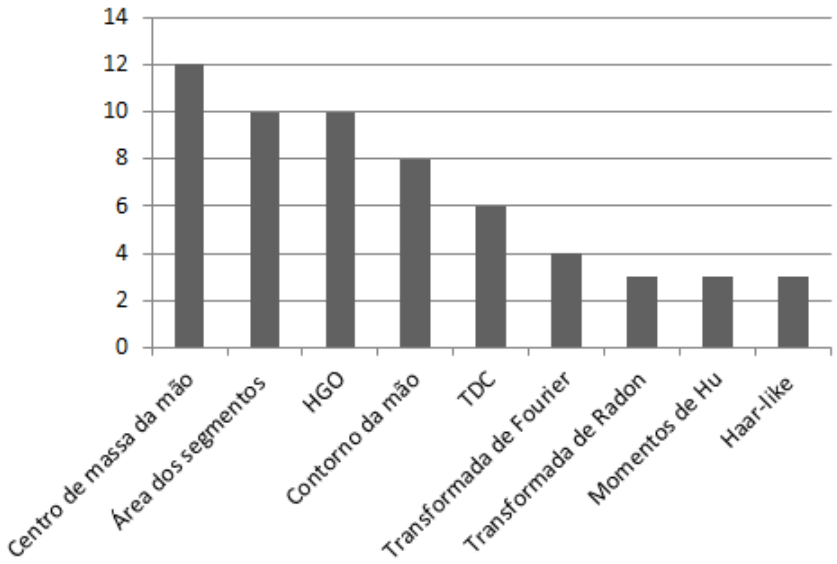


3.2.2 Características

Para realizar o reconhecimento dos parâmetros relacionados à realização de sinais em língua sinalizada e das letras e/ou palavras expressas em língua de sinais, é necessário conhecer algumas características das sequências de imagens. As principais características utilizadas pelos trabalhos selecionados são apresentadas pelo gráfico da Figura 17, sabendo-se que grande parte dos trabalhos selecionados não especificam as características utilizadas (40%).

Observa-se que o centro de massa da mão é bastante utilizado, principalmente para reconhecer a trajetória realizada pelas mãos durante a expressão dos sinais (movimento), como em Akmeliawati, Ooi e Kuang (2007), Hienz, Grobel e Offner (1996), Bauer e Hienz (2000), Madeo (2011), em que o cálculo do centro de massa das mãos é facilitado pelo uso de luvas coloridas. A área dos segmentos (mãos e/ou face) também é bastante utilizada, como em Hieu e Nitsuwat (2008), Zhang e Zhang (2010a), Grobel e Assan (1997), assim como Histogramas de Gradientes Orientados (HGO), utilizado em Kim, Shakhnarovich

Figura 17 – Distribuição de artigos por principais características utilizadas.



e Livescu (2013), Cooper, Pugeault e Bowden (2011), Buehler, Zisserman e Everingham (2009), Thangali e Sclaroff (2009). A ideia principal deste descritor é que a forma e a aparência de objetos em uma imagem podem ser descritos por meio da distribuição dos gradientes de intensidade dos pixels ou pelas direções das bordas (GRITTI et al., 2008).

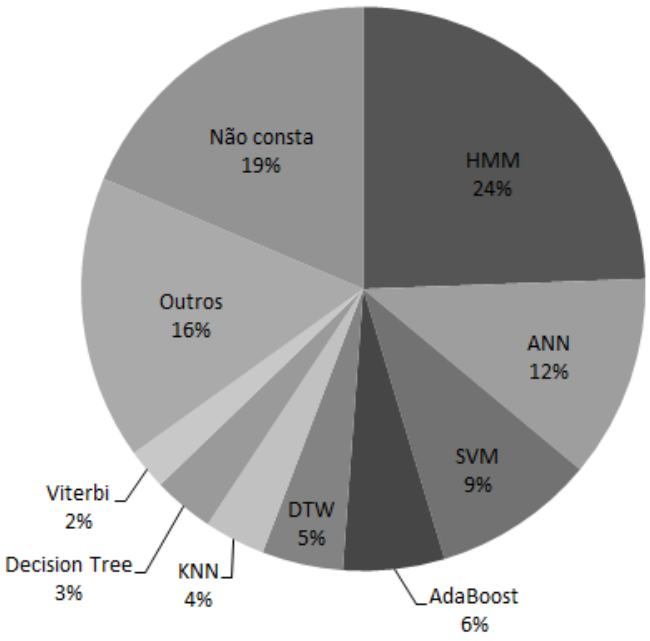
O contorno da mão (borda) é também uma característica bastante encontrada nos trabalhos selecionados. Para detectar a borda das mãos o principal algoritmo utilizado é o Sobel, presente em Isaacs e Foo (2004), Pistori e Neto (2004), Pavan e Modesto (2010). Alguns trabalhos representam o contorno das mãos por meio de assinaturas (representação da borda em uma função unidimensional), utilizadas em Futane e Dharaskar (2012), Gonçalves et al. (2012), Peres et al. (2006).

Também foram encontrados alguns trabalhos que optam por utilizar como características os coeficientes da Transformada Discreta do Cosseno (TDC), que converte um bloco de pixels em uma matriz de coeficientes, descorrelacionando a informação da imagem, utilizado por Paulraj et al. (2009), Assaleh et al. (2008), El-Jaber, Assaleh e Shanableh (2010), Paulraj et al. (2008), Shanableh e Assaleh (2007a), Shanableh e Assaleh (2007b), sendo que os dois últimos trabalhos também utilizam a Transformada de Radon, também presente em Madani e Nahvi (2013).

3.2.3 Técnicas de reconhecimento

Pelos artigos analisados, além das técnicas de processamento de imagens apresentadas anteriormente na subseção 3.2.1, nota-se que a maioria dos trabalhos utiliza também algoritmos e técnicas de aprendizado de máquina para o reconhecimento de língua de sinais, principalmente *Hidden Markov Model (HMM)*, seguido por *Artificial Neural Network (ANN)* e *Support Vector Machine (SVM)*. A distribuição dos artigos pelo tipo de técnica de aprendizado de máquina utilizada é apresentada pelo gráfico da Figura 18.

Figura 18 – Distribuição de artigos por algoritmos e técnicas de aprendizado de máquina.



A utilização de *Hidden Markov Model (HMM)* como técnica de reconhecimento de sinais é feita em diversos trabalhos, tais como: Assaleh et al. (2008), Sandjaja e Marcos (2009), Goh e Holden (2006), Starner e Pentland (1995). O trabalho de Assaleh et al. (2008) é um sistema para reconhecer sinais contínuos em Língua de Sinais Árabe por meio da análise do movimento das mãos, sem restrições, tais como o uso de luvas coloridas. Os resultados experimentais mostram uma taxa média de reconhecimento de palavra de 94% e de 75% para o reconhecimento de frases, utilizando 40 frases formadas por 80 palavras, gravadas 19 vezes cada por apenas um indivíduo em vídeos com resolução de 720 x 528 pixels. O trabalho proposto por Sandjaja e Marcos (2009) apresenta uma taxa de precisão média de 85,52%, utilizando *5-fold cross-validation*, para reconhecer números em Língua

Filipina de Sinais a partir de 500 vídeos com resolução de 640 x 480 pixels, gravados em um estúdio com luvas multicoloridas. Em Goh e Holden (2006) é apresentado um sistema para o reconhecimento de Língua de Sinais Australiana, que atinge 97,15% de precisão para reconhecer letras e 88,61% para palavras, utilizando configuração de mão, orientação e movimento como parâmetros e 40 palavras, sendo quatro sequências de vídeo independentes para cada palavra (160 sequências), gravadas em estúdio sem o uso de luvas coloridas, mas com iluminação e fundo controlados e restrições para a roupa do indivíduo que expressa os sinais. Uma boa taxa de 99,2% é encontrada em Starner e Pentland (1995), no reconhecimento de 40 palavras em Língua de Sinais Americana, gravadas com resolução de 320 x 243 pixels em um estúdio com luvas coloridas, utilizando configuração de mão, ponto de articulação e movimento como parâmetros.

HMM também são utilizadas por Davydov, Nikolski e Pasichnyk (2010), Bauer e Hienz (2000), Bauer, Hienz e Kraiss (2000). Em Davydov, Nikolski e Pasichnyk (2010) é alcançada uma taxa de reconhecimento de sinais de 91,7%, utilizando configuração de mão, ponto de articulação e movimento como parâmetros e um conjunto de teste composto por 85 sinais do dicionário de Língua de Sinais Ucraniana frequentemente utilizados, expressados por meio de 160 configurações de mão principais, determinadas por um profissional em língua de sinais, sem a utilização de luvas coloridas. A mesma taxa de acerto é atingida no trabalho de Bauer e Hienz (2000), utilizando configuração de mão, orientação e ponto de articulação como parâmetros e 97 sinais em Língua de Sinais Alemã expressados por apenas um indivíduo, mas gravado com o uso de luvas coloridas, fundo com cor uniforme e algumas restrições para a roupa do indivíduo. Em Bauer, Hienz e Kraiss (2000) é utilizado o mesmo conjunto de dados e as mesmas restrições que em Bauer e Hienz (2000), mas apresentam uma taxa um pouco maior, de 93,2%, por meio do uso de um modelo bigram, que é um modelo estatístico no qual é assumido que a ocorrência de um sinal depende da ocorrência do sinal anterior.

O uso de *Artificial Neural Network (ANN)* apresentou resultados promissores em reconhecimento de sinais estáticos. Em Isaacs e Foo (2004) é apresentado um sistema que além de usar ANN, utiliza um algoritmo genético para melhorar o pré-processamento realizado nas imagens, capaz de reconhecer 24 sinais estáticos do alfabeto da Língua de Sinais Americana, atingindo uma precisão de 99,9%, por meio de vídeos gravados em estúdio, sem o uso de luvas. ANNs também são utilizadas para reconhecer gestos em Língua Malaia de Sinais por Paulraj et al. (2009), Paulraj et al. (2008), Akmeliawati,

Ooi e Kuang (2007). O trabalho proposto por Paulraj et al. (2009) atinge uma taxa de reconhecimento de 81,07%, utilizando 14 gestos gravados com resolução de 320 x 240 pixels, realizados repetidamente 5 vezes por 2 indivíduos, sem o uso de luvas coloridas. Paulraj et al. (2008) apresentam uma taxa de classificação de 92,07%, utilizando 32 gestos diferentes gravados em estúdio com resolução de 320 x 240 pixels, sem o uso de luvas, mas com restrições para a roupa do indivíduo. Já o sistema proposto por Akmeliawati, Ooi e Kuang (2007) atinge taxas de 99,33%, 95,67% e 95% para reconhecer números, letras e palavras, respectivamente, utilizando 49 gestos estáticos e dinâmicos gravados em estúdio com luvas multicoloridas em vídeos com resolução de 352 x 288 pixels.

Alguns trabalhos optam por utilizar SVM para o reconhecimento de sinais, tais como: Quan (2010), Elakkiya, Selvamani e Kanimozhi (2014), Mushfieldt, Ghaziasgar e Connan (2013), Ari, Uyar e Akarun (2008). O primeiro trabalho apresenta uma taxa média de reconhecimento de 95,55%, utilizando 5.850 imagens com fundo branco de 30 letras em Língua de Sinais Chinesa (195 para cada letra). Em Elakkiya, Selvamani e Kanimozhi (2014) é apresentado um *framework* para segmentação e rastreamento de objetos de pele em vídeos de língua de sinais por meio de SVM e características *Haar-like*. Este trabalho atinge uma taxa de aproximadamente 90%, utilizando vídeos gravados em estúdio sem a utilização de luvas coloridas com resolução de 640 x 480 pixels. O trabalho de Mushfieldt, Ghaziasgar e Connan (2013) propõe uma abordagem para reconhecer expressões faciais em Língua de Sinais Sul Africana na presença de rotações e oclusões parciais da face. Os dados de treinamento utilizados consistem em sequências de vídeo de 10 indivíduos de gêneros mistos e tons de pele variados realizando 6 expressões diferentes, utilizando as bases de dados BU-3DFE, *Cohn-Kanade*, JAFFE e *Multi-Pie*. O sistema atinge uma taxa de reconhecimento de 85% para imagens faciais frontais e uma precisão de reconhecimento média de 80% para os rostos rotacionados em 60 graus. Ari, Uyar e Akarun (2008) também apresentam um sistema de reconhecimento automático de expressões faciais, que atinge uma taxa média de reconhecimento de 90%, utilizando um conjunto de dados que contém sete expressões faciais utilizadas em Língua de Sinais Turca, gravadas cinco vezes cada por 11 indivíduos diferentes (6 do sexo feminino, 5 do masculino).

O uso do algoritmo *AdaBoost* também foi encontrado com resultados promissores em alguns trabalhos selecionados, assim como em Han, Awad e Sutherland (2013), Wu e Nagahashi (2013). Em Wu e Nagahashi (2013) ele é utilizado num sistema para detecção e rastreamento das mãos em vídeos de língua de sinais. Os resultados experimentais

mostram que o método proposto obtém uma taxa média de detecção de mãos de 96,3%, e mais de 91,2% das mãos rastreadas são extraídas em tamanho adequado, utilizando 657 *frames* de 3 vídeos gravados em estúdio sem o uso de luvas coloridas. Já em Han, Awad e Sutherland (2013) ele é utilizado para reconhecer sinais dinâmicos em Língua de Sinais Britânica, atingindo uma taxa de reconhecimento de 97,6%, utilizando ponto de articulação, orientação e movimento como parâmetros, a partir de 20 sinais diferentes sinalizados 10 vezes cada por dois indivíduos, resultando em 200 amostras de vídeos com mais de 10.000 *frames*, gravados em estúdio sem o uso de luvas.

Yang, Ahuja e Tabb (2002) apresentam um sistema para extração de trajetórias de movimento e sua aplicação para reconhecimento de sinais utilizando *Time-Delay Neural Network (TDNN)*. Os experimentos demonstram que gestos podem ser reconhecidos usando trajetórias de movimento com uma precisão de 99,21%, utilizando *5-fold cross-validation*, a partir de vídeos de 40 gestos em Língua de Sinais Americana gravados em estúdio com resolução de 160 x 120 pixels e sem o uso de luvas coloridas, com uma média de 60 *frames* cada. Lichtenauer et al. (2007) também apresentam um sistema de reconhecimento de sinais com base no movimento, utilizando *Dynamic Time Warping (DTW)*. O sistema apresentado neste trabalho é capaz de detectar corretamente 95% dos sinais de teste utilizando *7-fold cross-validation*, a partir de um conjunto de 120 gestos dinâmicos em Língua de Sinais Holandesa realizados por 70 pessoas diferentes, gravados em estúdio por duas câmeras com resolução de 640 x 480 pixels, sem o uso de luvas coloridas. Já o trabalho de Zhang e Zhang (2010b) propõe um sistema diferenciado para realizar o reconhecimento de sinais dinâmicos. O sistema proposto converte as características extraídas dos vídeos para *strings* e utiliza distância de edição para realizar o reconhecimento dos sinais, atingindo uma taxa média de reconhecimento de 95,5%, utilizando ponto de articulação e movimento como parâmetros e uma base de dados construída a partir de um conjunto de vídeos em Língua de Sinais Chinesa extraídos de programas de televisão de Pequim.

Alguns outros trabalhos selecionados apresentam apenas sistemas específicos de segmentação ou reconhecimento de parâmetros relacionados à realização de sinais em língua sinalizada, tais como: Soontranon, Aramvith e Chalidabhongse (2004), Grobel e Hienz (1996), Dimov, Marinov e Zlateva (2007), Ong e Bowden (2004). O primeiro trabalho descreve um método de detecção do rosto e rastreamento da mão para um sistema de reconhecimento de língua de sinais, a partir de quatro vídeos em Língua de Sinais Tailandesa gravados em estúdio sem o uso de luvas, com resolução de 240 x 320 pixels.

A segmentação é feita pela função da elipse, encontrada por meio da distribuição dos componentes de cromaticidade Cb e Cr dos pixels da imagem, atingindo uma taxa média de reconhecimento de 89,8% para a detecção da região de pele (face e mãos). Após as regiões com tom de pele serem segmentadas, o sistema tem como objetivo detectar as características da face usando diferença de luminosidade e as das mãos usando o método esqueleto. O segundo trabalho apresenta um sistema para o reconhecimento de 32 configurações de mão diferentes utilizadas na Língua de Sinais Alemã. O modelo apresentado baseia-se nas características das áreas coloridas e sobre as relações entre essas áreas, atingindo uma taxa de reconhecimento de 94%, a partir de um classificador baseado em regras e utilizando vídeos com resolução de 768 x 512 pixels gravados em estúdio com luvas multicoloridas. Dimov, Marinov e Zlateva (2007) também propõem um sistema para reconhecimento de sinais estáticos (letras), mas de uma maneira mais intrusiva, em que os vídeos são gravados em estúdio, sem o uso de luvas, mas em um ambiente controlado onde apenas as mãos e a face aparecem. O reconhecimento é feito utilizando CBIR, atingindo também uma taxa de reconhecimento de aproximadamente 94%, a partir de 7 sinais/letras do alfabeto da Língua de Sinais Búlgara gravados em 344 imagens (49 imagens em média por letra). Já Ong e Bowden (2004) apresentam um sistema capaz de reconhecer as mãos em vídeos com uma taxa elevada de 99,8% e as configurações de mão com uma taxa de 97,4%, utilizando *Boosted Classifier Tree*. Para realizar os experimentos do reconhecimento da mão foram utilizadas 5.013 imagens em tons de cinza, extraídas de vários vídeos gravados em estúdio sem o uso de luvas por diferentes indivíduos (2.504 amostras para treinamento e 2.509 para teste) e mais 900 imagens para testar o reconhecimento das configurações de mão (300 grupos de formas diferentes).

Por fim, o movimento do gesto que une dois sinais consecutivos (movimento epêntese), um dos problemas difíceis enfrentados em reconhecimento automatizado de língua de sinais, é tratado no trabalho de Yang, Sarkar e Loeding (2007), que obteve 83% de taxa de reconhecimento utilizando o algoritmo *Level Building*, a partir de frases contínuas expressadas em sequências de imagens em Língua de Sinais Americana com resolução de 460 x 290, gravadas em estúdio sem o uso de luvas. O vocabulário utilizado de 39 sinais é composto pelos principais sinais que uma pessoa surda precisa para se comunicar com o pessoal de segurança nos aeroportos, expressados em 25 diferentes frases gravadas cinco vezes cada.

3.2.4 LIBRAS

Dos artigos selecionados, 11% propuseram sistemas de reconhecimento de sinais utilizando a Língua Brasileira de Sinais (LIBRAS). No trabalho de Madeo et al. (2010) *Fuzzy Learning Vector Quantization (FLQV)* não supervisionadas e supervisionadas são utilizadas para compor um comitê de máquinas, cujo objetivo é reconhecer imagens estáticas (configurações de mão) e movimentos. O comitê foi capaz de reconhecer 20 configurações de mão e cinco movimentos utilizados para representar as letras do alfabeto em LIBRAS, com precisão de 85% e 91,7%, respectivamente, por meio de vídeos gravados em um estúdio com luvas coloridas.

Em Digiampietri et al. (2012) é apresentado um sistema de informação para o reconhecimento automático de LIBRAS, também desenvolvido pela aluna deste projeto, fundamentado em dois pilares: um ambiente configurável e extensível para o gerenciamento de experimentos de processamento de línguas de sinais baseado no uso de *workflows* científicos e um conjunto de módulos desenvolvidos especificamente para o processamento de imagens e vídeos, composto por métodos para a segmentação e classificação de imagens. O sistema proposto atinge uma taxa de acerto de 87,67% para reconhecer as 26 letras do alfabeto, gravadas em estúdio com o uso de luvas multicoloridas, utilizando como características a área proporcional de cada segmento (dedos e palma da mão) e a posição relativa de cada segmento em relação ao centro de gravidade da mão.

Já o trabalho proposto por Souza, Dias e Pistori (2007) atinge uma taxa de reconhecimento de 80,1% utilizando HMM e 47 gestos extraídos do dicionário trilíngue de LIBRAS. Para os experimentos foram utilizados vídeos gravados por três indivíduos sem o uso de luvas coloridas, em um ambiente com fundo estático e uniforme, em que cada um executou sete vezes cada gesto. Com isso, o banco de imagens dedicado à experimentação possuía 21 amostras para cada gesto, totalizando 987 amostras.

Pistori e Neto (2004) e Gonçalves et al. (2012) apresentam sistemas para reconhecer configurações de mão utilizadas em LIBRAS. Em Pistori e Neto (2004) é apresentado um sistema que atinge uma taxa de precisão de 95,02%, utilizando o algoritmo *AdapTree* e um conjunto com 270 amostras de nove sinais alfabéticos diferentes, gravadas em estúdio sem o uso de luvas (66% da base de dados para treinamento e 34% para teste), enquanto que o trabalho de Gonçalves et al. (2012) atinge uma taxa de 97,6% utilizando uma técnica de *Content-Based Image Retrieval - CBIR* e 83 amostras gravadas em estúdio com

luvas coloridas. Por fim, em Madeo (2011) é apresentado um protótipo de uma aplicação educativa e inclusiva, que atinge 84,3% de precisão a partir de vídeos em LIBRAS gravados em estúdio com luvas coloridas, utilizando *Neuro-fuzzy* como técnica de reconhecimento e configuração de mão, orientação e movimento como parâmetros.

3.3 Conclusão

Notou-se a partir dos artigos analisados que o reconhecimento de sinais dinâmicos em língua de sinais ainda é pouco desenvolvido, assim como o reconhecimento dos movimentos dos gestos que une dois sinais consecutivos. Os principais resultados tratam apenas o reconhecimento de sinais estáticos ou o reconhecimento de sinais gravados com restrições, como a utilização de luvas coloridas e/ou ambientes controlados, objetivando facilitar o processo de reconhecimento.

Quanto aos trabalhos que tratam de LIBRAS, os melhores resultados encontrados também são apenas sobre reconhecimento de sinais estáticos ou de parâmetros (configurações de mão e movimento) ou então utilizam luvas coloridas e/ou ambientes controlados na gravação dos vídeos (a maioria).

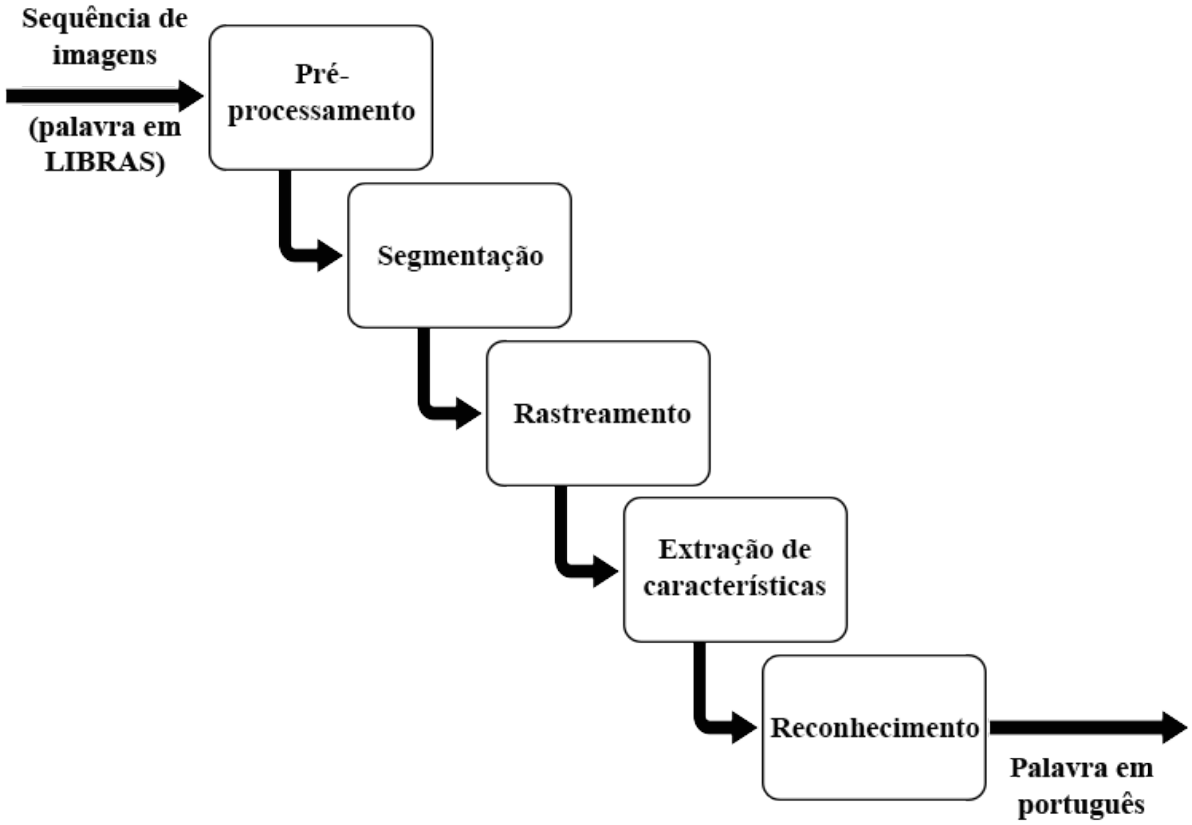
Notou-se também nos trabalhos analisados que existem poucas bases de imagens sobre língua de sinais. A maioria dos trabalhos constrói seu próprio conjunto de imagens para realizar os experimentos ou então não apresenta a base de imagens utilizada. Mas das poucas bases de imagens encontradas, nenhuma é brasileira e a maioria é de expressões faciais.

Outro ponto que chamou a atenção é a ausência de sistemas completos, que utilizam todos os cinco parâmetros relacionados à realização de sinais. Nenhum trabalho foi encontrado com esta característica, a maioria dos trabalhos selecionados foca apenas em dois ou três parâmetros para realizar o reconhecimento de sinais. Também notou-se poucos trabalhos que estudam expressões não manuais em reconhecimento de sinais. Os trabalhos encontrados tratam apenas o reconhecimento de expressões faciais especificamente, mas não a utilizam como um parâmetro, ou seja, não utilizam a influência destas expressões para realizar o reconhecimento de sinais (ARI; UYAR; AKARUN, 2008; MUSHFIELDT; GHAZIASGAR; CONNAN, 2013). Este parâmetro em particular é importante ao se pensar em reconhecimento de sinais e, até o momento, foi pouco trabalhado.

4 Sistema de reconhecimento

Neste capítulo serão apresentados o banco de imagens construído e as etapas principais do sistema proposto, correspondendo cada uma a um módulo do mesmo, de forma a elucidar ao leitor o funcionamento do sistema como um todo. Essas etapas também são apresentadas pelo diagrama de fluxo de execução do sistema da Figura 19.

Figura 19 – Diagrama de fluxo de execução do sistema.



Fonte: Elaborada pela autora.

4.1 Banco de Imagens

A fim de desenvolver e testar o sistema de reconhecimento de LIBRAS implementado, foi construído um banco de imagens, composto por 20 sequências de imagens (vídeos) gravadas por 10 indivíduos (quatro homens e seis mulheres), sendo dois vídeos gravados com cada indivíduo.

Os vídeos foram gravados em ambientes não controlados, sem nenhuma restrição. Os indivíduos sinalizaram um conjunto de 30 palavras básicas escolhidas por uma especialista em LIBRAS, são elas: casa, sol, luz, pessoa, cachorro, homem, mulher, aluno, bom, dia, tarde, noite, mãe, pai, filho, tio, sobrinho, cunhado, primo, sogro, irmão, nascer, morrer, viver, cozinha, banheiro, televisão, computador, mesa e faca. Destaca-se que dentre essas palavras, há algumas sinalizadas de maneira parecida. Observando os exemplos de sequências de imagens para cada palavra do banco no apêndice A, é possível perceber uma semelhança considerável entre as palavras "mulher" e "mãe", "pai" e "homem", "sol" e "dia".

As palavras escolhidas foram sinalizadas uma vez em cada vídeo. Desta forma, cada indivíduo sinalizou o mesmo conjunto de palavras em dois vídeos diferentes, gravados em dias e fundos diferentes e utilizando roupas distintas, a fim de proporcionar uma maior variabilidade. Vale destacar ainda, que a maioria dos vídeos foi gravada em locais com fundos não estáticos.

As gravações dos vídeos foram feitas utilizando uma câmera digital da marca Sony, modelo DSC-W530 de 14,1 megapixels, operando no modo de configuração automático. Cada vídeo foi editado (dividido) em 30 vídeos menores, de forma a separar a sinalização de cada uma das 30 palavras em cada um dos vídeos, com duração de aproximadamente dois segundos cada um, com 29 *frames* por segundo e resolução de 640 x 480 pixels cada *frame*.

A Figura 20 apresenta alguns exemplos de *frames* extraídos dos vídeos gravados, que formam o banco de imagens.

Vale ressaltar que esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa Envolvendo Seres Humanos da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP), conforme o parecer presente no anexo A deste documento. Os sujeitos aptos à pesquisa foram convidados a participarem por meio do Termo de Consentimento Livre e Esclarecido (TCLE), detalhado no apêndice B.

Durante o desenvolvimento deste projeto foram procurados bancos de imagens públicos, que contivessem sequências de imagens (vídeos) relacionadas a LIBRAS, mas nenhum banco público foi encontrado e por isso a necessidade da produção de um banco de imagens nesta dissertação. A falta de bancos/bases de dados de referência para línguas de sinais ou mesmo de ferramentas que facilitem a criação e disponibilização destes bancos

Figura 20 – Exemplos de imagens do banco de imagens.

Fonte: Elaborada pela autora.

tem motivado trabalhos recentes para o oferecimento deste tipo de recurso (WAGNER et al., 2012).

4.2 Pré-processamento

O objetivo principal desta primeira etapa do sistema é de pré-processar as sequências de imagens segmentando a região dos objetos em movimento (neste caso, os braços e as mãos), a fim de facilitar o processo de identificação e rastreamento das mãos realizado posteriormente. Primeiramente, é aplicada uma técnica de equalização de histograma nas sequências de imagens, apresentada na subseção 4.2.1, com o objetivo de melhorar a sua qualidade. Em seguida, a segmentação da região das mãos é realizada utilizando o algoritmo de subtração de fundo descrito na subseção 4.2.2. Após a aplicação deste algoritmo, de forma a melhorar o resultado obtido, é aplicada a técnica de fechamento, apresentada na subseção 4.2.3. Por fim, a fim de minimizar os ruídos, é aplicado o filtro da mediana, descrito na subseção 4.2.4.

4.2.1 Equalização de histograma

Analisando as sequências de imagens do banco, percebeu-se a necessidade de compatibilizar algumas distorções que ocorreram durante a etapa de aquisição dessas

imagens, principalmente por parte de diferenças de iluminação de ambiente, que poderiam comprometer as próximas etapas do sistema.

Desta forma, a fim de realçar o contraste das sequências de imagens, primeiramente foi utilizada a técnica de equalização de histograma, apresentada anteriormente na subseção 2.3.2. Existem vários métodos empregados para a realização da equalização de histograma. Neste trabalho foi utilizada a técnica de equalização adaptativa de histograma com limitação de contraste (*CLAHE - Contrast Limited Adaptive Histogram Equalization*).

Nesta técnica, em vez de ser calculado um histograma global, é calculado um histograma local para a vizinhança de cada pixel. Além disso, esta técnica corta no histograma local todas as intensidades em que o número de pixels esteja acima de um limite pré-definido. As intensidades que estiverem acima do limite são redistribuídas no histograma pelos tons vizinhos.

Para implementar a técnica CLAHE, foi utilizada uma função já pronta do pacote de aplicativos "ij.jar" do programa ImageJ (ZUIDERVELD, 1994). A função tem três parâmetros:

- *block size*: tamanho da região local em torno do pixel para a qual o histograma é equalizado;
- *histogram bins*: número de barras do histograma utilizado para a equalização de histograma;
- *max slope*: limita o estiramento do contraste na função de transferência de intensidade.

Para este trabalho, os parâmetros foram definidos com os valores "63", "256" e "3", respectivamente.

A Figura 21 apresenta o resultado obtido pela aplicação da técnica CLAHE em algumas imagens do banco de imagens.

4.2.2 Subtração de fundo

Conforme visto anteriormente na subseção 2.3.3, a subtração de fundo é uma técnica utilizada para separar objetos ou partes de objetos de interesse do restante da imagem. No contexto deste trabalho, a subtração de fundo foi aplicada para separar a área da região das mãos dos elementos estáticos do vídeo, resultando numa segunda sequência

Figura 21 – Resultado obtido pela técnica CLAHE: (a) *frames* originais; (b) *frames* com histograma equalizado.

Fonte: Elaborada pela autora.

de imagens, contendo apenas a área da região das mãos, que seria então submetida ao processo de rastreamento.

Foi utilizada uma função da biblioteca OpenCV para implementar a subtração de fundo, a função *"BackgroundSubtractorMOG2"*, em que é construído internamente um modelo adaptativo de mistura de gaussianas (*MoG - Mixture of Gaussian*) para subtração de fundo com detecção de sombras, baseado em Zivkovic (2004), Zivkovic e Heijden (2006). Nesse método, cada pixel é modelado como uma MoG e, em cada iteração, é calculada a probabilidade do pixel pertencer ao plano de fundo.

O modelo da função *"BackgroundSubtractorMOG2"* já está implementado de forma otimizada. Desta forma, o código para este algoritmo chama a seguinte função ajustando os seguintes parâmetros:

BackgroundSubtractorMOG2 (int history, float varThreshold, boolean bShadowDetection)

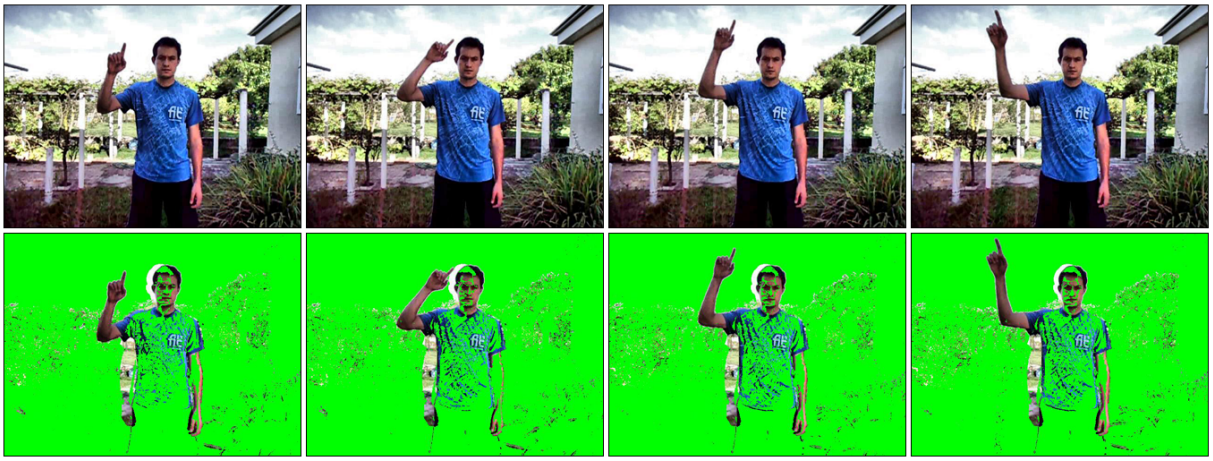
- *history*: quantidade de *frames* usados no histórico. O padrão é zero. Se diferente de zero, especifica o número de imagens anteriores para agregar como plano de fundo para comparação. Um valor diferente de zero realmente só importa para vídeos que precisam calcular um fundo em constante mudança a partir de *frames* mais recentes do vídeo;
- *varThreshold*: limiar do quadrado da distância de *Mahalanobis* entre o pixel e o modelo para decidir se o pixel é bem descrito pelo modelo de fundo. Este parâmetro não afeta a atualização de fundo. Um valor típico poderia ser de σ 4, (*varThreshold* = 4 * 4 = 16). Os valores mais elevados reduzem a sensibilidade e, portanto, o ruído;
- *bShadowDetection*: define se a detecção de sombra deve ser ativada (*true* ou *false*).

Os parâmetros foram definidos para este trabalho como "0", "32" e "*false*", respectivamente.

Uma característica importante deste algoritmo é que é selecionado um número apropriado de distribuição gaussiana para cada pixel, diferente da função *"BackgroundSubtractorMOG"*, em que é utilizado um K fixo de distribuições gaussianas em todo o algoritmo, fornecendo, desta forma, uma melhor adaptabilidade a diferentes cenas, por exemplo, devido a mudanças de iluminação.

A Figura 22 contém um exemplo de resultado gerado após a aplicação do método de subtração de fundo proposto em uma sequência do banco de imagens com histograma equalizado. Analisando os *frames* gerados, é possível perceber que a imagem resultante produz uma imagem do objeto de interesse (neste caso, as mãos do indivíduo) com pequenos buracos e/ou pedaços destruídos, que são os objetos principais que serão rastreados e utilizados no reconhecimento dos sinais, além também de ruídos no fundo da cena, devido a este não ser totalmente estático, podendo desta forma, prejudicar as próximas etapas do sistema. A fim de remover estes ruídos e melhorar o resultado obtido, também foi utilizada a técnica de processamento de imagens denominada fechamento, apresentada em seguida na subseção 4.2.3.

Figura 22 – Resultado obtido pelo método de subtração de fundo: (a) *frames* com histograma equalizado; (b) *frames* com fundo removido.



Fonte: Elaborada pela autora.

4.2.3 Fechamento

Analisando os resultados obtidos pela técnica de subtração de fundo definida anteriormente na subseção 4.2.2, é possível perceber alguns ruídos e defeitos nas imagens geradas, principalmente devido ao fato dos fundos dos ambientes em que foram gravados os vídeos não serem estáticos. Para corrigir esses problemas, foi utilizada a técnica de fechamento.

Conforme apresentado na subseção 2.3.4, a técnica de fechamento é utilizada para reparar imagens, suavizando contornos, unindo quebras estreitas e golfos longos e delgados e removendo os pixels ruidosos do interior do objeto. O fechamento é obtido a partir do encadeamento do filtro morfológico de dilatação, seguido pelo de erosão.

Neste trabalho, os filtros de dilatação e erosão foram implementados utilizando a biblioteca de funções OpenCV, ajustando as seguintes funções e parâmetros:

cvDilate(CvArr src, CvArr dst, IplConvKernel element, int iterations)

cvErode(CvArr src, CvArr dst, IplConvKernel element, int iterations)

- *src*: imagem de origem;
- *dst*: imagem de destino;
- *element*: elemento utilizado para a erosão ou dilatação. Se for igual a *null*, um elemento estruturante retangular 3 x 3 é utilizado;

- *iterations*: número de vezes que a erosão ou dilatação é aplicada.

O filtro de dilatação foi aplicado três vezes e o de erosão 15 vezes, com o valor do parâmetro *element* igual a *null*.

A Figura 23 apresenta um exemplo de resultado gerado com a aplicação do filtro de dilatação seguido pelo de erosão em uma sequência do banco de imagens com fundo removido pelo método apresentado anteriormente na subseção 4.2.2.

Figura 23 – Resultado obtido pela técnica de fechamento: (a) *frames* com histograma equalizado; (b) *frames* após a subtração de fundo; (c) *frames* após a aplicação do filtro de dilatação; (d) *frames* após a aplicação do filtro de erosão.

Fonte: Elaborada pela autora.

4.2.4 Filtro da mediana

Com o objetivo de suavizar os ruídos, mas preservando as bordas e detalhes finos das imagens geradas após a aplicação do processo de fechamento, foi utilizado o filtro da mediana, apresentado anteriormente com mais detalhes na subseção 2.3.5.

O filtro da mediana foi aplicado por meio da função *cvSmooth* da biblioteca OpenCV, utilizando o parâmetro CV_MEDIAN e janela 3x3. A Figura 24 apresenta um exemplo de resultado obtido pela aplicação deste filtro em uma parte de uma imagem gerada pela técnica de fechamento.

Figura 24 – Resultado obtido pelo filtro da mediana: (a) parte de um *frame* obtido pela técnica de fechamento; (b) resultado da filtragem pelo filtro da mediana com máscara 3x3.

Fonte: Elaborada pela autora.

4.3 Segmentação

A etapa de segmentação consiste em identificar as regiões de pele humana nas imagens, com o objetivo de facilitar a detecção das mãos nas sequências de imagens para realizar o rastreamento, além de preparar as imagens para a fase de extração de características. Para isso, foram implementadas as principais técnicas utilizadas nos trabalhos encontrados na revisão bibliográfica que tratam de segmentação de pele. Essas técnicas foram testadas e a que obteve melhor resultado foi utilizada neste trabalho.

Primeiramente, foram testados diferentes algoritmos de aprendizado de máquina presentes no pacote de software Weka (HALL et al., 2009). Com base nos resultados destes testes foi feita a escolha do algoritmo implementado, que tem como objetivo receber as

cores de um dado pixel e retornar a classe desse pixel, classificando-o como “pele” ou “não pele” (neste trabalho, vamos chamar de “fundo”).

Para o teste dos classificadores do Weka foi construído um conjunto de dados com 4.000 instâncias (pixels). Cada instância é composta por quatro atributos, as cores *RGB* (Vermelho [*Red*], Verde [*Green*] e Azul [*Blue*]) e a luminância (Y), calculada por meio da Equação 1, a fim de lidar com as variações de iluminação da pele humana; e, além destes atributos, um rótulo com a classe (“pele” ou “fundo”). Para a construção do conjunto de dados foram extraídos aleatoriamente 4.000 pixels (2.000 classificados como “pele” e 2.000 como “fundo”) de 200 *frames* retirados dos 20 vídeos gravados para a construção do banco de imagens (10 *frames* de cada vídeo), segmentados manualmente por um único indivíduo (a autora desta dissertação) e validados por outro (o orientador).

Os algoritmos de classificação foram executados utilizando a técnica de validação *10-fold cross-validation*, descrita na subseção 2.7.1. Primeiramente, os classificadores foram testados com um conjunto de dados que não possuía o atributo Y e posteriormente utilizando o atributo Y, a fim de avaliar o impacto que este atributo teria na classificação dos pixels. Para avaliação foram calculados os valores médios de acurácia, sensibilidade e especificidade, detalhados anteriormente na seção 2.6. Espera-se obter um alto valor para a sensibilidade e para a especificidade, para que o classificador identifique corretamente os pixels que são “pele” e aqueles que são “fundo”, além de uma acurácia mais próxima de 100%. Os 10 melhores resultados obtidos em cada um dos testes são apresentados em ordem decrescente nas tabelas 3 e 4, respectivamente.

Tabela 3 – Os 10 melhores resultados obtidos pelos classificadores do Weka (sem a utilização do atributo Y).

Classificador	**Acurácia**	**Sensibilidade**	**Especificidade**
functions.MultilayerPerceptron	97,80%	98,55%	97,05%
trees.FT	97,80%	98,95%	96,65%
trees.RandomForest	97,80%	98,70%	96,90%
functions.Logistic	97,73%	99,10%	96,35%
meta.MultiClassClassifier	97,73%	99,10%	96,35%
functions.SimpleLogistic	97,70%	99,35%	96,05%
trees.LMT	97,70%	99,35%	96,05%
rules.NNge	97,60%	98,40%	96,80%
meta.END	97,58%	98,55%	96,60%
meta.RotationForest	97,53%	98,90%	96,15%

Tabela 4 – Os 10 melhores resultados obtidos pelos classificadores do Weka (com a utilização do atributo Y).

Classificador	**Acurácia**	**Sensibilidade**	**Especificidade**
meta.RotationForest	98,25%	99,30%	97,20%
functions.Logistic	98,23%	99,20%	97,25%
meta.MultiClassClassifier	98,23%	99,20%	97,25%
functions.MultilayerPerceptron	98,20%	99,25%	97,15%
meta.ThresholdSelector	98,20%	99,25%	97,15%
functions.SimpleLogistic	98,18%	99,55%	96,80%
trees.LMT	98,17%	99,55%	96,80%
meta.Bagging	98,15%	99,30%	97,00%
trees.RandomForest	98,12%	98,90%	97,35%
rules.NNge	98,08%	98,50%	97,65%

A partir dos dados apresentados pelas tabelas 3 e 4, é possível perceber a melhora dos resultados com a utilização da luminância (Y) como atributo além das cores RGB. Também é possível observar que o classificador que obteve melhor resultado foi o *RotationForest*, sendo o classificador escolhido para ser utilizado.

Pelo fato do Weka possuir uma API de fácil utilização, o algoritmo selecionado (o *RotationForest*) não foi reimplementado, em vez disso, ele é executado pelo Weka. Foi desenvolvida uma ferramenta que converte a floresta resultante da execução do *RotationForest* em um sistema especialista que recebe as cores RGB e a luminância (Y) de um dado pixel e retorna a classe desse pixel ("pele" ou "fundo"). O segmentador então chama este sistema especialista para classificar cada pixel da imagem. Esse segmentador também será chamado como *RotationForest* neste trabalho, a fim de simplificar a sua identificação.

Além da ferramenta que utiliza o algoritmo *RotationForest* para realizar a classificação dos pixels, também foram implementados mais quatro algoritmos simples de segmentação de pele humana encontrados na literatura, que recebe as cores RGB de um dado pixel e o classifica como "pele" ou "fundo" com base em conjuntos de regras.

O primeiro algoritmo, proposto por Kovac, Peer e Solina (2003), é um conjunto de quatro regras:

1. $R > 95 \quad \text{e} \quad G > 40 \quad \text{e} \quad B > 20$.
2. $max(R, G, B) - min(R, G, B) > 15$.
3. $|R - G| > 15$.
4. $R > G \quad \text{e} \quad R > B$.

Se as quatro regras forem verdadeiras, o pixel é classificado como "pele", caso contrário, é classificado como "fundo". Para facilitar a sua identificação, neste trabalho chamaremos esse segmentador de *Kovac*.

O segundo algoritmo é uma regra muito simples proposta por Al-Shehri (2004), que considera apenas o valor de R e G. Esta regra classifica o pixel como "pele" quando $R - G$ é maior que 20 e menor que 80 ($20 < R - G < 80$), caso contrário, classifica como "fundo". Neste trabalho, esse segmentador será identificado como *Al-Shehri*.

No terceiro algoritmo, proposto por Osman, Hitam e Ismail (2012), duas regras são verificadas para realizar a classificação do pixel:

1. $0.0 \leq \frac{R-G}{R+G} \leq 0.5$.
2. $\frac{B}{R+G} \leq 0.5$.

Se as duas regras forem verdadeiras, o pixel é classificado como "pele", caso contrário, é classificado como "fundo". Identificaremos esse segmentador como *Osman* neste trabalho.

Por fim, o quarto algoritmo, proposto por Swift (2005), é composto pelas seguintes regras:

1. $B > R$.
2. $G < B$.
3. $G > R$.
4. $B < (\frac{1}{4})R$.
5. $B > 200$.

Se pelo menos uma das cinco regras for verdadeira, o pixel é classificado como "fundo", caso contrário, é classificado como "pele". Neste trabalho, esse segmentador será identificado como *Swift*.

Diversos testes foram realizados com os métodos de segmentação implementados, a fim de escolher o melhor para ser utilizado neste trabalho, os quais serão detalhados posteriormente na seção 5.1.

4.4 Rastreamento

Esta etapa do sistema tem como objetivo isolar apenas as regiões de pontos da imagem pertencentes a mão do indivíduo para a extração de características. Para atingir

este objetivo, foi implementado um algoritmo para rastrear as mãos do indivíduo nas sequências de imagens e gerar novas imagens delimitadas pela região em que elas se encontram, que dispensa o auxílio de dispositivos ou marcadores, deixando o corpo do indivíduo livre.

O rastreador de mãos implementado compara o conteúdo dos pixels de uma região de referência (da mão de uma imagem de referência) com os pixels das próximas imagens do vídeo e, para cada imagem subsequente a imagem de referência, encontra a posição na qual há menor diferença entre os pixels de referência e os pixels da imagem atual. Neste cálculo de menor distância é acrescentada uma penalidade proporcional à distância euclidiana entre a possível posição da mão na imagem atual e a posição da mão na imagem anterior. A função de rastreamento possui os seguintes parâmetros:

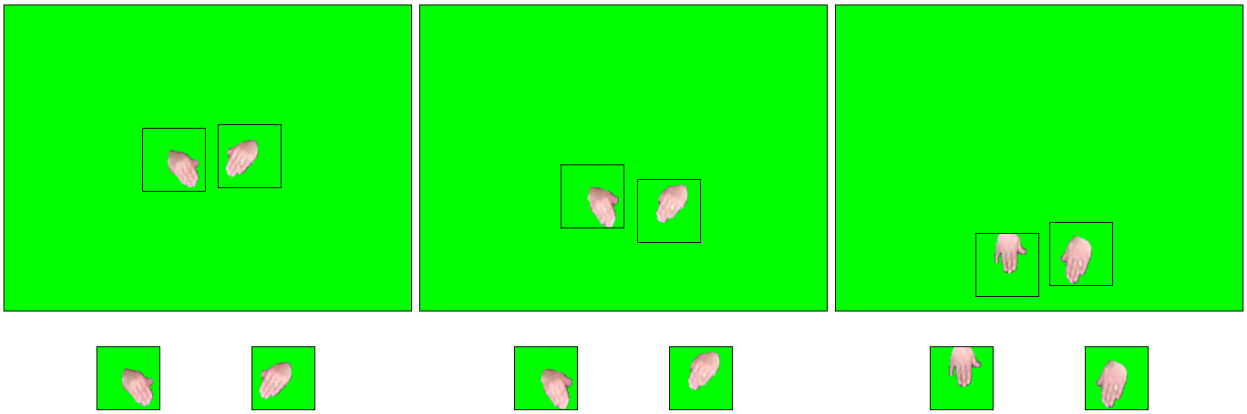
rastreamento(Image[] video, int frameReferencia, int larguraRegiao, Point posicaoDaMao, int janelaDeBusca)

- *video*: arranjo de imagens correspondendo ao vídeo de entrada;
- *frameReferencia*: posição da imagem de referência no arranjo de imagens;
- *larguraRegiao*: tamanho da largura (e altura, pois a região será quadrada) da região utilizada para comparar os valores dos pixels;
- *posicaoDaMao*: posição x e y da mão na imagem de referência.
- *janelaDeBusca*: janela de busca na qual será procurada pela mão nas imagens subsequentes a imagem de referência.

O rastreamento das mãos foi utilizado neste trabalho com os seguintes valores para seus parâmetros: *larguraRegiao = 5*, isto é, foi utilizado um quadrado de 25 pixels (5x5). A posição inicial da mão foi identificada manualmente. A *janelaDeBusca* utilizada foi de 100 pixels (isto é, a mão rastreada poderia estar até 100 pixels deslocada no eixo x e/ou no eixo y em relação à mão da imagem anterior). O resultado do rastreador é uma tabela contendo para cada imagem após a imagem de referência, a posição identificada como a mão (o centro da região mais parecida, segundo os critérios utilizados, em relação à região identificada como mão na imagem de referência).

A Figura 25 apresenta um exemplo de resultado gerado com a aplicação do método de rastreamento de mãos implementado em uma sequência de imagens pré-processadas e segmentadas.

Figura 25 – Exemplo de resultado gerado pelo rastreador.



Fonte: Elaborada pela autora.

Durante a execução do rastreador, ocorreram alguns casos em que o rastreador se perdeu algumas vezes (em aproximadamente 70% das sequências de imagens de cada filmagem do banco). Nestes casos, o rastreador foi corrigido manualmente, reiniciando a sua execução no *frame* em que o erro ocorreu.

4.5 Extração de características

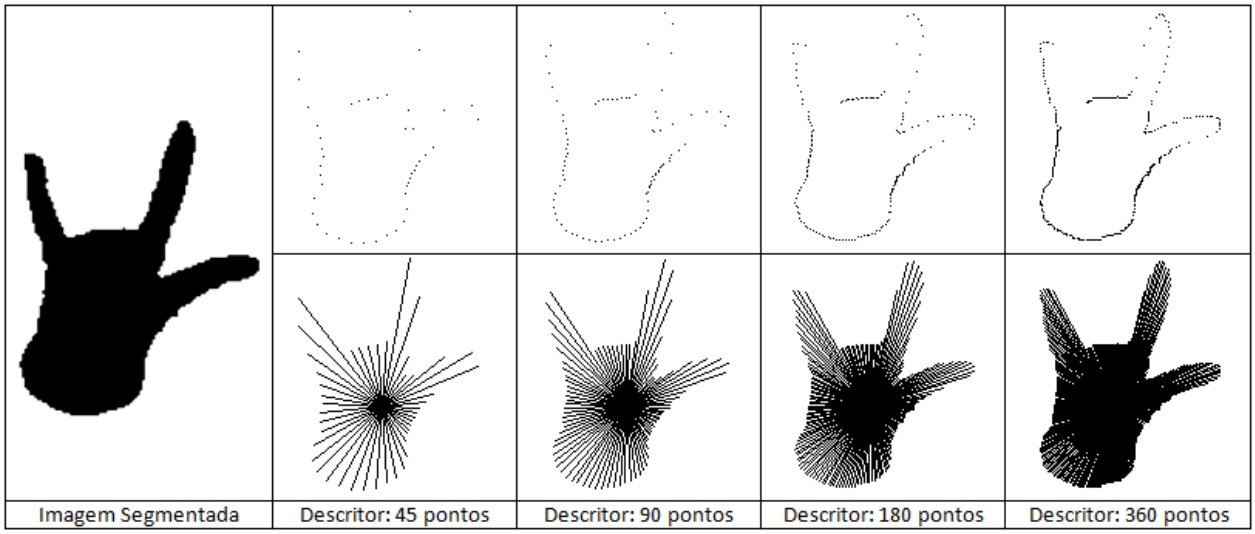
A etapa de extração de características consiste em capturar propriedades relevantes das sequências de imagens, a fim de fornecer informações para realizar o reconhecimento do sinal executado.

Com base nos trabalhos revisados, neste projeto foram extraídas as seguintes características das sequências de imagens: forma das mãos, frequência de pixels, deslocamento das mãos, distância entre as mãos e distância entre as mãos e a face; totalizando nove vetores de características.

A forma das mãos do indivíduo foi representada utilizando um extrator desenvolvido em um projeto anterior (DIGIAMPIETRI et al., 2012). Este extrator representa a mão por meio dos raios angulares dos contornos externos das mãos. Para calcular esses raios, primeiramente, dada uma sequência de imagens gerada pelo rastreador, cada imagem é convertida para preto e branco (os pixels com cor de pele pertencentes a mão são convertidos para a cor preta, enquanto que os pixels restantes pertencentes ao fundo são convertidos para a cor branca). Em seguida, são detectados o centro de massa da imagem

e o contorno externo da mão presente na imagem. A partir do centro de massa, são então calculados os tamanhos dos x raios que saem deste centro e atingem o contorno externo da imagem. Por exemplo, se $x = 36$ então, a cada 10 graus, será calculado o valor do raio entre o centro de massa da imagem e seu contorno externo. Por fim, esses raios são normalizados para valerem de 0 a 1. Este reduz as informações de uma imagem de, por exemplo, 307.200 pixels (imagem de 480 x 640 pontos) para x pontos. Além disso, esse extrator consegue recuperar imagens de maneira invariante ao seu tamanho à rotação (ao se utilizar uma correta medida de distância na comparação entre os vetores de x características) (DIGIAMPIETRI et al., 2012). A Figura 26 apresenta exemplos da aplicação deste extrator para diferentes valores de x. Nesta dissertação o valor utilizado para x foi 180.

Figura 26 – Exemplos da aplicação do extrator de forma utilizado.



Fonte: (DIGIAMPIETRI et al., 2012)

A característica de frequência de pixels consiste nas porcentagens de pixels pertencentes a cor de pele e ao fundo em cada imagem da sequência de imagens gerada pelo rastreador.

Para a extração das demais características (deslocamentos e distâncias), durante a aplicação do rastreador foi obtido o centro de massa das mãos em cada imagem da sequência de imagens. A partir dos centros de massas foram mensurados os deslocamentos das mãos, as distâncias entre as mãos e as distâncias entre as mãos e a face por meio da distância euclidiana entre os centros de massas.

Considerando os pontos $A = (a_1, a_2, ..., a_n)$ e $B = (b_1, b_2, ..., b_n)$, a distância euclidiana entre esses pontos em um espaço n-dimensional é calculada a partir da equação 11.

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{11}$$

Para a obtenção do centro de massa da face, antes de aplicar as etapas de pré-processamento, segmentação e rastreamento anteriormente definidas, foi utilizada uma implementação do método proposto por Viola e Jones (2001) para detecção de faces, presente na biblioteca OpenCV. Este método foi escolhido por ser capaz de detectar faces com precisão, alta taxa de acerto, baixa taxa de falsos positivos e baixo custo computacional, obtendo neste trabalho 100% de sucesso para detectar as faces dos indivíduos em todas as sequências de imagens do banco.

A Figura 27 apresenta alguns exemplos de resultados obtidos com a aplicação deste método. Foi considerado como centro de massa o centro do quadrado que delimita a face.

Figura 27 – Exemplos de resultados gerados pelo método de detecção de face utilizado.

Fonte: Elaborada pela autora.

Após terem sido extraídos, todos os vetores de características foram normalizados. A normalização tem como objetivo ajustar as escalas de valores dos dados para o mesmo intervalo, minimizando os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis. Neste trabalho, os valores dos vetores de características foram ajustados para uma faixa entre 0 e 1, utilizando o método de normalização pelo valor máximo dos elementos, que consiste em dividir cada valor pelo maior valor. Desta forma, cada valor de cada vetor de característica foi dividido pelo máximo valor encontrado em seu vetor.

4.6 Reconhecimento

Para o reconhecimento e classificação das palavras expressas em LIBRAS nas sequências de imagens, foi utilizada a técnica de distância de edição, apresentada anteriormente na subseção 2.5.3.

Primeiramente, é calculada a distância de edição entre os nove vetores de características da palavra sinalizada com cada uma das demais amostras de palavras do banco. Em seguida, é utilizado um classificador binário que irá verificar se há um possível casamento entre a palavra sinalizada e cada palavra do banco, com base nas nove distâncias calculadas. Por fim, verifica-se qual palavra do banco recebeu mais votos pelo classificador (isto é, teve mais instâncias para as quais o classificador indicou um possível casamento) e esta palavra será então apresentada como o resultado do reconhecimento.

Neste trabalho foi utilizado como classificador o método *Random Forest* proposto por Breiman (2001), a partir do pacote de software Weka. Este método consiste da criação de uma série de árvores de decisão construídas durante o seu treinamento. Cada uma dessas árvores de decisão define a classe à qual pertencerá um determinado objeto. O objeto então pertencerá à classe que foi retornada como resposta pela maioria das árvores pertencentes à floresta. Este método vem sendo utilizado em diversas áreas do aprendizado de máquina, apresentando excelentes resultados.

Para validar a solução proposta, foi produzido um conjunto de dados da seguinte forma. Para cada uma das sequências de imagens (palavras) extraídas de cada um dos vídeos, foram calculadas as distâncias de edição das suas características com as características das demais amostras de palavras restantes. O arquivo resultante é composto por 13 colunas, com os seguintes atributos: **id**, **palavra sinalizada**, **palavra comparada**, **distância de edição dos deslocamentos da mão direita**, **distância de edição dos deslocamentos da mão esquerda**, **distância de edição das distâncias entre a mão direita e a face**, **distância de edição das distâncias entre a mão esquerda e a face**, **distância de edição das distâncias entre as mãos**, **distância de edição das frequências de pixels da mão direita**, **distância de edição das frequências de pixels da mão esquerda**, **distância de edição dos raios angulares do contorno externo da mão direita**, **distância de edição dos raios angulares do contorno externo da mão esquerda** e **classe**.

O **id** é constituído pelo número da palavra sinalizada (que varia de 01 a 30) mais o número do vídeo em que ela foi sinalizada (que varia de 01 a 20). Por exemplo, o **id** da palavra 01 do vídeo 12 é igual a 0112. O atributo **palavra sinalizada** é o nome da palavra sinalizada ("casa", por exemplo), enquanto que **palavra comparada** é o nome da palavra com a qual a palavra sinalizada está sendo comparada. As demais colunas, com exceção da última, são as distâncias de edição entre os vetores de características da palavra sinalizada e da palavra comparada. A última coluna (**classe**) contém a classificação da instância ("*true*", se a palavra sinalizada for igual a palavra comparada, "*false*", caso contrário).

Para a construção dos conjuntos foram utilizadas quatro diferentes valores para o parâmetro "penalidade", recebido pelo método que calcula a distância de edição: 0,05, 0,1, 0,2 e 0,3. Dos 20 vídeos gravados para a construção do banco de imagens, os de número 8 e 9 não foram utilizados, pois obtiveram péssimos resultados após a etapa de segmentação, não sendo possível a execução das demais etapas do sistema, conforme será apresentado adiante na seção 5.1.

Com os conjuntos de dados gerados, utilizou-se a validação cruzada em dois subconjuntos (*2-fold cross-validation*) para verificar a solução do sistema. Os subconjuntos foram gerados da seguinte forma: o primeiro continha metade dos vídeos gravados e o segundo continha a outra metade. Optou-se por utilizar este número reduzido de subconjuntos por se entender que a utilização de outra estratégia como utilizar todos os vídeos menos um para treinamento (*leave-one-out*) poderia melhorar de maneira artificial os resultados dos testes.

Não fez parte do escopo desta dissertação identificar o melhor valor para a "penalidade" nos cálculos de distância, porém foram testados e avaliados alguns diferentes valores que, conforme será apresentado no próximo capítulo, tiveram influência nos resultados do classificador, porém o resultado final do reconhecimento com os quatro valores utilizados foi o mesmo e por isso será apresentado apenas o resultado de reconhecimento usando-se "penalidade" 0,2.

Os resultados dos testes realizados com os conjuntos de dados construídos serão detalhados adiante na seção 5.2.

5 Experimentos, resultados e discussão

Neste capítulo serão apresentados e discutidos os experimentos realizados, destacando os principais fatores que interferiram no resultado final do sistema, ou seja, no reconhecimento dos sinais. A seção 5.1 contém os experimentos realizados com os métodos de segmentação implementados, enquanto que a seção 5.2, apresenta os resultados obtidos no reconhecimento dos sinais.

5.1 Segmentação

Com o objetivo de escolher o segmentador utilizado para segmentar as sequências de imagens pré-processadas, foram realizados alguns testes com os cinco métodos de segmentação selecionados e detalhados anteriormente na seção 4.3.

Primeiramente, a fim de verificar o efeito da etapa de pré-processamento na segmentação das imagens, os cinco segmentadores foram testados utilizando um conjunto de 200 *frames* originais extraídos dos 20 vídeos gravados para a construção do banco de imagens (10 *frames* cada vídeo), sem serem pré-processados. Para avaliar os resultados dos testes, foram usadas as medidas padrão de acurácia, sensibilidade e especificidade, além do índice *overlap*. Os resultados obtidos em cada um dos testes são apresentados na tabela 5.

Tabela 5 – Resultados obtidos pelos segmentadores implementados utilizando *frames* originais sem pré-processamento.

Segmentador	**Acurácia**	**Sensibilidade**	**Especificidade**	**Overlap**
Kovac	96,92%	51,04%	97,91%	0,26
Al-Shehri	95,62%	54,33%	96,54%	0,25
Osman	70,67%	79,85%	70,42%	0,06
Swift	78,81%	57,74%	79,16%	0,05
RotationForest	48,11%	95,17%	47,12%	0,04

Em seguida, os cinco segmentadores foram testados novamente, utilizando os mesmos 200 *frames* utilizados anteriormente, mas após terem passado pela etapa de pré-processamento. Os resultados obtidos são apresentados na tabela 6.

Analisando os resultados apresentados nas tabelas 5 e 6, percebeu-se que com a aplicação da etapa de pré-processamento nas sequências de imagens os resultados dos segmentadores melhoraram consideravelmente, sendo esta uma etapa indispensável para a execução do sistema proposto.

Tabela 6 – Resultados obtidos pelos segmentadores implementados utilizando *frames* pré-processados.

Segmentador	**Acurácia**	**Sensibilidade**	**Especificidade**	**Overlap**
Kovac	98,86%	70,50%	99,49%	0,59
Al-Shehri	98,62%	70,20%	99,27%	0,54
Osman	98,07%	80,86%	98,42%	0,48
Swift	98,04%	54,05%	98,92%	0,34
RotationForest	96,07%	94,84%	96,09%	0,36

A Figura 28 apresenta exemplos de resultados gerados com a aplicação dos cinco métodos de segmentação implementados em imagens pré-processadas do banco de imagens.

Apesar do algoritmo *RotationForest* ter apresentado o melhor resultado para classificar os pixels, comparado com o restante dos algoritmos do Weka, conforme apresentado anteriormente na seção 4.3, analisando os resultados na tabela 6 e as imagens geradas pelos segmentadores, apresentadas na Figura 28, percebeu-se que o melhor resultado obtido para segmentar os *frames* do banco de imagens construído, foi o gerado pelo segmentador *Kovac*, chegando mais próximo do resultado esperado (segmentado manualmente). Desta forma, o segmentador escolhido para ser utilizado neste trabalho foi o *Kovac*.

Analisando visualmente os resultados gerados com a aplicação do método *Kovac* em todas as 20 filmagens feitas para a construção do banco de imagens, percebeu-se que as sequências de imagens resultantes das filmagens de número 8 e 9 apresentaram alguns problemas, que impediram a execução do rastreamento das mãos e, consequentemente, o reconhecimento dos sinais. A filmagem 8, por ter sido feita em ambiente fechado com pouca iluminação e com um indivíduo utilizando roupas complexas com cores próximas ao tom de pele, obteve sequências de imagens com muitos ruídos após a segmentação, prejudicando a identificação das mãos. Já a filmagem 9, devido a sua execução ter sido realizada em ambiente aberto e bastante iluminado, com a aplicação do segmentador obteve-se sequências de imagens com a maior parte dos pixels pertencentes aos segmentos de pele classificados como fundo, impedindo também a identificação das mãos. Desta forma, como não foi possível executar as outras etapas do sistema, posteriores ao da segmentação, as filmagens 8 e 9 foram excluídas dos testes finais do sistema.

Além das filmagens 8 e 9, observou-se que grande parte das sequências de imagens resultantes das filmagens de número 2, 16 e 19 também apresentaram alguns problemas após a aplicação do segmentador, que não impediram a execução das etapas seguintes do sistema, mas poderiam influenciar negativamente no resultado final. Com base nos

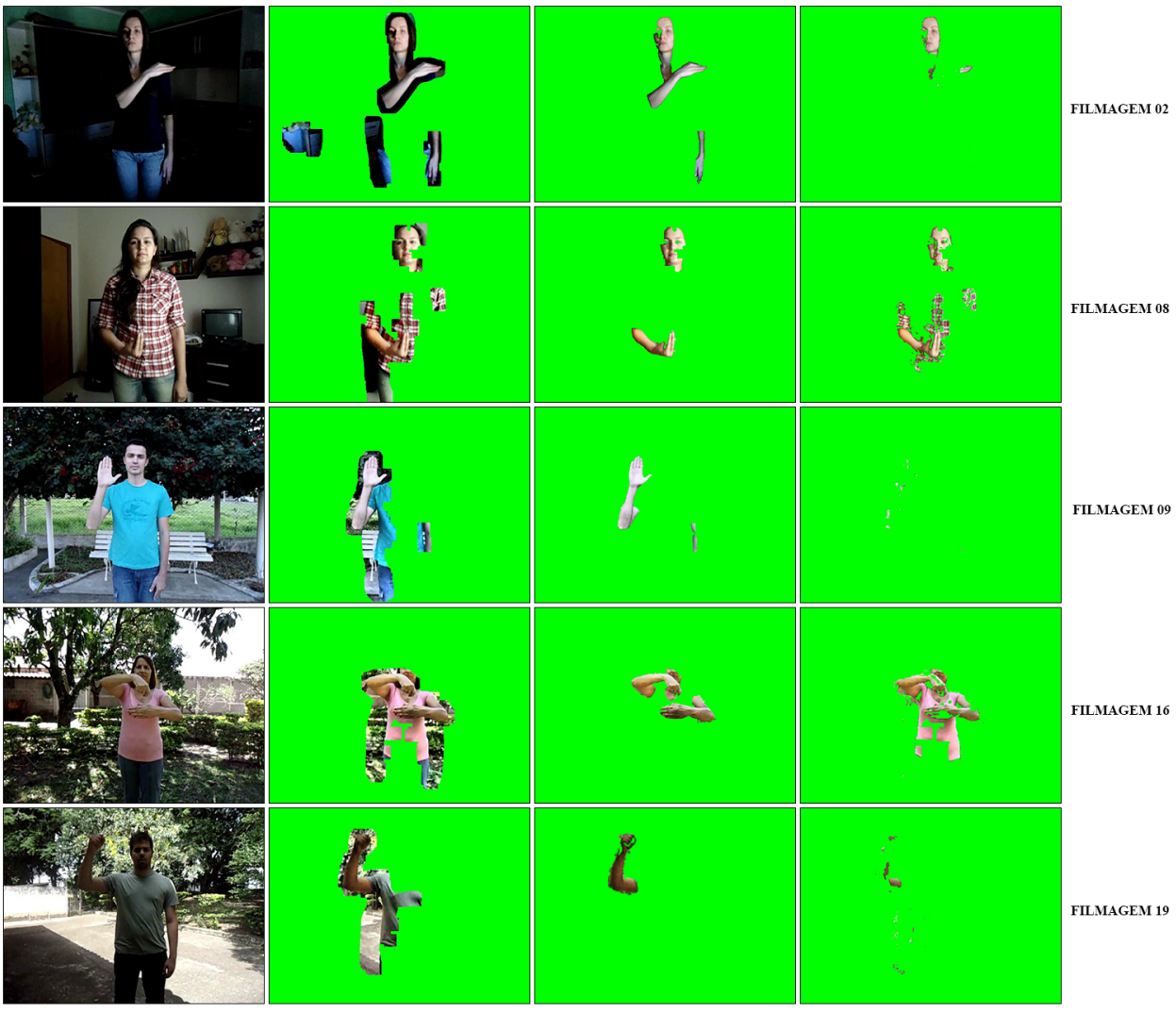
Figura 28 – Exemplos de resultados gerados pelos segmentadores implementados: (a) *frames* pré-processados; (b) *frames* esperados (segmentação manual); (c) *Kovac*; (d) *Al-Shehri*; (e) *Osman*; (f) *Swift*; (g) *RotationForest*.

Fonte: Elaborada pela autora.

resultados dos testes realizados com os segmentadores, infere-se que estes problemas de segmentação também foram obtidos devido a influência das roupas que os indivíduos utilizaram durante a filmagem e do ambiente em que foram filmados os vídeos.

A Figura 29 apresenta exemplos de resultados gerados com a aplicação do método *Kovac* em imagens extraídas das filmagens 2, 8, 9, 16 e 19 do banco de imagens.

Figura 29 – Exemplos de resultados gerados com a aplicação do método *Kovac* em imagens das filmagens 2, 8, 9, 16 e 19: (a) *frames* originais; (b) *frames* pré-processados; (c) *frames* esperados (segmentação manual); (d) *frames* segmentados pelo método *Kovac*.



Fonte: Elaborada pela autora.

A fim de verificar o efeito que as filmagens 8 e 9 causaram nos resultados obtidos pelos segmentadores, os cinco segmentadores foram testados sem estas filmagens, utilizando

um conjunto de 180 *frames* pré-processados extraídos dos 18 vídeos restantes (10 *frames* cada vídeo). Os segmentadores também foram testados sem as filmagens 2, 16 e 19, além das filmagens 8 e 9, utilizando um conjunto de 150 *frames* pré-processados extraídos dos 15 vídeos restantes (10 *frames* cada vídeo). Os resultados obtidos em cada um dos testes são apresentados nas tabelas 7 e 8, respectivamente.

Tabela 7 – Resultados obtidos pelos segmentadores implementados sem as filmagens 8 e 9, utilizando *frames* pré-processados.

Segmentador	**Acurácia**	**Sensibilidade**	**Especificidade**	**Overlap**
Kovac	99,02%	72,64%	99,56%	0,61
Al-Shehri	98,82%	72,87%	99,36%	0,57
Osman	98,28%	83,02%	98,56%	0,50
Swift	98,22%	55,23%	99,02%	0,35
RotationForest	96,15%	96,41%	96,13%	0,36

Tabela 8 – Resultados obtidos pelos segmentadores implementados sem as filmagens 2, 8, 9, 16 e 19, utilizando *frames* pré-processados.

Segmentador	**Acurácia**	**Sensibilidade**	**Especificidade**	**Overlap**
Kovac	99,20%	75,25%	99,67%	0,65
Al-Shehri	98,96%	74,91%	99,43%	0,59
Osman	98,52%	80,64%	98,84%	0,52
Swift	98,34%	50,20%	99,21%	0,34
RotationForest	96,35%	95,88%	96,34%	0,36

Com a exclusão das filmagens 8 e 9, percebeu-se que o melhor resultado obtido continuou sendo o do método *Kovac*, que aumentou sua taxa média de acurácia de 98,86% para 99,02% e, seu índice *overlap*, de 0,59 para 0,61. Com a exclusão das filmagens 2, 16 e 19, além as filmagens 8 e 9, a taxa de acurácia média e o índice *overlap* aumentaram ainda mais, para 99,20% e 0,65, respectivamente. Desta forma, acredita-se que as filmagens 2, 16 e 19 possam interferir negativamente no resultado final do sistema.

Mesmo não utilizando o mesmo conjunto de imagens, percebe-se que a taxa de acerto atingida pelo segmentador implementado neste trabalho é superior às taxas apresentadas nos trabalhos encontrados na revisão sistemática, que tratam de segmentação de pele humana. Por exemplo, no trabalho de Han, Awad e Sutherland (2009), com a utilização do classificador SVM para segmentar a região de pele, classificando cada pixel das imagens como pele e não pele, é atingido uma taxa de classificação média de apenas 76,77%, a partir de um conjunto com 240 *frames* da base de dados ECHO[1]. Já em Soontranon,

[1] www.let.ru.nl/sign-lang/echo

Aramvith e Chalidabhongse (2004), a segmentação é realizada por meio da função da elipse, encontrada a partir da distribuição dos componentes de cromaticidade Cb e Cr dos pixels da imagem, atingindo uma taxa média de reconhecimento de 89,8% para a detecção da região de pele (face e mãos), utilizando quatro vídeos em Língua de Sinais Tailandesa gravados em estúdio sem o uso de luvas, com resolução de 240 x 320 pixels.

5.2 Reconhecimento

Primeiramente, foi testado o classificador *Random Forest* utilizando os conjuntos de dados detalhados anteriormente na seção 4.6, que foram construídos variando o atributo "penalidade" com os valores 0,05, 0,1, 0,2 e 0,3, totalizando quatro conjuntos. Para avaliar o classificador, foi utilizada a técnica de validação *2-fold cross-validation* e foram calculados os valores médios de acurácia, sensibilidade e especificidade, apresentados na Tabela 9. Analisando esses resultados, percebeu-se que o conjunto gerado utilizando o atributo "penalidade" com valor igual a 0,2 foi o que obteve os melhores resultados.

Tabela 9 – Resultados obtidos pelo classificador *Random Forest*.

Penalidade	**Acurácia**	**Sensibilidade**	**Especificidade**
0,05	98,11%	44,57 %	99,86 %
0,1	98,15%	49,21 %	99,76 %
0,2	98,17%	49,81 %	99,76 %
0,3	98,16%	49,57 %	99,75 %

O reconhecimento das palavras foi testado utilizando os quatro valores diferentes para o atributo "penalidade" e, apesar desses valores terem influenciado nos resultados do classificador, o resultado final obtido para o reconhecimento das palavras foi o mesmo para os quatro valores. Desta forma, será apresentado apenas o resultado obtido utilizando "penalidade" igual a 0,2.

Com a exclusão das filmagens 8 e 9, dos 18 vídeos utilizados para testar o reconhecimento das palavras, foram extraídas 422 amostras de palavras (sequência de imagens). Vale ressaltar que há apenas uma amostra de cada uma das 30 palavras escolhidas em cada um dos vídeos, totalizando 540 amostras. Mas destas 540 amostras, 118 apresentaram problemas durante a segmentação, que prejudicaram a etapa de rastreamento. Desta forma, para testar o reconhecimento das palavras foram utilizadas 422 amostras, com uma média de aproximadamente 14 amostras por palavra. As filmagens que tiveram mais

amostras excluídas devido ao péssimo resultado durante a segmentação foram as de número 2 e 19, com 19 e 15 amostras excluídas cada, respectivamente. Conforme apresentado anteriormente na seção 5.1, acredita-se que os problemas de segmentação obtidos nessas filmagens foram por influência das roupas que os indivíduos utilizaram durante a filmagem e do ambiente em que os vídeos foram filmados. As filmagens 5, 13, 17 e 20 não tiveram nenhuma amostra excluída e as filmagens 4, 11 e 14 tiveram apenas uma amostra excluída cada uma. O restante teve em média nove amostras excluídas cada. Desta forma, das 600 amostras originalmente filmadas, 60 foram excluídas devido à exclusão dos vídeos 8 e 9 e mais 118 foram excluídas dos demais vídeos. Conforme apresentado, esta exclusão foi feita devido à incapacidade do sistema em segmentar corretamente as mãos destas amostras (em especial, devido a problemas nas etapas de rastreamento e segmentação). Assim, considerando-se as 600 amostras pode-se considerar que o sistema foi capaz de segmentar minimamente bem 422 (70,3%) ou se foram considerados apenas os 18 vídeos, o sistema foi capaz de segmentar 78,1% das amostras (422 de 540).

O sistema reconheceu corretamente todas as 422 amostras, atingindo 100% de acerto. A Tabela 10 contém uma parte dos resultados obtidos durante o reconhecimento (30 amostras). A tabela completa com todas as 422 amostras encontra-se no apêndice C.

Houve quatro casos, das 422 amostras, em que a porcentagem de votos corretos indicado pelo classificador foi de 50%, por exemplo, a amostra de número 87 da Tabela 10, em que o classificador retornou quatro classificações iguais a "*true*", destas classificações, duas foram atribuídas corretamente e duas incorretamente. Mas os 50% de votos incorretos não foram todos para uma mesma palavra em nenhum dos casos, portanto, não houve dúvida no reconhecimento, lembrando que a palavra apresentada como resultado do reconhecimento é a que obteve mais votos corretos pelo classificador.

Apesar da pressuposição de que as sequências de imagens das filmagens 2, 16 e 19 geradas após as etapas de segmentação e rastreamento poderiam prejudicar o resultado final do sistema, todas as amostras de palavras utilizadas e extraídas dessas sequências foram reconhecidas corretamente.

Pelo fato de não terem sido encontrados outros trabalhos que tratem de reconhecimento de palavras em LIBRAS na revisão sistemática, uma breve comparação com as técnicas melhor sucedidas no reconhecimento de línguas de sinais é apresentada a seguir. A principal limitação da técnica proposta neste trabalho é quanto a capacidade de segmentação sem a imposição de restrições. Como um fator adicional, destaca-se que os

Tabela 10 – Parte dos resultados obtidos para o reconhecimento das palavras.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
74	aluno	aluno	7	7	100%
75	banheiro	banheiro	5	5	100%
76	bom	bom	7	7	100%
77	cachorro	cachorro	7	10	70%
78	casa	casa	8	9	88,89%
79	computador	computador	8	9	88,89%
80	cozinha	cozinha	8	8	100%
81	cunhado	cunhado	6	7	85,71%
82	dia	dia	7	7	100%
83	faca	faca	5	6	83,33%
84	filho	filho	7	7	100%
85	homem	homem	10	11	90,91%
86	irmao	irmao	5	7	71,43%
87	luz	luz	2	4	50%
88	mae	mae	8	8	100%
89	mesa	mesa	5	5	100%
90	morrer	morrer	8	8	100%
91	mulher	mulher	6	6	100%
92	nascer	nascer	8	8	100%
93	noite	noite	8	8	100%
94	pai	pai	7	9	77,78%
95	pessoa	pessoa	6	7	85,71%
96	primo	primo	8	8	100%
97	sobrinho	sobrinho	6	8	75%
98	sogro	sogro	6	7	85,71%
99	sol	sol	7	11	63,64%
100	tarde	tarde	6	7	85,71%
101	televisao	televisao	9	10	90%
102	tio	tio	7	7	100%
103	viver	viver	8	9	88,89%

vídeos foram gravados por mais de um indivíduo, característica importante que não foi avaliada por todos os trabalhos correlatos.

Ao se considerar apenas as amostras que puderam ser segmentadas, nota-se que a técnica utilizada para realizar o reconhecimento de palavras neste trabalho, atingiu uma taxa de acerto superior as encontradas nos trabalhos revisados que tratam de reconhecimento de palavras em outras língua de sinais. Por exemplo, Starner e Pentland (1995) apresentam uma taxa de 99,2% de acerto no reconhecimento de 40 palavras em Língua de Sinais Americana, utilizando HMM e configuração de mão, ponto de articulação e movimento como parâmetros, mas foram gravadas com resolução de 320 x 243 pixels em

um estúdio com luvas coloridas, que facilitam bastante a segmentação e o reconhecimento das mãos. Goh e Holden (2006) apresentam um sistema para o reconhecimento de Língua de Sinais Australiana que utiliza HMM e configuração de mão, orientação e movimento como parâmetros, atingindo 88,61% de precisão para reconhecer um conjunto de 40 palavras, sendo quatro sequências de vídeo independentes para cada palavra (160 sequências), gravadas em estúdio sem o uso de luvas coloridas, mas com iluminação e fundo controlados e restrições para a roupa do indivíduo que expressa os sinais. O sistema desenvolvido por Assaleh et al. (2008) para reconhecer sinais contínuos em Língua de Sinais Árabe utilizando também HMM, por meio da análise do movimento das mãos, sem restrições, tais como o uso de luvas coloridas, atinge uma taxa média de reconhecimento de palavra de 94%, a partir de 40 frases formadas por 80 palavras, gravadas 19 vezes cada por apenas um indivíduo, em vídeos com resolução de 720 x 528 pixels.

6 Conclusões

Neste capítulo serão destacadas as principais contribuições deste trabalho e apresentadas algumas possibilidades de trabalhos futuros.

6.1 Contribuições

O principal objetivo e contribuição deste trabalho foi o desenvolvimento de um sistema para o reconhecimento automático de LIBRAS, capaz de analisar vídeos (sequências de imagens) de pessoas se comunicando em LIBRAS, reconhecer os sinais expressos nesses vídeos e apresentar a palavra correspondente em língua portuguesa, sem a utilização de dispositivos eletromecânicos, o uso de luvas coloridas ou a exigência de gravações de alta qualidade em laboratórios com ambientes controlados. Pretende-se facilitar desta forma, a comunicação entre surdos e/ou deficientes auditivos e pessoas que não conhecem uma língua gestual, proporcionando uma maior inclusão destes com o restante da sociedade.

Para atingir este objetivo, primeiramente foi realizada uma revisão sistemática acerca dos trabalhos feitos sobre a área temática deste trabalho. Com base na revisão, foram escolhidas e implementadas algumas técnicas de processamento de imagens que constituíram a etapa de pré-processamento do sistema. Conforme constatado com os experimentos realizados, o pré-processamento das imagens foi fundamental para obter uma boa acurácia no resultado do segmentador implementado para identificar as regiões de pele humana nas imagens e, consequentemente, nos resultados finais do sistema. Com a execução da etapa de pré-processamento, a acurácia do segmentador aumentou de 96,92% para 98,86% e o índice *overlap* mais que dobrou, aumentando de 0,26 para 0,59. Também foi construído um banco de imagens para ser utilizado no desenvolvimento e nos testes do sistema. Após a execução da etapa de segmentação, percebeu-se que duas das 20 filmagens gravadas para a construção do banco obtiveram péssimos resultados, influenciados pelas roupas que os indivíduos utilizaram e pelo ambiente em que as filmagens foram realizadas. Com a exclusão dessas duas filmagens, a taxa média de acurácia e o índice *overlap* do segmentador aumentaram para 99,02% e 0,61, respectivamente. Desta forma, evidenciou-se que a segmentação de pele humana em sequências de imagens gravadas em ambientes não controlados, sem nenhuma restrição, é uma tarefa bastante complexa e sensível à alguns fatores, como a roupa do indivíduo, a iluminação e o fundo em que as sequências

de imagens foram gravadas. Acredita-se que, por esses motivos, muitos dos trabalhos encontrados na literatura, que tratam de reconhecimento de sinais utilizando a abordagem visual, utilizam luvas coloridas e/ou ambientes controlados na gravação dos vídeos.

Durante o desenvolvimento desta dissertação foram encontradas algumas dificuldades, como a indisponibilidade de bancos de imagens públicos que contivessem sequências de imagens (vídeos) relacionadas a LIBRAS, necessitando assim, a construção de um banco de imagens para ser utilizado no desenvolvimento deste trabalho. Além disso, não foram encontrados trabalhos que tratassem exatamente do reconhecimento de sinais dinâmicos em LIBRAS utilizando a abordagem visual sem restrições, a fim de realizar uma comparação entre os resultados. Dos trabalhos que tratam de LIBRAS encontrados durante a revisão sistemática, os que apresentaram os melhores resultados são apenas sobre reconhecimento de sinais estáticos ou de parâmetros (configurações de mão e movimento), ou então, utilizam luvas coloridas e/ou ambientes controlados na gravação dos vídeos.

Apesar dos obstáculos encontrados, o objetivo proposto neste trabalho foi alcançado com êxito, pois mais de 70% das amostras de palavras sinalizadas em diferentes ambientes, por diferentes pessoas e sem a imposição de restrições puderam ser rastreadas e segmentadas. Para estas palavras (422 amostras de palavras do banco de imagens construído), o reconhecimento atingiu 100% de acerto.

Acredita-se que todas as etapas do sistema proposto são contribuições para trabalhos futuros da área de reconhecimento de sinais, além de poderem contribuir para outros tipos de trabalhos que envolvam processamento de imagens, segmentação de pele humana, rastreamento de objetos, entre outros. Além disso, como contribuição adicional, espera-se que o banco de imagens relacionadas a LIBRAS construído neste trabalho também possa ser utilizado em trabalho futuros.

6.2 Trabalhos futuros

Destaca-se que, devido aos desafios encontrados no reconhecimento de línguas de sinais, não esperava-se que o sistema produzido fosse uma solução definitiva para o reconhecimento automático de LIBRAS, mas sim uma ferramenta para auxiliar neste processo, que poderá ser aperfeiçoada em trabalhos futuros.

Os principais pontos a serem melhorados no sistema são os relacionados ao rastreador e ao segmentador. Como trabalho futuro, sugere-se testar outros rastreadores de objetos

e/ou melhorar o rastreador utilizado neste trabalho. Além disso, os testes realizados no sistema proposto utilizaram apenas as sequências de imagens gravadas para a construção do banco de imagens deste trabalho. Sendo assim, o uso de outros bancos de imagens e/ou aumentar o número de sequências de imagens do banco construído com a gravação de novos vídeos são desejáveis. Também recomenda-se expandir o conjunto de palavras, assim como a quantidade de indivíduos (voluntários) utilizados na gravação dos vídeos.

Também sugere-se como trabalho futuro explorar o uso de técnicas de processamento de imagens específicas para cada vídeo, de acordo com as suas características (resolução, ambiente em que foi filmado, iluminação, etc.). Sugere-se ainda que seja realizado um estudo da técnica de reconhecimento proposta nesse sistema utilizando vídeos gravados com algumas restrições, assim como são utilizados em muitos trabalhos da literatura, a fim de realizar uma comparação dos resultados.

O sistema proposto realiza o reconhecimento dos sinais utilizando apenas os parâmetros configuração de mão, ponto de articulação, movimento e orientação para realizar o reconhecimento dos sinais. Desta forma, outra possível evolução deste trabalho, seria incorporar expressões não manuais, um parâmetro importante ao se pensar em reconhecimento de sinais e, até o momento, foi pouco tratado nos trabalhos de reconhecimento de sinais encontrados na literatura.

Referências[1]

AKMELIAWATI, R.; OOI, M.; KUANG, Y. C. Real-time malaysian sign language translation using colour segmentation and neural network. In: *Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE*. [S.l.: s.n.], 2007. p. 1–6. ISSN 1091-5281. Citado 3 vezes nas páginas 46, 49 e 50.

AL-SHEHRI, S. A. A simple and novel method for skin detection and face locating and tracking. In: *Computer Human Interaction, 6th Asia Pacific Conference, APCHI 2004, Rotorua, New Zealand, June 29 - July 2, 2004, Proceedings*. [s.n.], 2004. p. 1–8. Disponível em: <http://dx.doi.org/10.1007/978-3-540-27795-8_1>. Citado na página 66.

ALBRES, N. A. *Surdos & Inclusão Educacional*. [S.l.]: Editora Arara Azul, 2010. Citado na página 15.

ARI, I.; UYAR, A.; AKARUN, L. Facial feature tracking and expression recognition for sign language. In: *Computer and Information Sciences, 2008. ISCIS '08. 23rd International Symposium on*. [S.l.: s.n.], 2008. p. 1–6. Citado 2 vezes nas páginas 50 e 54.

ASSALEH, K. et al. Vision-based system for sontinuous arabic sign language recognition in user dependent mode. In: *Mechatronics and Its Applications, 2008. ISMA 2008. 5th International Symposium on*. [S.l.: s.n.], 2008. p. 1–5. Citado 3 vezes nas páginas 47, 48 e 81.

BAUER, B.; HIENZ, H. Relevant features for video-based continuous sign language recognition. In: *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. [S.l.: s.n.], 2000. p. 440–445. Citado 2 vezes nas páginas 46 e 49.

BAUER, B.; HIENZ, H.; KRAISS, K. F. Video-based continuous sign language recognition using statistical methods. In: *Pattern Recognition, 2000. Proceedings. 15th International Conference on*. [S.l.: s.n.], 2000. v. 2, p. 463–466 vol.2. ISSN 1051-4651. Citado na página 49.

BRADSKI, G.; KAEHLER, A. *Learning OpenCV: Computer Vision with the OpenCV Library*. [S.l.]: O'Reilly Media, Inc., 2008. ISBN 978-0-596-51613-0. Citado na página 38.

BREIMAN, L. Random forests. *Machine Learning*, Kluwer Academic Publishers, v. 45, n. 1, p. 5–32, 2001. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023-/A%3A1010933404324>. Citado na página 71.

BRITTO, G. R. de. *Desenvolvimento de algoritmo para "tracking"de veículos*. 2011. Trabalho de Conclusão de Curso, Escola Politécnica da Universidade de São Paulo, Departamento de Engenharia de Sistemas Eletrônicos, São Paulo, Brasil. Citado na página 28.

BUEHLER, P.; ZISSERMAN, A.; EVERINGHAM, M. Learning sign language by watching tv (using weakly aligned subtitles). In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. [S.l.: s.n.], 2009. p. 2961–2968. ISSN 1063-6919. Citado na página 47.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

CHANDA, P.; AUEPHANWIRIYAKUL, S.; THEERA-UMPON, N. Thai sign language translation system using upright speed-up robust feature and c-means clustering. In: *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*. [S.l.: s.n.], 2012. p. 1–6. ISSN 1098-7584. Citado 2 vezes nas páginas 43 e 45.

CHAVEIRO, N. et al. Mitos da língua de sinais na perspectiva de docentes da universidade federal de goiás. *Revista Virtual de Cultura Surda e Diversidade*, n. 5, 2009. Editora Arara Azul. Disponível em: <http://www.editora-arara-azul.com.br/revista/compar3.php>. Citado na página 20.

COOPER, H.; PUGEAULT, N.; BOWDEN, R. Reading the signs: A video based sign dictionary. In: *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. [S.l.: s.n.], 2011. p. 914–919. Citado na página 47.

DAVYDOV, M. V.; NIKOLSKI, I. V.; PASICHNYK, V. V. Real-time ukrainian sign language recognition system. In: *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*. [S.l.: s.n.], 2010. v. 1, p. 875–879. Citado na página 49.

DIAS, J. M. S. et al. Ogre - open gestures recognition engine. In: *Computer Graphics and Image Processing, 2004. Proceedings. 17th Brazilian Symposium on*. [S.l.: s.n.], 2004. p. 33–40. ISSN 1530-1834. Citado na página 45.

DIGIAMPIETRI, L. et al. Um sistema de informação extensível para o reconhecimento automático de libras. In: *SBSI 2012 - Trilhas Técnicas (Technical Tracks)*. São Paulo, SP, Brazil: [s.n.], 2012. Citado 6 vezes nas páginas 16, 18, 22, 53, 68 e 69.

DIMOV, D.; MARINOV, A.; ZLATEVA, N. Cbir approach to the recognition of a sign language alphabet. In: *Proceedings of the 2007 International Conference on Computer Systems and Technologies*. New York, NY, USA: ACM, 2007. (CompSysTech '07), p. 96:1–96:9. ISBN 978-954-9641-50-9. Disponível em: <http://doi.acm.org/10.1145/1330598-.1330700>. Citado 3 vezes nas páginas 44, 51 e 52.

EL-JABER, M.; ASSALEH, K.; SHANABLEH, T. Enhanced user-dependent recognition of arabic sign language via disparity images. In: *Mechatronics and its Applications (ISMA), 2010 7th International Symposium on*. [S.l.: s.n.], 2010. p. 1–4. Citado 2 vezes nas páginas 45 e 47.

ELAKKIYA, R.; SELVAMANI, K.; KANIMOZHI, S. A framework for recognizing and segmenting sign language gestures from continuous video sequence using boosted learning algorithm. In: *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*. [S.l.: s.n.], 2014. p. 498–503. Citado na página 50.

FACON, J. Morfologia matemática: teoria e exemplos. *Editora Universitária Champagnat da Pontifícia Universidade Católica do Paraná*, 1996. Citado na página 29.

FERREIRA, A. L. et al. *Aprendendo Libras: Módulo 2*. 1ª. ed. [S.l.]: EDUFRN, 2011. 64 p. Citado na página 21.

FILHO, O. M.; NETO, H. V. *Processamento Digital de Imagens*. Rio de Janeiro, Brazil: Editora Brasport, 1999. Citado 4 vezes nas páginas 24, 26, 30 e 31.

FUTANE, P. R.; DHARASKAR, R. V. Video gestures identification and recognition using fourier descriptor and general fuzzy minmax neural network for subset of indian sign language. In: *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*. [S.l.: s.n.], 2012. p. 525–530. Citado na página 47.

GOH, P.; HOLDEN, E. Dynamic fingerspelling recognition using geometric and motion features. In: *Image Processing, 2006 IEEE International Conference on*. [S.l.: s.n.], 2006. p. 2741–2744. ISSN 1522-4880. Citado 5 vezes nas páginas 43, 45, 48, 49 e 81.

GOLDFELD, M. *A criança surda: linguagem e cognição numa perspectiva sociointeracionista*. 2ª. ed. São Paulo: Plexus, 1997. Citado na página 20.

GONÇALVES, V. M. et al. Desempenho de funções de similaridade em cbir no contexto de teste de software: Um estudo de caso em segmentação de imagens de gestos de libras. In: *Proceedings of the VIII Workshop de Visão Computacional (WVC)*. [S.l.]: Brazilian Computer Society, 2012. p. 1–6. Citado 3 vezes nas páginas 45, 47 e 53.

GONZALEZ, R. C.; WOODS, R. E. *Processamento de Imagens Digitais*. [S.l.]: Edgard Blucher, 2000. Citado 5 vezes nas páginas 24, 25, 28, 30 e 31.

GRITTI, T. et al. Local features based facial expression recognition with face registration errors. In: *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*. [S.l.: s.n.], 2008. p. 1–8. Citado na página 47.

GROBEL, K.; ASSAN, M. Isolated sign language recognition using hidden markov models. In: *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*. [S.l.: s.n.], 1997. v. 1, p. 162–167 vol.1. ISSN 1062-922X. Citado 2 vezes nas páginas 22 e 46.

GROBEL, K.; HIENZ, H. Video-based handshape recognition using a handshape structure model in real time. In: *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*. [S.l.: s.n.], 1996. v. 3, p. 446–450 vol.3. ISSN 1051-4651. Citado na página 51.

GRUSZAUSKAS, N. P. et al. Performance of breast ultrasound computer-aided diagnosis: dependence on image selection. *Acad. Radiol*, v. 15(10), p. 1234–1245, 2008. Citado na página 36.

HABILI, N.; LIM, C. C.; MOINI, A. Segmentation of the face and hands in sign language video sequences using color and motion cues. *Circuits and Systems for Video Technology, IEEE Transactions on*, v. 14, n. 8, p. 1086–1097, Aug 2004. ISSN 1051-8215. Citado na página 45.

HALL, M. et al. The weka data mining software: An update. *SIGKDD Explorations*, v. 11, p. 10–18., 2009. Citado 2 vezes nas páginas 37 e 63.

HAN, J.; AWAD, G.; SUTHERLAND, A. Automatic skin segmentation and tracking in sign language recognition. *Computer Vision, IET*, v. 3, n. 1, p. 24–35, March 2009. ISSN 1751-9632. Citado 2 vezes nas páginas 45 e 77.

HAN, J.; AWAD, G.; SUTHERLAND, A. Boosted subunits: A framework for recognising sign language from videos. *Image Processing, IET*, v. 7, n. 1, p. 70–80, February 2013. ISSN 1751-9659. Citado 2 vezes nas páginas 50 e 51.

HIENZ, H.; GROBEL, K.; OFFNER, G. Real-time hand-arm motion analysis using a single video camera. In: *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*. [S.l.: s.n.], 1996. p. 323–327. Citado 3 vezes nas páginas 22, 43 e 46.

HIEU, D. V.; NITSUWAT, S. Image preprocessing and trajectory feature extraction based on hidden markov models for sign language recognition. In: *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD '08. Ninth ACIS International Conference on*. [S.l.: s.n.], 2008. p. 501–506. Citado 3 vezes nas páginas 43, 45 e 46.

IBGE, I. B. de Geografia e E. Características gerais da população, religião e pessoas com deficiência. *Censo demográfico 2010*, p. 1–215, 2010. Disponível em: <http://biblioteca-.ibge.gov.br/visualizacao/periodicos/94/cd_2010_religiao_deficiencia.pdf>. Citado na página 15.

ISAACS, J.; FOO, S. Hand pose estimation for american sign language recognition. In: *System Theory, 2004. Proceedings of the Thirty-Sixth Southeastern Symposium on*. [S.l.: s.n.], 2004. p. 132–136. ISSN 0094-2898. Citado 2 vezes nas páginas 47 e 49.

JUNQUEIRA, D. M.; BRAUN, R. L.; VERLI, H. Alinhamentos. In: _____. *Bioinformática: da Biologia a Flexibilidade Molecular*. 1ª. ed. Porto Alegre - RS: Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, 2014. cap. 3, p. 39–61. Disponível em: <http://www.ufrgs.br/bioinfo/ebook/>. Citado 3 vezes nas páginas 32, 33 e 34.

KELLY, D. et al. Analysis of sign language gestures using size functions and principal component analysis. In: *Machine Vision and Image Processing Conference, 2008. IMVIP '08. International*. [S.l.: s.n.], 2008. p. 31–36. Citado na página 45.

KELLY, D.; MCDONALD, J.; MARKHAM, C. Continuous recognition of motion based gestures in sign language. In: *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. [S.l.: s.n.], 2009. p. 1073–1080. Citado na página 45.

KIM, T.; SHAKHNAROVICH, G.; LIVESCU, K. Fingerspelling recognition with semi-markov conditional random fields. In: *Computer Vision (ICCV), 2013 IEEE International Conference on*. [S.l.: s.n.], 2013. p. 1521–1528. ISSN 1550-5499. Citado 2 vezes nas páginas 45 e 47.

KLIMA, E.; BELLUGI, U. *The Signs of Language*. [S.l.]: Cambridge University Press, 1979. 432 p. Citado 2 vezes nas páginas 20 e 21.

KOVAC, J.; PEER, P.; SOLINA, F. Human skin color clustering for face detection. In: *EUROCON 2003. Computer as a Tool. The IEEE Region 8*. [S.l.: s.n.], 2003. v. 2, p. 144–148 vol.2. Citado na página 65.

LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, v. 10, p. 707, 1966. Citado 2 vezes nas páginas 33 e 34.

LICHTENAUER, J. F. et al. Sign language detection using 3d visual cues. In: *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. [S.l.: s.n.], 2007. p. 435–440. Citado na página 51.

LIRA, G. A.; SOUZA, T. A. F. *Dicionário da língua brasileira de sinais*. [S.l.], 2008. Versão 2.1. Disponível em: <http://www.acessibilidadebrasil.org.br/libras/>. Citado na página 21.

MADANI, H.; NAHVI, M. Isolated dynamic persian sign language recognition based on camshift algorithm and radon transform. In: *Pattern Recognition and Image Analysis (PRIA), 2013 First Iranian Conference on*. [S.l.: s.n.], 2013. p. 1–5. Citado 2 vezes nas páginas 45 e 47.

MADEO, R. C. B. Brazilian sign language multimedia hangman game: A prototype of an educational and inclusive application. In: *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, NY, USA: ACM, 2011. (ASSETS '11), p. 311–312. ISBN 978-1-4503-0920-2. Disponível em: <http://doi.acm.org/10.1145/2049536.2049623>. Citado 2 vezes nas páginas 46 e 54.

MADEO, R. C. B. et al. A committee machine implementing the pattern recognition module for fingerspelling applications. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2010. (SAC '10), p. 954–958. ISBN 978-1-60558-639-7. Disponível em: <http://doi.acm.org/10.1145/1774088.1774287>. Citado na página 53.

MUSHFIELDT, D.; GHAZIASGAR, M.; CONNAN, J. Robust facial expression recognition in the presence of rotation and partial occlusion. In: *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*. New York, NY, USA: ACM, 2013. (SAICSIT '13), p. 186–193. ISBN 978-1-4503-2112-9. Disponível em: <http://doi.acm.org/10.1145/2513456.2513493>. Citado 3 vezes nas páginas 45, 50 e 54.

NAVARRO, G. A guided tour to approximate string matching. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 33, n. 1, p. 31–88, mar. 2001. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/375360.375365>. Citado na página 33.

NETO, J. a. P. S.; OQUENDO, L. Estudo do estado da arte das técnicas de reconhecimento das línguas de sinais por computador. In: *Anais do VI Congresso Tecnológico INFOBRASIL TI & TELECOM*. Fortaleza: [s.n.], 2013. Citado na página 15.

ONG, E. J.; BOWDEN, R. A boosted classifier tree for hand shape detection. In: *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. [S.l.: s.n.], 2004. p. 889–894. Citado 2 vezes nas páginas 51 e 52.

OSMAN, G.; HITAM, M. S.; ISMAIL, M. N. Enhanced skin colour classifier using RGB ratio model. *CoRR*, abs/1212.2692, 2012. Disponível em: <http://arxiv.org/abs/1212-.2692>. Citado na página 66.

PAULRAJ, M. P. et al. Extraction of head and hand gesture features for recognition of sign language. In: *Electronic Design, 2008. ICED 2008. International Conference on*. [S.l.: s.n.], 2008. p. 1–6. Citado 4 vezes nas páginas 43, 47, 49 e 50.

PAULRAJ, M. P. et al. Gesture recognition system for kod tangan bahasa melayu (ktbm) using neural network. In: *Signal Processing Its Applications, 2009. CSPA 2009. 5th International Colloquium on*. [S.l.: s.n.], 2009. p. 19–22. Citado 3 vezes nas páginas 47, 49 e 50.

PAVAN, A.; MODESTO, F. C. Reconhecimento de gestos com segmentação de imagens dinâmicas aplicadas a libras. In: *Proceedings of the VII Workshop de Realidade Virtual e Aumentada*. [S.l.]: Brazilian Computer Society, 2010. Citado na página 47.

PERES, S. M. et al. Libras signals recognition: A study with learning vector quantization and bit signature. In: *Neural Networks, 2006. SBRN '06. Ninth Brazilian Symposium on*. [S.l.: s.n.], 2006. p. 119–124. Citado na página 47.

PISTORI, H.; NETO, J. J. An experiment on handshape sign recognition using adaptive technology: Preliminary results. In: *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA)*. São Luis, Maranhão, Brazil: Springer, 2004. (Lecture Notes in Computer Science, v. 3171), p. 464–473. ISBN 3-540-23237-0. Citado 2 vezes nas páginas 47 e 53.

QUADROS, R. M.; KARNOPP, L. B. *Língua de Sinais Brasileira - Estudos linguísticos*. [S.l.]: Editora Artmed, 2004. Citado na página 15.

QUAN, Y. Chinese sign language recognition based on video sequence appearance modeling. In: *Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on*. [S.l.: s.n.], 2010. p. 1537–1542. Citado na página 50.

RADHA, V.; KRISHNAVENI, M. Threshold based segmentation using median filter for sign language recognition system. In: *Nature Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*. [S.l.: s.n.], 2009. p. 1394–1399. Citado na página 45.

RIBEIRO, H. L.; GONZAGA, A. Hand image segmentation in video sequence by gmm: A comparative analysis. In: *Computer Graphics and Image Processing, 2006. SIBGRAPI '06. 19th Brazilian Symposium on*. [S.l.: s.n.], 2006. p. 357–364. ISSN 1530-1834. Citado na página 45.

ROUSSOS, A. et al. Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos. *J. Mach. Learn. Res.*, JMLR.org, v. 14, n. 1, p. 1627–1663, Jan 2013. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation-.cfm?id=2567709.2567716>. Citado na página 45.

SANDJAJA, I. N.; MARCOS, N. Sign language number recognition. In: *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*. [S.l.: s.n.], 2009. p. 1503–1508. Citado na página 48.

SANTOS, L. D. M. et al. Procedimentos de validação cruzada em mineração de dados para ambiente de computação paralela. In: *ERAD 2009 - 9a. Escola Regional de Alto Desempenho - Arquiteturas Multicore, Caxias do Sul-RS, Brasil*. [S.l.: s.n.], 2009. Citado na página 37.

SHANABLEH, T.; ASSALEH, K. Two tier feature extractions for recognition of isolated arabic sign language using fisher's linear discriminants. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. [S.l.: s.n.], 2007. v. 2, p. II–501–II–504. ISSN 1520-6149. Citado na página 47.

SHANABLEH, T.; ASSALEH, K. Video-based feature extraction techniques for isolated arabic sign language recognition. In: *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*. [S.l.: s.n.], 2007. p. 1–4. Citado na página 47.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology*, Vol. 147, p. 195–197, 1981. Citado na página 33.

SOARES, L. A. M. *Um Estudo de Domínio da Língua Brasileira de Sinais a fim de Colher Requisitos para a Criação de um Modelo Computacional Descritivo desse Idioma*. 2014. Monografia, Programa de Pós-graduação em Ciência da Computação, Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE), Recife, Brasil. Disponível em: <http://www.cin.ufpe.br/~in1020/arquivos/monografias/2013_2/lucas.pdf>. Citado na página 21.

SOONTRANON, N.; ARAMVITH, S.; CHALIDABHONGSE, T. H. Face and hands localization and tracking for sign language recognition. In: *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on*. [S.l.: s.n.], 2004. v. 2, p. 1246–1251 vol.2. Citado 2 vezes nas páginas 51 e 78.

SOUZA, K. P.; DIAS, J. B.; PISTORI, H. Reconhecimento automático de gestos da língua brasileira de sinais utilizando visão computacional. In: *Proceedings of the III Workshop de Visão Computacional*. [S.l.: s.n.], 2007. Citado na página 53.

STARNER, T.; PENTLAND, A. Real-time american sign language recognition from video using hidden markov models. In: *Computer Vision, 1995. Proceedings., International Symposium on*. [S.l.: s.n.], 1995. p. 265–270. Citado 3 vezes nas páginas 48, 49 e 80.

STOKOE, W. C. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of Deaf Studies and Deaf Education*, v. 10, n. 1, p. 3–37, 2005. Citado 2 vezes nas páginas 20 e 21.

SWIFT, D. *Evaluating graphic image files for objectionable content*. Google Patents, 2005. US Patent 6,895,111. Disponível em: <http://www.google.com/patents/US6895111>. Citado na página 66.

TEODORO, B. *Desenvolvimento de Ferramenta para Análise de Sequências de Imagens e Vídeos Digitais em LIBRAS*. 2012. Trabalho de Conclusão de Curso, Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP), São Paulo, Brasil. Citado 3 vezes nas páginas 16, 18 e 22.

TEODORO, B.; DIGIAMPIETRI, L. A local alignment based sign language recognition system. In: FRERY, S. M. A. C. (Ed.). *Workshop of Works in Progress (WIP) in SIBGRAPI 2013 (XXVI Conference on Graphics, Patterns and Images)*. Arequipa, Peru: [s.n.], 2013. Disponível em: <http://www.ucsp.edu.pe/sibgrapi2013/eproceedings/>. Citado na página 22.

THANGALI, A.; SCLAROFF, S. An alignment based similarity measure for hand detection in cluttered sign language video. In: *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*. [S.l.: s.n.], 2009. p. 89–96. ISSN 2160-7508. Citado na página 47.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc., 2006. ISBN 0123695317. Citado na página 31.

TORRE-UGARTE-GUANILO, M. C. De-la; TAKAHASHI, R. F.; BERTOLOZZI, M. R. Revisão sistemática: noções gerais. *Revista da Escola de Enfermagem da USP*, scielo,

v. 45, p. 1260 – 1266, 10 2011. ISSN 0080-6234. Disponível em: <http://www.scielo.br-/scielo.php?script=sci_arttext&pid=S0080-62342011000500033&nrm=iso>. Citado na página 17.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. [S.l.: s.n.], 2001. v. 1, p. I–511–I–518 vol.1. ISSN 1063-6919. Citado na página 70.

WAGNER, P. K. et al. Uma ferramenta para construção de conjuntos de dados de referência para sistemas de análise de gestos baseados em imagens. In: *VIII Simpósio Brasileiro de Sistemas de Informação*. São Paulo: [s.n.], 2012. v. 1, p. 607–618. Citado na página 57.

WILCOX, S.; WILCOX, P. P. *Aprender a Ver*. Editora Arara Azul, 2005. Disponível em: <http://editora-arara-azul.com.br/pdf/livro2.pdf>. Citado na página 15.

WU, S.; NAGAHASHI, H. Real-time 2d hands detection and tracking for sign language recognition. In: *System of Systems Engineering (SoSE), 2013 8th International Conference on*. [S.l.: s.n.], 2013. p. 40–45. Citado 2 vezes nas páginas 45 e 50.

YANG, M.-H.; AHUJA, N.; TABB, M. Extraction of 2d motion trajectories and its application to hand gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 24, n. 8, p. 1061–1074, Aug 2002. ISSN 0162-8828. Citado na página 51.

YANG, R.; SARKAR, S.; LOEDING, B. Enhanced level building algorithm for the movement epenthesis problem in sign language recognition. In: *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*. [S.l.: s.n.], 2007. p. 1–8. ISSN 1063-6919. Citado na página 52.

YAZADI, F. Cyberglove systems cyberglove ii wireless data glove user guide. CyberGlove Systems LLC. 2009. Citado na página 22.

ZHANG, S.; ZHANG, B. Using hmm to sign language video retrieval. In: *Computational Intelligence and Natural Computing Proceedings (CINC), 2010 Second International Conference on*. [S.l.: s.n.], 2010. v. 1, p. 55–59. Citado na página 46.

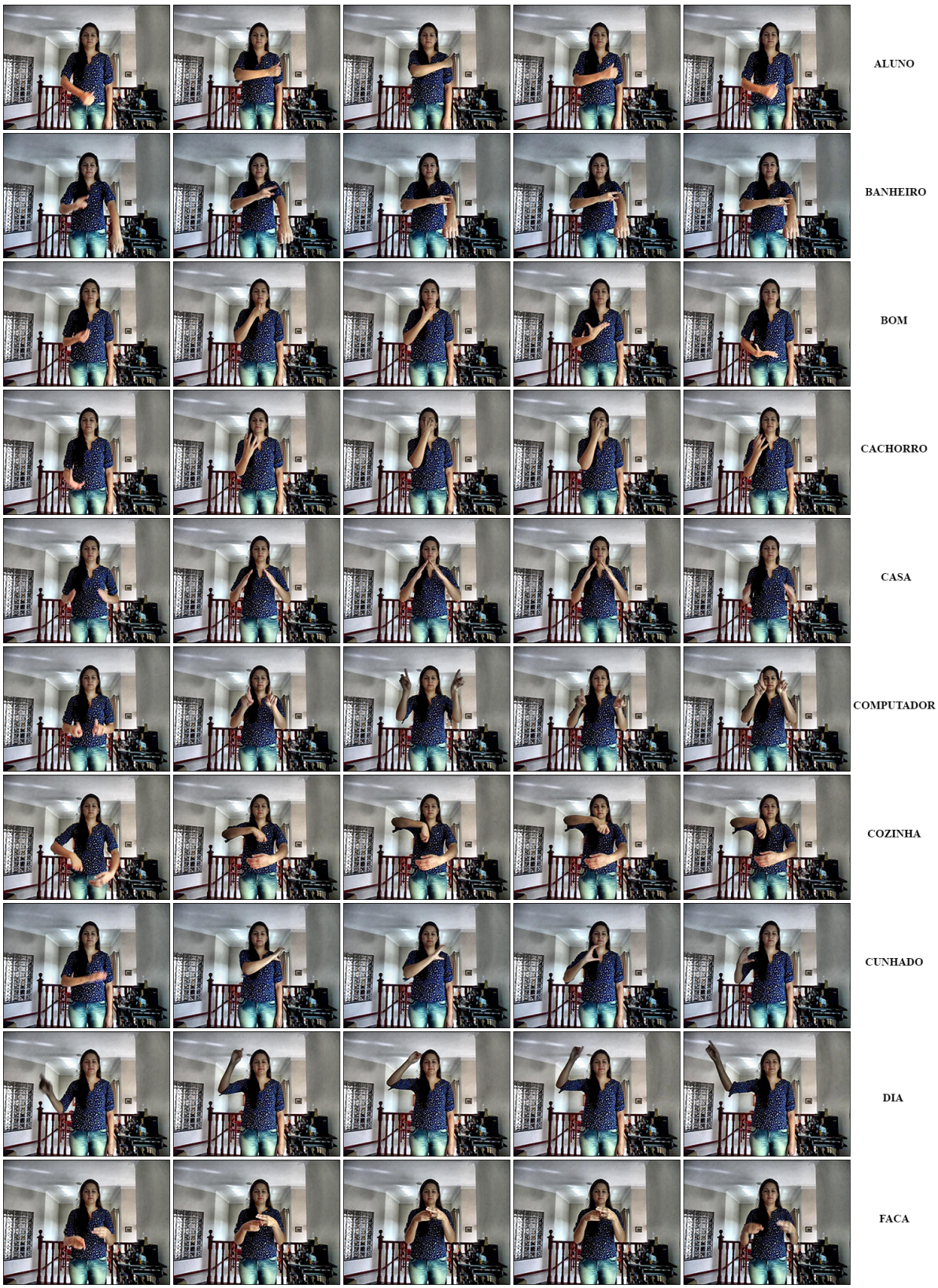
ZHANG, S.; ZHANG, S. Hand language video retrieval based on string ed. In: *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*. [S.l.: s.n.], 2010. v. 3, p. V3–39–V3–43. ISSN 2154-7491. Citado na página 51.

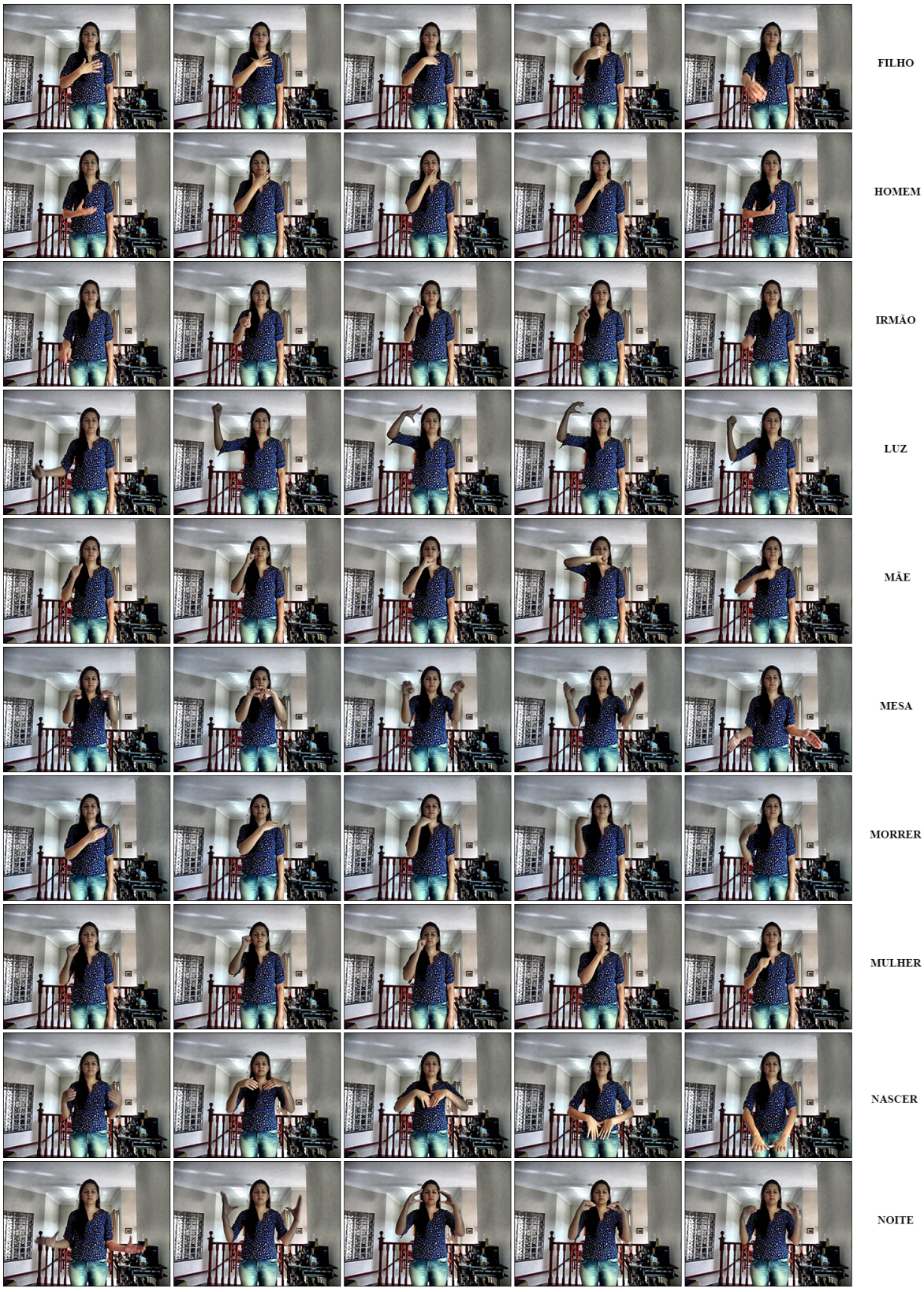
ZIVKOVIC, Z. Improved adaptive gaussian mixture model for background subtraction. In: *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02*. Washington, DC, USA: IEEE Computer Society, 2004. (ICPR '04), p. 28–31. ISBN 0-7695-2128-2. Disponível em: <http://dx.doi.org/10.1109/ICPR.2004.479>. Citado na página 59.

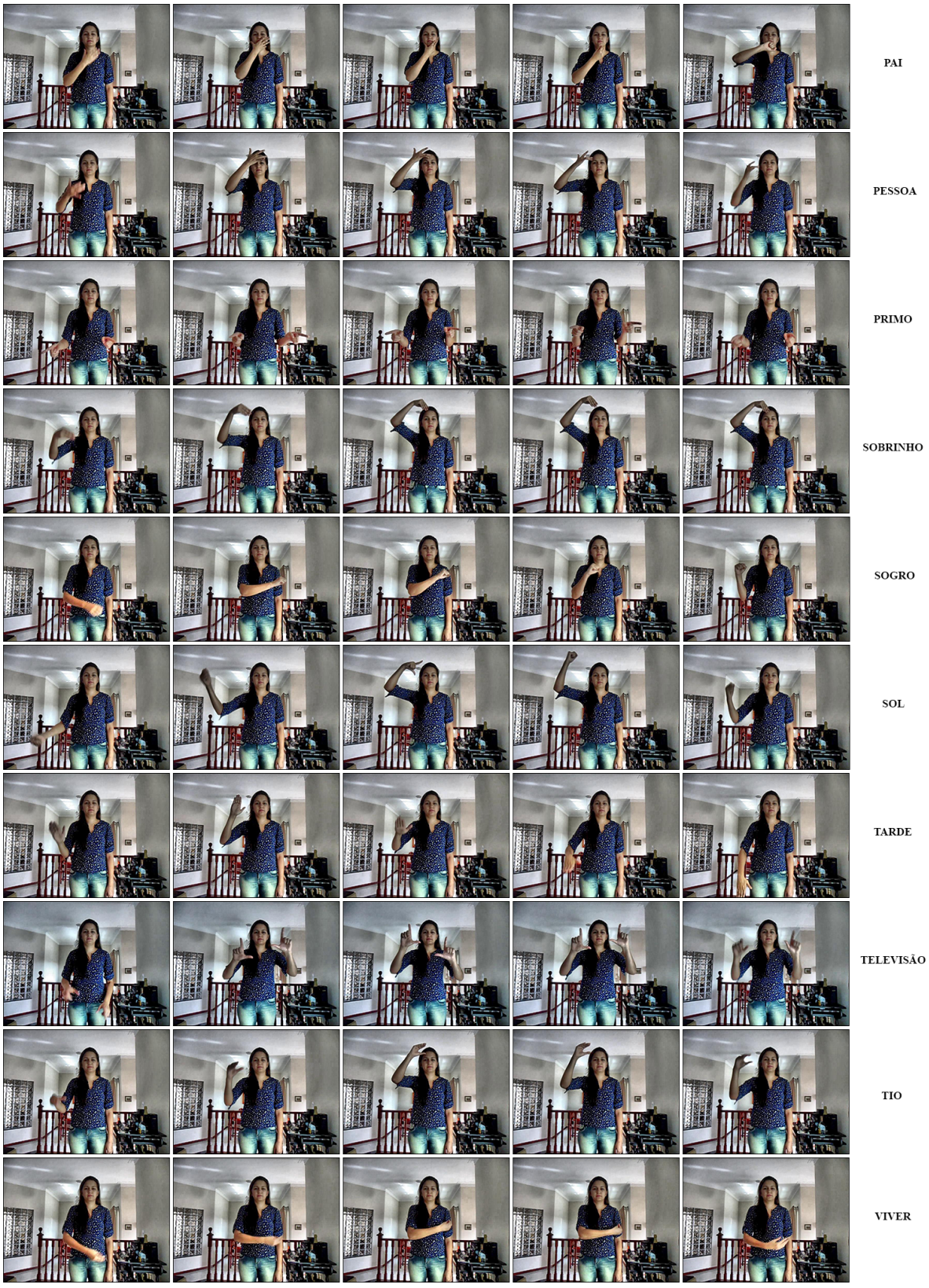
ZIVKOVIC, Z.; HEIJDEN, F. van der. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett.*, Elsevier Science Inc., New York, NY, USA, v. 27, n. 7, p. 773–780, maio 2006. ISSN 0167-8655. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2005.11.005>. Citado na página 59.

ZUIDERVELD, K. Graphics gems iv. In: HECKBERT, P. S. (Ed.). San Diego, CA, USA: Academic Press Professional, Inc., 1994. cap. Contrast Limited Adaptive Histogram Equalization, p. 474–485. ISBN 0-12-336155-9. Disponível em: <http://dl.acm.org-/citation.cfm?id=180895.180940>. Citado na página 58.

Apêndice A – Exemplos de sequências de imagens para cada palavra do banco.




FILHO
HOMEM
IRMÃO
LUZ
MÃE
MESA
MORRER
MULHER
NASCER
NOITE


PAI
PESSOA
PRIMO
SOBRINHO
SOGRO
SOL
TARDE
TELEVISÃO
TIO
VIVER

Apêndice B – Termo de Consentimento Livre e Esclarecido (TCLE)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Convidamos o (a) Sr (a) para participar da pesquisa “Desenvolvimento de um Sistema de Reconhecimento Automático de Língua Brasileira de Sinais”, sob a responsabilidade da pesquisadora Beatriz Tomazela Teodoro, a qual pretende desenvolver um sistema de informação para analisar e reconhecer sinais dinâmicos em sequências de imagens (vídeos) relacionadas à Língua Brasileira de Sinais (LIBRAS) e traduzir automaticamente as palavras expressadas nas sequências dessas imagens para o português.

Sua participação é voluntária e se dará por meio da gravação de dois vídeos curtos, em que você expressará um pequeno conjunto de palavras em LIBRAS. Os vídeos gravados farão parte de um banco de dados que será utilizado para treinar e testar o sistema de reconhecimento em desenvolvimento e em futuros trabalhos científicos da área, sem fins lucrativos. As sequências de imagens (vídeos) poderão ser publicadas em artigos científicos, sem qualquer tipo de compromisso e necessidade de notificação a você.

Os riscos eminentes da sua participação na pesquisa são mínimos. Os únicos riscos vislumbrados, apesar de improváveis são, o de você sofrer algum mal estar ou haver algum desconforto muscular durante a gravação dos vídeos, por exemplo uma câimbra ou um "mal jeito". Se você aceitar participar, estará ajudando a atingir o objetivo principal da pesquisa, o de proporcionar uma maior inclusão dos surdos e/ou deficientes auditivo com a sociedade, facilitando a comunicação destes com pessoas que não tem o conhecimento de LIBRAS.

Se depois de consentir sua participação o Sr (a) desistir de continuar participando, tem o direito e a liberdade de retirar seu consentimento em qualquer fase da pesquisa, seja antes ou depois da coleta dos dados, independente do motivo e sem nenhum prejuízo a sua pessoa. O (a) Sr (a) não terá nenhuma despesa e também não receberá nenhuma remuneração. Os resultados da pesquisa serão analisados e publicados, mas sua identidade não será divulgada, sendo guardada em sigilo. Para qualquer outra informação, o (a) Sr (a) poderá entrar em contato com a pesquisadora pelo e-mail beatriz.teodoro@usp.br ou pelo telefone (14) 38452078, ou poderá entrar em contato com o Comitê de Ética em Pesquisa da Escola de Artes, Ciências e Humanidades – CEP/EACH, na Av. Arlindo Béttio, 1000, Sala T14 - I1, Ermelino Matarazzo, São Paulo-SP, telefone (11) 3091-1046, e-mail cep-each@usp.br.

Consentimento Pós–Informação

Eu,_____, RG n° _____, fui informado sobre o que a pesquisadora quer fazer e porque precisa da minha colaboração, e entendi a explicação. Por isso, eu concordo em participar do projeto, sabendo que não vou ganhar nada e que posso sair quando quiser. Este documento é emitido em duas vias que serão ambas assinadas por mim e pela pesquisadora, ficando uma via com cada um de nós.

_____ _____ , ___ / ___ / _____.

Assinatura do participante

_____

Assinatura do(a) pesquisador(a) responsável

Apêndice C – Resultados obtidos para o reconhecimento das palavras

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
1	aluno	aluno	5	5	100%
2	bom	bom	6	7	85,71%
3	cachorro	cachorro	7	7	100%
4	faca	faca	8	8	100%
5	filho	filho	8	8	100%
6	homem	homem	6	6	100%
7	irmao	irmao	7	7	100%
8	mae	mae	8	9	88,89%
9	morrer	morrer	8	8	100%
10	mulher	mulher	7	7	100%
11	nascer	nascer	9	9	100%
12	pai	pai	6	7	85,71%
13	viver	viver	6	6	100%
14	computador	computador	7	7	100%
15	dia	dia	7	7	100%
16	irmao	irmao	6	7	85,71%
17	luz	luz	8	9	88,89%
18	noite	noite	7	11	63,64%
19	sobrinho	sobrinho	4	4	100%
20	sogro	sogro	5	5	100%
21	sol	sol	5	5	100%
22	tarde	tarde	8	8	100%
23	televisao	televisao	9	10	90%
24	tio	tio	8	9	88,89%
25	aluno	aluno	6	6	100%
26	bom	bom	7	7	100%
27	cachorro	cachorro	5	6	83,33%
28	casa	casa	8	10	80%
29	cunhado	cunhado	5	5	100%
30	filho	filho	7	7	100%
31	homem	homem	7	7	100%
32	irmao	irmao	7	7	100%
33	mae	mae	8	8	100%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
34	morrer	morrer	7	8	87,50%
35	mulher	mulher	7	10	70%
36	nascer	nascer	10	10	100%
37	noite	noite	7	9	77,78%
38	pai	pai	6	8	75%
39	pessoa	pessoa	8	9	88,89%
40	primo	primo	8	8	100%
41	sogro	sogro	6	9	66,67%
42	tarde	tarde	7	7	100%
43	tio	tio	6	6	100%
44	viver	viver	7	7	100%
45	aluno	aluno	7	7	100%
46	banheiro	banheiro	7	9	77,78%
47	bom	bom	7	7	100%
48	cachorro	cachorro	5	7	71,43%
49	casa	casa	9	14	64,29%
50	computador	computador	5	5	100%
51	cozinha	cozinha	9	9	100%
52	cunhado	cunhado	6	8	75%
53	faca	faca	7	8	87,50%
54	filho	filho	7	7	100%
55	homem	homem	9	9	100%
56	irmao	irmao	8	8	100%
57	luz	luz	6	6	100%
58	mae	mae	7	8	87,50%
59	mesa	mesa	6	6	100%
60	morrer	morrer	9	12	75%
61	mulher	mulher	5	7	71,43%
62	nascer	nascer	10	10	100%
63	noite	noite	8	9	88,89%
64	pai	pai	4	7	57,14%
65	pessoa	pessoa	7	9	77,78%
66	primo	primo	8	8	100%
67	sobrinho	sobrinho	5	5	100%
68	sogro	sogro	6	7	85,71%
69	sol	sol	6	6	100%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
70	tarde	tarde	8	8	100%
71	televisao	televisao	6	8	75%
72	tio	tio	8	11	72,73%
73	viver	viver	4	5	80%
74	aluno	aluno	7	7	100%
75	banheiro	banheiro	5	5	100%
76	bom	bom	7	7	100%
77	cachorro	cachorro	7	10	70%
78	casa	casa	8	9	88,89%
79	computador	computador	8	9	88,89%
80	cozinha	cozinha	8	8	100%
81	cunhado	cunhado	6	7	85,71%
82	dia	dia	7	7	100%
83	faca	faca	5	6	83,33%
84	filho	filho	7	7	100%
85	homem	homem	10	11	90,91%
86	irmao	irmao	5	7	71,43%
87	luz	luz	2	4	50%
88	mae	mae	8	8	100%
89	mesa	mesa	5	5	100%
90	morrer	morrer	8	8	100%
91	mulher	mulher	6	6	100%
92	nascer	nascer	8	8	100%
93	noite	noite	8	8	100%
94	pai	pai	7	9	77,78%
95	pessoa	pessoa	6	7	85,71%
96	primo	primo	8	8	100%
97	sobrinho	sobrinho	6	8	75%
98	sogro	sogro	6	7	85,71%
99	sol	sol	7	11	63,64%
100	tarde	tarde	6	7	85,71%
101	televisao	televisao	9	10	90%
102	tio	tio	7	7	100%
103	viver	viver	8	9	88,89%
104	aluno	aluno	7	7	100%
105	cachorro	cachorro	7	8	87,50%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
106	casa	casa	7	7	100%
107	computador	computador	5	7	71,43%
108	cozinha	cozinha	9	10	90%
109	dia	dia	7	7	100%
110	faca	faca	8	9	88,89%
111	filho	filho	9	9	100%
112	homem	homem	8	8	100%
113	irmao	irmao	7	8	87,50%
114	luz	luz	4	4	100%
115	mae	mae	7	8	87,50%
116	mesa	mesa	6	6	100%
117	mulher	mulher	8	8	100%
118	pai	pai	5	5	100%
119	pessoa	pessoa	6	6	100%
120	sobrinho	sobrinho	5	6	83,33%
121	sogro	sogro	5	5	100%
122	sol	sol	6	6	100%
123	televisao	televisao	9	9	100%
124	tio	tio	5	7	71,43%
125	viver	viver	9	9	100%
126	aluno	aluno	6	6	100%
127	banheiro	banheiro	5	5	100%
128	bom	bom	6	7	85,71%
129	cachorro	cachorro	9	11	81,82%
130	casa	casa	5	7	71,43%
131	cozinha	cozinha	9	9	100%
132	cunhado	cunhado	6	6	100%
133	faca	faca	5	9	55,56%
134	filho	filho	6	6	100%
135	homem	homem	6	6	100%
136	irmao	irmao	6	7	85,71%
137	mae	mae	4	5	80%
138	mulher	mulher	6	8	75%
139	nascer	nascer	11	11	100%
140	pai	pai	6	6	100%
141	primo	primo	8	8	100%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
142	viver	viver	4	4	100%
143	aluno	aluno	9	11	81,82%
144	bom	bom	7	7	100%
145	dia	dia	5	5	100%
146	faca	faca	6	6	100%
147	filho	filho	8	9	88,89%
148	homem	homem	8	8	100%
149	irmao	irmao	5	6	83,33%
150	luz	luz	8	8	100%
151	mae	mae	7	7	100%
152	mesa	mesa	4	6	66,67%
153	morrer	morrer	8	8	100%
154	mulher	mulher	6	7	85,71%
155	pessoa	pessoa	8	8	100%
156	primo	primo	9	9	100%
157	sobrinho	sobrinho	4	7	57,14%
158	sogro	sogro	6	7	85,71%
159	sol	sol	8	9	88,89%
160	tarde	tarde	8	9	88,89%
161	televisao	televisao	8	9	88,89%
162	tio	tio	7	7	100%
163	aluno	aluno	9	10	90%
164	banheiro	banheiro	7	8	87,50%
165	bom	bom	8	8	100%
166	cachorro	cachorro	8	8	100%
167	casa	casa	7	7	100%
168	computador	computador	4	4	100%
169	cozinha	cozinha	5	5	100%
170	dia	dia	4	4	100%
171	faca	faca	6	9	66,67%
172	filho	filho	6	6	100%
173	homem	homem	8	8	100%
174	irmao	irmao	5	6	83,33%
175	luz	luz	5	5	100%
176	mae	mae	8	8	100%
177	mesa	mesa	5	5	100%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
178	morrer	morrer	3	3	100%
179	mulher	mulher	7	7	100%
180	nascer	nascer	7	7	100%
181	noite	noite	7	7	100%
182	pai	pai	7	8	87,50%
183	pessoa	pessoa	8	10	80%
184	primo	primo	8	8	100%
185	sobrinho	sobrinho	7	9	77,78%
186	sogro	sogro	5	5	100%
187	sol	sol	5	5	100%
188	tarde	tarde	8	8	100%
189	televisao	televisao	5	5	100%
190	tio	tio	9	9	100%
191	viver	viver	8	8	100%
192	aluno	aluno	5	5	100%
193	banheiro	banheiro	3	6	50%
194	bom	bom	6	6	100%
195	cachorro	cachorro	6	6	100%
196	cozinha	cozinha	5	5	100%
197	cunhado	cunhado	4	4	100%
198	faca	faca	6	6	100%
199	filho	filho	8	9	88,89%
200	homem	homem	8	8	100%
201	irmao	irmao	6	8	75%
202	mae	mae	8	8	100%
203	morrer	morrer	4	6	66,67%
204	mulher	mulher	8	10	80%
205	nascer	nascer	7	7	100%
206	pai	pai	5	5	100%
207	pessoa	pessoa	5	5	100%
208	primo	primo	7	7	100%
209	sogro	sogro	7	10	70%
210	tarde	tarde	6	6	100%
211	televisao	televisao	6	6	100%
212	viver	viver	6	6	100%
213	aluno	aluno	7	8	87,50%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
214	banheiro	banheiro	3	4	75%
215	bom	bom	4	4	100%
216	cachorro	cachorro	7	7	100%
217	casa	casa	7	9	77,78%
218	computador	computador	5	5	100%
219	cozinha	cozinha	4	5	80%
220	cunhado	cunhado	4	8	50%
221	dia	dia	7	7	100%
222	faca	faca	9	9	100%
223	filho	filho	8	8	100%
224	homem	homem	7	7	100%
225	irmao	irmao	8	9	88,89%
226	luz	luz	6	9	66,67%
227	mae	mae	5	5	100%
228	mesa	mesa	5	5	100%
229	morrer	morrer	6	6	100%
230	mulher	mulher	6	7	85,71%
231	nascer	nascer	6	6	100%
232	noite	noite	6	8	75%
233	pai	pai	3	3	100%
234	pessoa	pessoa	6	6	100%
235	primo	primo	5	5	100%
236	sobrinho	sobrinho	4	4	100%
237	sogro	sogro	6	6	100%
238	sol	sol	4	5	80%
239	tarde	tarde	5	5	100%
240	televisao	televisao	7	7	100%
241	tio	tio	6	6	100%
242	viver	viver	4	4	100%
243	aluno	aluno	7	7	100%
244	banheiro	banheiro	5	5	100%
245	bom	bom	6	6	100%
246	cachorro	cachorro	5	5	100%
247	casa	casa	8	9	88,89%
248	computador	computador	4	4	100%
249	cozinha	cozinha	4	5	80%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
250	cunhado	cunhado	3	4	75%
251	dia	dia	5	5	100%
252	faca	faca	4	4	100%
253	filho	filho	5	6	83,33%
254	homem	homem	7	8	87,50%
255	irmao	irmao	7	8	87,50%
256	luz	luz	4	4	100%
257	mae	mae	7	7	100%
258	mesa	mesa	6	6	100%
259	mulher	mulher	8	8	100%
260	nascer	nascer	5	5	100%
261	noite	noite	5	7	71,43%
262	pai	pai	4	4	100%
263	pessoa	pessoa	5	5	100%
264	primo	primo	7	8	87,50%
265	sobrinho	sobrinho	6	7	85,71%
266	sogro	sogro	8	10	80%
267	sol	sol	6	8	75%
268	tarde	tarde	5	5	100%
269	televisao	televisao	7	7	100%
270	tio	tio	6	7	85,71%
271	viver	viver	4	4	100%
272	aluno	aluno	5	5	100%
273	bom	bom	5	5	100%
274	cachorro	cachorro	5	5	100%
275	casa	casa	5	8	62,50%
276	cozinha	cozinha	7	7	100%
277	dia	dia	5	6	83,33%
278	filho	filho	6	7	85,71%
279	homem	homem	8	8	100%
280	irmao	irmao	6	6	100%
281	luz	luz	5	5	100%
282	mae	mae	6	6	100%
283	morrer	morrer	5	5	100%
284	mulher	mulher	6	6	100%
285	nascer	nascer	7	7	100%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
286	noite	noite	10	10	100%
287	pessoa	pessoa	4	4	100%
288	primo	primo	5	5	100%
289	sobrinho	sobrinho	6	6	100%
290	sogro	sogro	7	8	87,50%
291	sol	sol	5	8	62,50%
292	tarde	tarde	4	4	100%
293	televisao	televisao	6	8	75%
294	tio	tio	4	5	80%
295	viver	viver	5	6	83,33%
296	aluno	aluno	4	6	66,67%
297	bom	bom	2	2	100%
298	cachorro	cachorro	7	7	100%
299	casa	casa	8	9	88,89%
300	computador	computador	3	4	75%
301	cozinha	cozinha	6	6	100%
302	cunhado	cunhado	2	2	100%
303	dia	dia	1	3	33,33%
304	faca	faca	6	7	85,71%
305	filho	filho	6	7	85,71%
306	homem	homem	6	6	100%
307	irmao	irmao	4	5	80%
308	luz	luz	3	4	75%
309	mae	mae	5	5	100%
310	morrer	morrer	5	5	100%
311	mulher	mulher	5	5	100%
312	nascer	nascer	6	6	100%
313	noite	noite	7	8	87,50%
314	pai	pai	5	5	100%
315	pessoa	pessoa	5	5	100%
316	primo	primo	6	6	100%
317	sobrinho	sobrinho	5	6	83,33%
318	sol	sol	4	5	80%
319	tarde	tarde	5	5	100%
320	televisao	televisao	7	7	100%
321	tio	tio	8	10	80%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
322	viver	viver	5	5	100%
323	aluno	aluno	7	8	87,50%
324	banheiro	banheiro	3	4	75%
325	bom	bom	5	5	100%
326	cachorro	cachorro	4	7	57,14%
327	casa	casa	6	8	75%
328	computador	computador	5	5	100%
329	cozinha	cozinha	6	6	100%
330	cunhado	cunhado	4	5	80%
331	dia	dia	5	5	100%
332	faca	faca	7	11	63,64%
333	filho	filho	6	6	100%
334	homem	homem	4	4	100%
335	irmao	irmao	7	9	77,78%
336	luz	luz	5	7	71,43%
337	mae	mae	6	6	100%
338	mesa	mesa	7	10	70%
339	morrer	morrer	4	4	100%
340	mulher	mulher	7	7	100%
341	nascer	nascer	7	7	100%
342	noite	noite	9	11	81,82%
343	pai	pai	5	5	100%
344	pessoa	pessoa	6	6	100%
345	primo	primo	5	6	83,33%
346	sobrinho	sobrinho	4	6	66,67%
347	sogro	sogro	5	6	83,33%
348	sol	sol	6	7	85,71%
349	tarde	tarde	4	5	80%
350	televisao	televisao	8	12	66,67%
351	tio	tio	5	6	83,33%
352	viver	viver	5	5	100%
353	aluno	aluno	6	6	100%
354	banheiro	banheiro	5	9	55,56%
355	bom	bom	6	6	100%
356	cachorro	cachorro	6	6	100%
357	casa	casa	6	8	75%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
358	computador	computador	5	5	100%
359	cozinha	cozinha	5	6	83,33%
360	dia	dia	5	5	100%
361	filho	filho	7	7	100%
362	homem	homem	7	7	100%
363	luz	luz	5	5	100%
364	mae	mae	7	7	100%
365	morrer	morrer	5	5	100%
366	mulher	mulher	4	4	100%
367	nascer	nascer	5	5	100%
368	noite	noite	7	8	87,50%
369	pai	pai	4	4	100%
370	pessoa	pessoa	4	4	100%
371	primo	primo	10	10	100%
372	sobrinho	sobrinho	4	7	57,14%
373	sol	sol	6	7	85,71%
374	tarde	tarde	5	5	100%
375	televisao	televisao	7	9	77,78%
376	tio	tio	7	9	77,78%
377	viver	viver	5	5	100%
378	aluno	aluno	6	6	100%
379	bom	bom	7	8	87,50%
380	cachorro	cachorro	6	7	85,71%
381	cozinha	cozinha	4	4	100%
382	cunhado	cunhado	2	2	100%
383	faca	faca	7	10	70%
384	filho	filho	9	10	90%
385	homem	homem	7	8	87,50%
386	irmao	irmao	8	10	80%
387	morrer	morrer	5	7	71,43%
388	mulher	mulher	5	5	100%
389	nascer	nascer	6	6	100%
390	primo	primo	5	5	100%
391	tio	tio	8	9	88,89%
392	viver	viver	6	7	85,71%
393	aluno	aluno	7	7	100%

Continua na próxima página.

Número da amostra	Palavra sinalizada	Palavra reconhecida	Votos corretos	Total de votos	Porcentagem
394	banheiro	banheiro	4	5	80%
395	bom	bom	7	8	87,50%
396	cachorro	cachorro	6	6	100%
397	casa	casa	5	5	100%
398	computador	computador	4	5	80%
399	cozinha	cozinha	5	5	100%
400	cunhado	cunhado	3	3	100%
401	dia	dia	6	6	100%
402	faca	faca	6	6	100%
403	filho	filho	8	12	66,67%
404	homem	homem	7	7	100%
405	irmao	irmao	7	7	100%
406	luz	luz	5	6	83,33%
407	mae	mae	6	8	75%
408	mesa	mesa	3	3	100%
409	morrer	morrer	4	7	57,14%
410	mulher	mulher	3	5	60%
411	nascer	nascer	10	10	100%
412	noite	noite	5	5	100%
413	pai	pai	5	5	100%
414	pessoa	pessoa	6	7	85,71%
415	primo	primo	5	5	100%
416	sobrinho	sobrinho	5	7	71,43%
417	sogro	sogro	7	7	100%
418	sol	sol	6	10	60%
419	tarde	tarde	8	8	100%
420	televisao	televisao	7	8	87,50%
421	tio	tio	7	7	100%
422	viver	viver	7	7	100%

Anexo A – Parecer Consubstanciado do Comitê de Ética em Pesquisa Envolvendo Seres Humanos da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP)

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Desenvolvimento de um Sistema de Reconhecimento Automático de Língua Brasileira de Sinais

Pesquisador: Beatriz Teodoro

Área Temática:

Versão: 2

CAAE: 36135214.4.0000.5390

Instituição Proponente: Escola de Artes, Ciências e Humanidades - EACH/USP

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 915.307

Data da Relatoria: 03/11/2014

Apresentação do Projeto:

Os autores referem que nos últimos anos, é possível observar que o empenho em facilitar a comunicação entre surdos e pessoas que não conhecem uma língua gestual tem aumentado, mas ainda há poucos ambientes acessíveis para os surdos. O número de pessoas isentas de deficiência auditiva que sabem se expressar utilizando língua de sinais é muito pequeno, tornando extremamente difícil a comunicação de surdos com estas pessoas. Além disso, o reconhecimento e tradução de língua de sinais é uma área bastante complexa e que está em estado inicial de desenvolvimento.

Assim, o foco do estudo proposto é o reconhecimento de Língua Brasileira de Sinais (LIBRAS), com a finalidade de simplificar a comunicação entre surdos conversando em LIBRAS e ouvintes que não conheçam esta língua. O reconhecimento será realizado através do processamento de imagens e vídeos digitais de pessoas se comunicando em LIBRAS, sem o uso de luvas coloridas e/ou luvas de dados e sensores, focando em sinais que utilizam apenas as mãos.

Objetivo da Pesquisa:

Objetivo primário: desenvolver um sistema para analisar e reconhecer sinais dinâmicos em sequências de imagens (vídeos) relacionadas à LIBRAS e traduzir automaticamente as palavras
expressas nas sequências dessas imagens para o português, sem o uso de luvas coloridas ou a

exigência de gravações de alta qualidade em estúdios

Avaliação dos Riscos e Benefícios:

Quanto aos riscos são descritos tanto no TCLE quanto no trabalho detalhado e nas informações básicas do projeto que são mínimos. Os únicos riscos vislumbrados, apesar de improváveis são, o de sofrer algum mal estar ou haver algum desconforto muscular durante a gravação dos vídeos, por exemplo uma câimbra ou um "mal jeito".

Os benefícios de acordo com os autores vislumbram proporcionar uma maior
inclusão dos surdos e/ou deficientes auditivos com a sociedade, facilitando a comunicação destes com pessoas que não tem o conhecimento de LIBRAS.

Comentários e Considerações sobre a Pesquisa:

Pesquisa com temática relevante uma vez que traz proposta de desenvolvimento de recurso para facilitar a comunicação entre surdos e pessoas que não conhecem uma língua gestual, através do desenvolvimento de um sistema de informação para o reconhecimento automático de LIBRAS.

Considerações sobre os Termos de apresentação obrigatória:

Apresenta carta de apresentação do protocolo de pesquisa ao Comitê de Ética em Pesquisa envolvendo seres humanos da EACH, folha de rosto assinada em 09/09/2014, termo de consentimento livre e esclarecido e projeto detalhado.

Recomendações:

Aprovação

Conclusões ou Pendências e Lista de Inadequações:

Não há pendências

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Considerações Finais a critério do CEP:

SAO PAULO, 15 de Dezembro de 2014

Assinado por:
Beatriz Aparecida Ozello Gutierrez
(Coordenador)

Endereço: Av. Arlindo Béttio, nº 1000
Bairro: Ermelino Matarazzo **CEP:** 03.828-000
UF: SP **Município:** SAO PAULO
Telefone: (11)3091-1046 **E-mail:** cep-each@usp.br