UMA ESTRATÉGIA PARA RECONHECIMENTO DE SINAIS DA LÍNGUA BRASILEIRA DE SINAIS UTILIZANDO APRENDIZADO PROFUNDO

ADA RAQUEL DOS SANTOS CRUZ

UMA ESTRATÉGIA PARA RECONHECIMENTO DE SINAIS DA LÍNGUA BRASILEIRA DE SINAIS UTILIZANDO APRENDIZADO PROFUNDO

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientadora: Eulanda Miranda dos Santos

Manaus

Maio de 2020

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Cruz, Ada Raquel dos Santos
C957e Uma estratégia para reconhecimento de sinais da Língua Brasileira de Sinais utilizando aprendizado profundo / Ada Raquel dos Santos Cruz . 2020
78 f.: il. color; 31 cm.

Orientadora: Eulanda Miranda dos Santos
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Reconhecimento de Gestos. 2. Língua Brasileira de Sinais. 3. Aprendizado Profundo. 4. Redes Neurais Convolutivas. 5. Tecnologia Assistiva. I. Santos, Eulanda Miranda dos. II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"UMA ESTRATÉGIA PARA RECONHECIMENTO DE SINAIS DA LÍNGUA BRASILEIRA DE SINAIS UTILIZANDO APRENDIZADO PROFUNDO"

ADA RAQUEL DOS SANTOS CRUZ

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Profa. Eulanda Miranda dos Santos - PRESIDENTE

Prof. Marco Antonio Pinheiro de Cristo - MEMBRO INTERNO

Prof. Tiago Maritan Ugulino de Araújo - MEMBRO EXTERNO

Manaus, 26 de Maio de 2020

Dedico este trabalho à minha família.

Agradecimentos

Agradeço em primeiro lugar ao Eterno Rei do Universo, bendito seja Ele, que me concedeu vida, me sustentou e me permitiu chegar a essa ocasião. Por ter sido meu apoio em momentos difíceis e não ter permitido que meus pés vacilassem. Por todas as oportunidades e escapes. Por tudo.

Agradeço à minha família, em especial à minha mãe, Ruth, que teve força não só para cuidar de meus irmãos e de mim, como também para nos incentivar e enfatizar a importância dos estudos, proporcionando e viabilizando sempre a liberdade de escolha que me ajudou a chegar até aqui. Agradeço aos meus irmãos, Renuel, Rebeca, Rute, Ariel, Hadassa e Antonio, que me ensinam muito mais do que possam imaginar, que em todos os momentos me nutriram com força, um pouco de estresse, paciência, muito amor e carinho. À mãe Ada, meu amor, por todo amor, carinho e conselhos. Ao meu presente, Everton, que esteve ao meu lado durante esse período e sempre se fez presente, me confortando em momentos difíceis, me ajudando com *insights* e iluminando a minha vida. Aos meus avós Manoel e Rosalba, por todo o amor e carinho, por me transmitirem força, fé e firmeza. Ao seu Edson e à dona Telma, por todo carinho e incentivo. Aos meus primos/irmãos Amel, Lhia, Jerusa e Júnior, e aos meus tios João Luiz, Ronilde e João Paulo, que me ajudaram sempre que puderam, antes e durante esse período, de várias maneiras, no mínimo e no máximo. Aos amigos da Congregação Israelita em Manaus, pelo companheirismo e força fundamentais. Ao meu pai Robson e à Daiane, pelo apoio e amizade.

Agradeço à professora Eulanda Santos, minha querida orientadora e conterrânea, que aceitou trabalhar comigo, mesmo de longe. Por ser paciente e exigente ao mesmo tempo. Por ter revisado as 300 versões da dissertação, atentando aos detalhes. Pelas conversas descontraídas e pelas três diferentes reuniões marcadas para o mesmo horário.

Agradeço ao meu professor e amigo, Marcelo Chamy, que me acompanha e acredita em mim desde a graduação. Quem me incentivou a entrar no mestrado e me apoiou durante o processo, me encorajando a dar o meu melhor.

Agradeço aos meus colegas de turma e a todos os que se tornaram amigos. Vocês

deixaram as coisas mais leves. Ao Ralph Breno e ao Anderson Pimentel, que me deram um importante suporte no início do mestrado, com dicas em disciplinas e conselhos. Também agradeço os parceiros do grupo de estudos, que me ajudaram muito no início, direta e indiretamente: Marcos Pereira, Yumi Ouchi, Ricardo Guimarães, Gabriel Leitão e Patrícia Chourio.

Agradeço ao Fagner Cunha e a todos os amigos e colegas de laboratório e almoço, especialmente o prezado e nobre Rayol Neto, o Léo Rodrigues, a Vitinha Aires, o Hendrio Luis, o Helmer Mourão, o Leandro Okimoto e o Felipe Lobo. Obrigada pela parceria e por terem me ajudado com os treinos da defesa!

Ao Thiago Marques, meu amigo, autor dos bordões inspiradores, que me repassou tanto conhecimento (é uma pena que meu cérebro só guarde as piadas). Já te falei isso, Thiago. Obrigada por me relembrar várias vezes que todo o fracasso começa [...].

Aos professores Marco Cristo e Tiago Maritan, por terem contribuído para o desenvolvimento e finalização deste trabalho. Também, a todos os professores do PPGI, por todo o conhecimento transmitido e pela disponibilidade. Em especial, agradeço ao professor Eduardo Souto, com quem comecei a aprender a trilhar o caminho da pesquisa, pondo em prática as suas valiosas dicas!

A todos os servidores do Instituto de Computação da UFAM, especialmente ao pessoal da secretaria, que juntamente com o professor Eduardo Feitosa, foram atenciosos desde o início do processo seletivo e sempre se mostraram dispostos a ajudar a todos: o Frank, o Robson, o Márcio e a Helen.

À Priscila e ao Alejandro, que me deram o suporte fundamental para que eu fosse capaz de concluir esse trabalho.

Finalmente, agradeço a todos os que contribuíram direta e indiretamente para o desenvolvimento e finalização deste trabalho. Muito obrigada a todos.

"Ben Zoma dizia: a quem se deve chamar sábio? Àquele que aprende com toda pessoa, conforme foi dito: "de todos os que me ensinaram obtive ensinamento"; decerto Teus testemunhos são minha meditação.

A quem se deve considerar forte? Àquele que domina sua má inclinação, como se diz na Escritura: aquele que é lento para a ira é melhor que o homem forte; e aquele que domina suas emoções é melhor do que o que conquista uma cidade.

A quem se deve considerar rico? Àquele que se alegra com o que possui, conforme se lê: feliz serás quando comeres do produto do esforço das tuas mãos e o bem estará contigo. Feliz serás nesse mundo e bem haverá para ti no outro mundo.

A quem se deve respeito? Àquele que respeita os seus semelhantes, conforme foi dito: respeitarei os que me respeitarem e os que me desprezarem serão desprezados."

(Pirkei Avot, Capítulo 4)

Resumo

As línguas de sinais são línguas naturais e vivas utilizadas para comunicação não oral entre surdos, deficientes auditivos e ouvintes. No Brasil, a língua de sinais, conhecida como Libras, é legalmente reconhecida como meio de expressão e comunicação dessa parcela da população, que equivale a cerca de 9 milhões de pessoas. Com o intuito de facilitar a comunicação entre surdos e ouvintes, algumas tecnologias assistivas de transcrição da Libras para o português foram desenvolvidas, especialmente baseadas em técnicas de visão computacional. Nesse cenário, as redes neurais convolutivas são amplamente utilizadas por apresentarem resultados considerados estado da arte em reconhecimento de gestos. Porém, ainda não foi encontrada uma solução efetiva para o problema, devido, principalmente: ao alto custo financeiro de implantação, como por exemplo, a utilização de dispositivos específicos de aquisição de dados; ao aspecto intrusivo, visto que algumas soluções utilizam sensores portáteis, como luvas equipadas com sensores de movimento; e a limitações técnicas, pois, apesar da maioria dos sinais da Libras possuir movimento, a área de reconhecimento de sinais da Libras é majoritariamente dominada por soluções que consideram apenas sinais estáticos. Além disso, poucos trabalhos exploram o impacto da utilização de técnicas como transferência de aprendizado e aumento de dados, ou fusão de diferentes canais de dados.

Assim, o desenvolvimento deste trabalho se apoia na necessidade da elaboração de um método para reconhecimento de sinais da Libras, que seja de baixo custo, não intrusivo e eficiente no reconhecimento de sinais que são executados em movimento. Como resultado, este trabalho apresenta uma estratégia para reconhecimento de sinais estáticos e dinâmicos da Libras combinando rede neural convolutiva tridimensional, fusão de dados de múltiplos canais e transferência de aprendizado.

Palavras-chave: Língua de sinais, Língua Brasileira de Sinais, Reconhecimento de Gestos, Tecnologia Assistiva, Aprendizado Profundo, Redes Neurais Convolutivas.

Abstract

Sign languages are natural and living languages used for non-verbal communication between deaf, hearing impaired and hearing people. In Brazil, the sign language known as Libras is legally recognized as the language of expression and communication for this group of the population, composed of approximately 9 million people. To facilitate communication between deaf and hearing people, some assistive technologies for transcription of Libras to Portuguese have been developed, especially using vision-based techniques. In this scenario, convolutional neural networks are widely used due to achieving results considered state of the art in the area of gesture recognition. However, an effective solution to this problem has not yet been found, mainly due to: the high financial cost of implementation, such as the use of specific data acquisition devices; the intrusive aspect, since some solutions use portable sensors, such as gloves equipped with motion sensors; and technical limitations, because, despite the majority of Libras signs being executed with motion, the Libras signs recognition area is especially dominated by solutions that consider only static signs. Besides, few studies explore the impact of using techniques such as transfer learning, data augmentation and fusion different data channels.

Thus, the development of this work is based on the need for a method to perform signs Libras recognition, which low cost, in a non-intrusive manner, and efficient in recognizing signs that are executed in motion. To accomplish this, this work aims to present a strategy for the recognition of static and dynamic Libras signs employing a three-dimensional convolutional neural network, fusion data from multiple channels, and transfer learning.

Keywords: Sign Language, Brazilian Sign Language, Gesture Recognition, Assistive Technology, Deep Learning, Convolutional Neural Networks..

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

As línguas de sinais são línguas naturais e vivas utilizadas em diversos países para comunicação não oral entre surdos, deficientes auditivos e ouvintes. Elas são compostas por aspectos gramaticais, léxicos, sintáticos, semânticos e pragmáticos (Capovilla e Raphael, 2004). No Brasil, a língua brasileira de sinais (Libras) é legalmente reconhecida como meio de expressão e comunicação dessa parcela da população (Brasil, 2005). Considerando essa realidade, a tecnologia tem sido utilizada como meio para facilitar a interação e a comunicação entre os indivíduos. Alguns exemplos disso são a disponibilização de dicionários *on-line* português-Libras[1], a publicação de vídeos com tradução de músicas em português para Libras[2], o lançamento de desenhos animados com diálogos em Libras[3], e aplicações de tradução automática de português para Libras[4].

Embora grande parte dos esforços seja direcionada à tradução português-Libras, algumas tecnologias assistivas de transcrição da Libras para o português — ou de reconhecimento de sinais da Libras — foram desenvolvidas, especialmente baseadas em técnicas de visão computacional (Leal, 2018; Voigt, 2018; Silva, 2018; Machado, 2018; Magalhaes, 2018). Entretanto, ainda não foi encontrada uma solução efetiva para o problema, devido, principalmente: às limitações técnicas, considerando que, apesar da maioria dos sinais da Libras possuir movimento, a área de reconhecimento de sinais da Libras é majoritariamente dominada por soluções que consideram apenas sinais estáticos; ao alto custo financeiro de implantação, como por exemplo, a utilização de dispositivos específicos de aquisição de dados; e ao aspecto intrusivo, visto que algumas soluções utilizam sensores portáteis, como luvas equipadas com sensores de movimento. Os dois últimos aspectos são muito importantes a se considerar quando se

[1] http://www.ines.gov.br/dicionario-de-libras/main_site/libras.htm
[2] https://www.youtube.com/channel/UC4x7_sGe2H4L9yvUjBxAyYA
[3] https://www.youtube.com/channel/UCJtOTvG4EvBGkvtTVVv8Lpg
[4] https://www.handtalk.me/br/Aplicativo

trata de ferramentas de tecnologia assistiva, cujo objetivo é melhorar a qualidade de vida de pessoas com necessidades especiais.

Portanto, o desenvolvimento deste trabalho se apoia na necessidade da elaboração de um método para reconhecimento de sinais da Libras que seja de baixo custo, não intrusivo e eficiente no reconhecimento de sinais que são executados em movimento. De modo geral, espera-se que o método proposto, ao atender a esses requisitos, melhore a experiência de comunicação entre surdos e ouvintes, sendo capaz de reconhecer uma ampla gama de sinais da Libras.

No restante deste capítulo são descritos a motivação social deste trabalho, na Seção 1.1; a definição dos desafios da tarefa de reconhecimento de gestos, na Seção 1.2; a justificativa da hipótese (Seção 1.3); o objetivo geral (Seção 1.6), e os objetivos específicos (Seção 1.4.1); a metodologia adotada (Seção 1.5); e a forma como está estruturada esta dissertação (Seção 1.7).

1.1 Motivação

A necessidade da comunicação remonta às primeiras sociedades, quando o homem utilizava pinturas rupestres como uma forma de comunicação e expressão (Mattelart e Mattelart, 2011). Com o passar dos anos, o sistema de comunicação foi aprimorado e, atualmente, a maioria das pessoas utiliza a língua falada para se comunicar. Entretanto, além da língua falada, há também a língua de sinais, um conjunto de formas gestuais utilizado para comunicação não oral entre surdos, deficientes auditivos e ouvintes. Assim como as línguas verbais, cada país possui sua língua de sinal particular.

O Censo (2010) realizou um levantamento do índice de brasileiros com deficiência auditiva: na época havia, aproximadamente, 9,7 milhões de pessoas com algum grau da deficiência. Dessas, 2,1 milhões apresentava grande dificuldade, ou nenhuma capacidade de audição, sendo cerca de 63 mil crianças em idade escolar. Desse grupo, a maioria utilizava a Libras como língua natural principal.

A Libras é legalmente reconhecida como meio de expressão e comunicação da comunidade surda brasileira (Brasil, 2002). Após uma reflexão sobre o assunto, emergem várias áreas do cotidiano das pessoas que devem ser ajustadas para atender às demandas de comunicação com surdos e deficientes auditivos. Além disso, os dados do Censo (2010) podem levar a uma discussão sobre quais são os dispositivos que legislam sobre a matéria, com a finalidade de elucidar o contexto atual das políticas públicas que objetivam viabilizar a integração dessa parcela da população.

O Decreto nº 3.298, de 20 de dezembro de 1999, que regulamenta a lei nº 7.853,

de 24 de outubro de 1989, dispõe sobre a Política Nacional para a Integração da Pessoa Portadora de Deficiência (Brasil, 1999). A artigo 2º reza que "cabe aos órgãos e às entidades do Poder Público assegurar à pessoa portadora de deficiência o pleno exercício de seus direitos básicos, inclusive dos direitos à educação, à saúde, ao trabalho, ao desporto, ao turismo, ao lazer, à previdência social, à assistência social, ao transporte, à edificação pública, à habitação, à cultura, ao amparo à infância e à maternidade, e de outros que, decorrentes da Constituição e das leis, propiciem seu bem-estar pessoal, social e econômico". É importante destacar o artigo 7 que, no inciso II, descreve o objetivo de "integração das ações dos órgãos e das entidades públicos e privados nas áreas de saúde, educação, trabalho, transporte, assistência social, edificação pública, previdência social, habitação, cultura, desporto e lazer, visando à prevenção das deficiências, à eliminação de suas múltiplas causas e à inclusão social", que pode ser alcançado, conforme discorrido no artigo 8, inciso IV, utilizando instrumentos obtidos por meio do "fomento da tecnologia de bioengenharia voltada para a pessoa portadora de deficiência, bem como a facilitação da importação de equipamentos".

No que diz respeito ao quadro geral da educação, por exemplo, o Governo Federal Brasileiro, por meio do Ministério da Educação (MEC), sancionou várias normas jurídicas (Brasil, 2002, 2005, 2003), como leis, decretos e portarias, com o objetivo de facilitar a inclusão e a permanência de alunos deficientes em todas as modalidades de ensino. No âmbito escolar, o Decreto nº 5.626 de 2005 (Brasil, 2005) regulamenta a lei nº 10.436, de 24 de abril de 2002 (Brasil, 2002), que torna obrigatória a inserção da Libras em todas as instituições de ensino, públicas e privadas, como disciplina curricular nos cursos de licenciatura e fonoaudiologia. No artigo 23, é determinado que as instituições federais de ensino devem proporcionar aos alunos surdos os serviços de tradutor e intérprete de Libras-português em sala de aula e em outros espaços educacionais, além de equipamentos e tecnologias que viabilizem o acesso à comunicação, à informação e à educação. No artigo 24, determina-se que em cursos de ensino a distância, os sistemas de acesso à informação devem exibir uma janela com tradutor e intérprete de Libras-português e subtitulação por meio do sistema de legenda.

Portanto, para dar cumprimento às normas referentes ao atendimento a estudantes integrantes da comunidade surda, três medidas se fazem necessárias: a capacitação de professores em Libras; a contratação de intérpretes para os alunos surdos ou com deficiência auditiva; e a aquisição de equipamentos e tecnologias assistivas, grupo no qual as aplicações para reconhecimento de língua de sinais (RLS) se enquadram.

Existem aplicações de transcrição automática da Libras para o português, especialmente baseadas em técnicas de visão computacional. Entretanto, ainda não foi encontrada uma solução efetiva devido a questões de inviabilidade da aplicação prática

das soluções propostas, derivada de três motivos principais: alto custo financeiro de implantação; métodos intrusivos, sendo necessário o uso de sensores pelo corpo (Silva, 2018); e limitações técnicas, como por exemplo, algumas abordagens que são limitadas ao reconhecimento de gestos estáticos, letras, números e poucas palavras (Barros Junior, 2016; Gonçalves et al., 2016).

1.2 Definição do Problema

A análise de movimento humano por meio de técnicas de visão computacional é um amplo domínio de estudo, que abrange muitas áreas de pesquisa e múltiplas sub-tarefas, incluindo problemas como, por exemplo, o reconhecimento de expressões faciais, de gestos manuais ou de expressão corporal, de ação e de atividades. Embora soluções semelhantes possam ser usadas para tratar cada um desses problemas, dado que são problemas que podem ser considerados similares, todos possuem peculiaridades que carecem de análise especial para que sistemas eficientes sejam elaborados.

Como toda língua humana, a língua de sinais passa pelo processo contínuo e gradual de variação e mudança, seja por motivações internas, seja por contato com outras línguas de sinais ou orais. Nesse sentido, elas possuem os mesmos universais linguísticos que caracterizam as línguas orais, com aspectos semânticos como a sinonímia, a homonímia e a polissemia (Xavier, 2006).

As informações da língua de sinais podem ser transmitidas usando gestos das mãos, posição da cabeça e partes do corpo. Assim sendo, a tarefa de RLS compreende dois desafios. O mais delicado decorre da heterogeneidade intraclasse, isto é, as diferentes execuções atribuídas a um sinal com o mesmo significado. O segundo desafio resulta da similaridade interclasses, ou seja, do alto grau de semelhança que pode existir entre sinais com diferentes significados. Além disso, essas semelhanças aumentam devido às dissimilaridades intraclasse anteriormente citadas. O problema se agrava quando considerados os sinais polissêmicos, cuja execução é idêntica, sendo o significado dependente do contexto.

Segundo Munib et al. (2007), há quatro componentes essenciais em sistemas de reconhecimento de gestos: modelagem de gesto, análise de gesto, reconhecimento de gesto e aplicações baseadas em gestos. As redes neurais profundas se destacam atualmente como técnicas fundamentais em tarefas de modelagem de objetos e de análise e reconhecimento de gestos e ações. Por esse motivo, este trabalho utiliza redes neurais profundas para reconhecimento de sinais da Libras.

Dentre os trabalhos existentes na literatura, a maioria não realiza o reconheci-

mento de sinais dinâmicos, ou seja, sinais que precisam ser executados com movimento para que o seu significado seja entendido corretamente. Em alguns trabalhos que realizam classificação de sinais dinâmicos, a aquisição dos dados é realizada de forma intrusiva, com o uso de luvas, fato que pode ser considerado uma desvantagem em relação aos trabalhos que utilizam apenas visão computacional. Além disso, dentre os métodos que classificam sinais dinâmicos, poucos aplicam métodos de modelagem temporal com redes neurais. Ainda mais reduzida é a quantidade de trabalhos que empregam a combinação de várias entradas, seja na forma de fusão de diferentes canais de dados, ou fusão de modalidades diferentes de dados, como dados RGB, de profundidade, de fluxo ótico — estimativa usada para mensurar o movimento em uma cena, melhor definida na Seção 2.4 — e pontos-chave das articulações.

Desse modo, um dos maiores desafios no RLS é tratar o aspecto dinâmico do gesto. Alguns trabalhos recentes tentam explorar a capacidade das redes neurais profundas seguindo a arquitetura tradicional no reconhecimento de sinais, enviando como entrada dados de movimento previamente quantificado. Outra estratégia empregada envolve o uso de modelos temporais como redes neurais recorrentes e redes com memória, as quais capturam relações temporais de longo alcance.

1.3 Justificativa

Considerar o aspecto temporal do sinal é importante porque a maioria dos sinais da Libras é composta por sinais dinâmicos. Entretanto, além de haver poucos trabalhos que envolvem o reconhecimento de sinais dinâmicos da Libras, as soluções propostas tratam uma quantidade muito reduzida de sinais. Sendo assim, a ampliação do conjunto de sinais reconhecidos já pode ser considerada como um avanço para a área. Este trabalho envolve o reconhecimento de sinais de uma base de dados composta por uma quantidade de classes muito superior aos conjuntos utilizados pelos trabalhos que compõem a literatura.

Também há na literatura soluções que envolvem o reconhecimento de línguas de sinais de outros países, em que foram identificadas melhorias nas técnicas empregadas, bem como na quantidade de sinais dinâmicos reconhecidos. Contudo, é difícil validar tais soluções no contexto de sinais da Libras, visto que os modelos utilizados não estão disponíveis publicamente e são de difícil reprodutibilidade.

A dimensão temporal em sequências normalmente faz com que o reconhecimento de gestos e ações seja um problema desafiador em termos de quantidade de dados a serem processados e complexidade do modelo. Entretanto, estudos recentes têm

apresentado avanços no processamento de imagem e vídeo. Embora as redes neurais convolutivas, do inglês *Convolutional Neural Network* (CNN), tenham sido projetadas, principalmente, para processamento espacial, elas são utilizadas com êxito para entradas de natureza sequencial, como texto e áudio. Esse sucesso das CNN em tarefas de extração e representação de características de imagens, associado ao êxito obtido por essas arquiteturas na tarefa de modelagem implícita de características temporais de vídeos realizada por modelos de CNN tridimensional (3D-CNN), são as principais razões para que esse tipo de rede seja considerado estado da arte no reconhecimento de gestos e ações humanas (Asadi-Aghbolaghi et al., 2017). Porém, poucas foram as soluções encontradas para o reconhecimento de sinais da Libras que utilizam modelos estado da arte de 3D-CNN.

Outro ponto importante a mencionar é que, apesar da base utilizada nesta dissertação possuir uma quantidade de classes significativamente superior às bases utilizadas em trabalhos relacionados, seria necessário que a base de dados fosse composta por milhares de instâncias por classe para possibilitar o uso de modelos de redes neurais profundas treinados do zero. Logo, também é necessário explorar diferentes estratégias de transferência de aprendizado de modelos pré-treinados em contextos similares, como por exemplo, em tarefas de reconhecimento de gestos e ações.

Além da transferência de aprendizado, a literatura que envolve aprendizado de máquina indica outra alternativa para contornar o problema da insuficiência de instâncias: o aumento artificial de dados, técnica que visa a ampliação do conjunto de treinamento mediante a aplicação de transformações geométricas e fotométricas em imagens. Tal técnica é pouco explorada na área de reconhecimento de gestos e ações em vídeo, tampouco foram encontrados trabalhos que a apliquem no contexto de reconhecimento de sinais da Libras.

Ademais, também não foram encontrados trabalhos envolvendo Libras que enfrentam o problema de reconhecimento a partir de uma abordagem multicanal que utilize dados RGB e de fluxo ótico. Neste trabalho, a proposta é explorar melhor o esse aspecto multicanal e a forma como essas informações se complementam.

A hipótese é a de que a combinação de técnicas de transferência de dados, aumento artificial de dados e fusão de diferentes canais de dados, pode melhorar os resultados em RLS. Portanto, o propósito deste trabalho é identificar o cenário em que modelos de 3D-CNN podem apresentar altas taxas de precisão, quando aplicados ao reconhecimento de sinais estáticos e dinâmicos da Libras, utilizando informações sobre dimensões espaciais e temporais. Os cenários se diferenciam conforme os canais de dados recebidos como entrada, a estratégia de transferência de aprendizado utilizada e o nível em que a fusão de dados é aplicada.

1.4 Objetivo Geral

O objetivo geral deste trabalho é identificar experimentalmente uma estratégia para reconhecimento de sinais estáticos e dinâmicos da Libras que utilize informações sobre dimensões espaciais e temporais, por meio de 3D-CNN, fusão de dados de múltiplos canais e transferência de aprendizado.

1.4.1 Objetivos Específicos

Para atingir o objetivo geral, pretende-se alcançar os seguintes objetivos específicos:

1. Identificar a melhor estratégia para transferência de aprendizado a partir de comparação entre modelos gerados para representação de características de problemas como reconhecimento de ações ou de gestos;

2. Identificar o cenário que apresenta melhores resultados para classificação de sinais da Libras utilizando vídeos representados por sequências de imagens RGB e fluxo ótico, separadamente;

3. Elevar as taxas de precisão ao fundir informações de diferentes canais, utilizando vídeos de imagens RGB e dados do fluxo ótico em conjunto;

1.5 Metodologia

Para obter-se uma visão geral do cenário atual em reconhecimento de sinais da Libras, foi realizada uma revisão da literatura, compreendendo trabalhos publicados entre 2016 e 2019. Em seguida, foi realizada uma revisão da literatura, conduzida com base na estrutura proposta por Kitchenham e Charters (2007), envolvendo trabalhos de RLS de outros países. Além disso, foi realizada uma pesquisa adicional para visualizar o progresso dos estudos que envolvem reconhecimento de gestos — não específicos de línguas de sinais — e ações.

A partir das revisões da bibliografia foi identificado que, dentre outras abordagens, os trabalhos mais recentes utilizam dois métodos principais. O primeiro é a modelagem da sequência temporal, para a qual as arquiteturas combinam modelos de redes bidimensionais com modelos de sequência temporal, como a cadeia de Markov ou redes neurais recorrentes. Outros trabalhos, que realizam o reconhecimento de gestos e ações, utilizam modelos de 3D-CNN. Tipicamente, para ambas as abordagens, a

entrada do modelo é obtida a partir de quatro fontes de dados: imagens RGB, imagens com informações de profundidade, vetores com informações sobre o esqueleto do emissor e imagens com informação de fluxo ótico.

Com base nessas informações, definiu-se para ser usado neste trabalho um modelo de 3D-CNN para reconhecimento do sinal, e dados RGB e de fluxo ótico para alimentar a rede neural. Sendo assim, durante o desenvolvimento desta pesquisa foram executadas as seguintes etapas:

1. Seleção do modelo e ajuste de parâmetros: a primeira atividade realizada consiste na seleção do modelo de rede neural utilizado — o Inflated-3D (Carreira e Zisserman, 2017), modelo referência em reconhecimento de ações. O segundo passo é o ajuste de hiperparâmetros desse modelo, para que o melhor conjunto de parâmetros seja encontrado, implicando em boas taxas de classificação. Essa etapa é realizada utilizando uma amostra da base de dados principal, composta por 50 classes de sinais. Durante essa etapa, também são identificadas as melhores formas de transferência de aprendizado e ajuste.

2. Comparação de resultados ao *baseline*: é realizada uma etapa de experimentos para validação dos resultados, frente aos resultados obtidos pelo trabalho de Machado (2018) ao utilizar uma amostra da base de dados principal, composta por 84 classes.

3. Validação de hiperparâmetros na base completa: os hiperparâmetros que obtiveram os melhores resultados na amostra de 84 classes da base são, então, utilizados em experimentos executados na base completa.

4. Aumento de dados: os experimentos com a base resultante do aumento de dados são iniciados.

5. Fusão de dados: por último, as previsões resultantes de modelos individuais são fundidas para que se resulte em uma contribuição mútua entre os diferentes canais de dados.

1.6 Contribuições

As principais contribuições deste trabalho são:

1. Especificação de uma estratégia para reconhecimento de sinais estáticos e dinâmicos da Libras que combine 3D-CNN, fusão de dados multicanais e transferência de aprendizado;

2. Disponibilização de um modelo pré-treinado em dados RGB, capaz de reconhecer 560 sinais da Libras;

3. Disponibilização de um modelo pré-treinado em dados de fluxo ótico, capaz de reconhecer 560 sinais da Libras;

1.7 Estrutura do Trabalho

O restante deste trabalho está organizado da seguinte maneira: o Capítulo 2 apresenta a fundamentação teórica necessária para o entendimento dos componentes da abordagem proposta. O Capítulo 3 expõe uma síntese de trabalhos recentes relacionados ao RLS que utilizam modelos de aprendizagem profunda, discutindo e apresentando um comparativo entre os trabalhos, além dos avanços obtidos no reconhecimento de ações em vídeo. No Capítulo 4 é apresentada a abordagem proposta. Os resultados dos experimentos são apresentados no Capítulo 5. Por fim, no Capítulo 6 são feitas as considerações finais desta dissertação.

Capítulo 2

Fundamentação Teórica

Este capítulo aborda os conceitos fundamentais para o entendimento e desenvolvimento das partes integrantes da abordagem proposta. São apresentados os conceitos universais de línguas de sinais e os específicos da língua de sinais brasileira. Também são analisados os métodos utilizados para aquisição dos dados em aplicações de RLS. Na sequência, são explicados os conceitos teóricos relacionados às redes neurais artificiais (RNA), com ênfase aos conceitos relacionados às redes neurais convolutivas, na Seção 2.3.1. Por fim, são descritos conceitos sobre fusão de dados, transferência de aprendizado, e aumento de dados. Além disso, para se entender as análises de resultados obtidos a partir dessas combinações, algumas métricas de avaliação, comumente utilizadas para avaliar o desempenho de modelos de redes neurais, são conceituadas.

2.1 Língua Brasileira de Sinais

A língua é um sistema de signos compartilhado por uma comunidade linguística comum, sendo a fala ou os sinais expressões de diferentes línguas. As línguas de sinais são utilizadas pela comunidade surda como forma de comunicação e expressão (Quadros, 2004), variando de acordo com o país. No Brasil, a língua de sinais utilizada pela comunidade surda é mais conhecida como Libras, difundida pela Federação Nacional de Educação e Integração de Surdos (FENEI). Entretanto, há também o termo Língua Brasileira de Sinais (LBS), que segue o padrão internacional para nomenclatura das línguas de sinais (Quadros, 2004).

Após estudos iniciados por Stokoe (1960), foram descobertas evidências de que as línguas de sinais possuem aspectos linguísticos equivalentes aos de línguas orais, compartilhando as restrições aplicadas às línguas faladas. Portanto, apesar de serem utilizados recursos diferentes de expressão, as línguas de sinais são estruturadas formal-

mente, possuindo parâmetros de execução, e reconhecidas pela linguística como língua na modalidade visuoespacial.

2.1.1 Parâmetros da Libras

Os trabalhos de Stokoe (1960) foram os pioneiros no tratamento linguístico às línguas de sinais, analisando aspectos morfológicos e fonológicos da Língua de Sinais Americana (LSA). Anos depois, outros estudos foram realizados por Klima e Bellugi (1979) sobre os atributos dos sinais. Como resultado das análises, foram definidos parâmetros para formação de sinais da LSA que são comuns à maioria das línguas de sinais existentes, incluindo a Libras.

Os sinais da Libras são formados por parâmetros que podem depender de aspectos manuais e expressões, faciais e/ou corporais (Felipe e Monteiro, 2007), que serão descritos a seguir.

2.1.1.1 Configuração das mãos

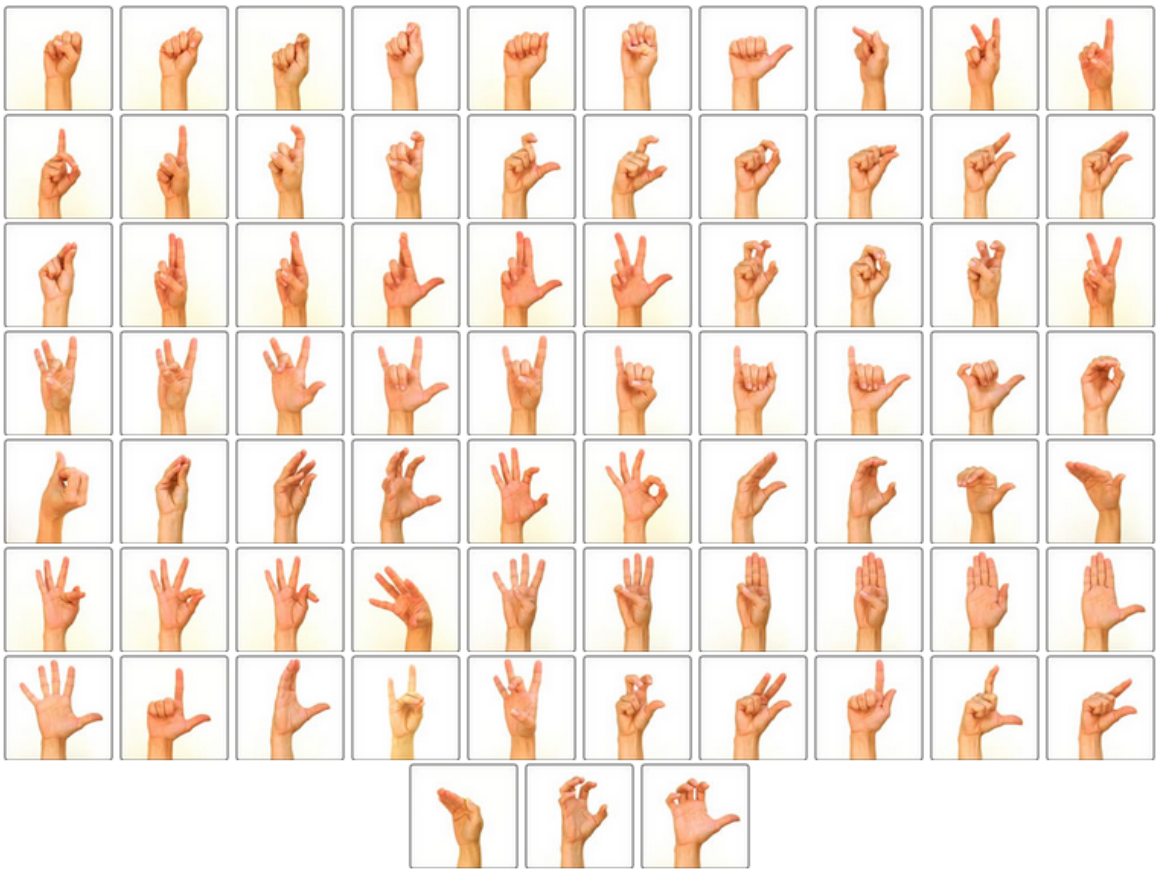
Essa característica está relacionada ao aspecto de uma ou duas mãos durante a execução do sinal. A Libras possui até 73 configurações[1] que podem ser feitas pela mão dominante ou pelas duas mãos, dependendo do sinal. Há configurações muito semelhantes, conforme apresentado na Figura 2.1, extraída do Dicionário Digital da INES[2]. Algumas letras do alfabeto são representadas por configurações. Por exemplo, na Figura 2.2, os sinais APRENDER, SÁBADO e DESODORANTE-SPRAY têm a mesma configuração de mão — que também representa a letra "s" — e são realizados na testa, na boca e na axila, respectivamente (Felipe e Monteiro, 2007). A configuração da mão também pode ser um marcador de gênero animado (pessoa e animais) ou inanimado (coisas) (Felipe, 2007).

2.1.1.2 Ponto de articulação

O ponto de articulação é a região onde incide a mão dominante (Stokoe Jr, 2005). A Figura 2.3 apresenta os sinais TRABALHAR, BRINCAR, PAQUERAR, que são feitos no espaço neutro vertical (do meio do corpo até à cabeça) e horizontal (à frente do emissor), e os sinais ESQUECER, APRENDER e DECORAR, que fazem parte dos sinais realizados com toque em alguma parte do corpo, que nesse exemplo, são realizados na testa (Felipe e Monteiro, 2007). O ponto de articulação também pode ser uma marcação de concordância verbal com o advérbio de lugar (Felipe, 2007).

[1] http://www.acessibilidadebrasil.org.br/libras_3/
[2] http://www.ines.gov.br/dicionario-de-libras/main_site/libras.htm

Figura 2.1: Configurações de mão extraídas do Dicionário Digital da INES.

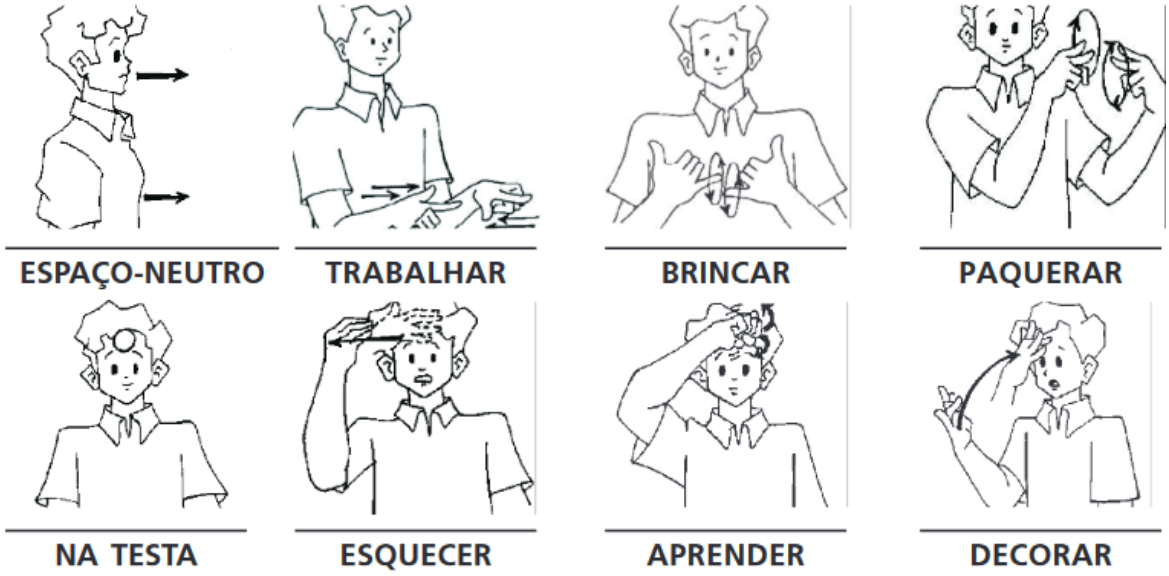


Fonte: Lira e Souza (2018).

Figura 2.2: Exemplo do uso da mesma configuração de mãos para emitir o sinal da letra “s” e de três diferentes palavras.

Fonte: Felipe e Monteiro (2007).

Figura 2.3: Representação do parâmetro “ponto de articulação”.



Fonte: Felipe e Monteiro (2007).

2.1.1.3 Movimento

Como mencionado, os sinais podem ter um movimento (sinais dinâmicos) ou não (sinais estáticos). Na Figura 2.4 são ilustrados outros exemplos de sinais com movimento (RIR, CHORAR e CONHECER), além de dois sinais (AJOELHAR e EM-PÉ) não têm movimento (Felipe e Monteiro, 2007). Esse parâmetro pode indicar uma raiz verbal na palavra e a alteração na frequência do movimento pode ser uma marca de aspecto temporal, de advérbio de modo ou um intensificador (Felipe, 2007).

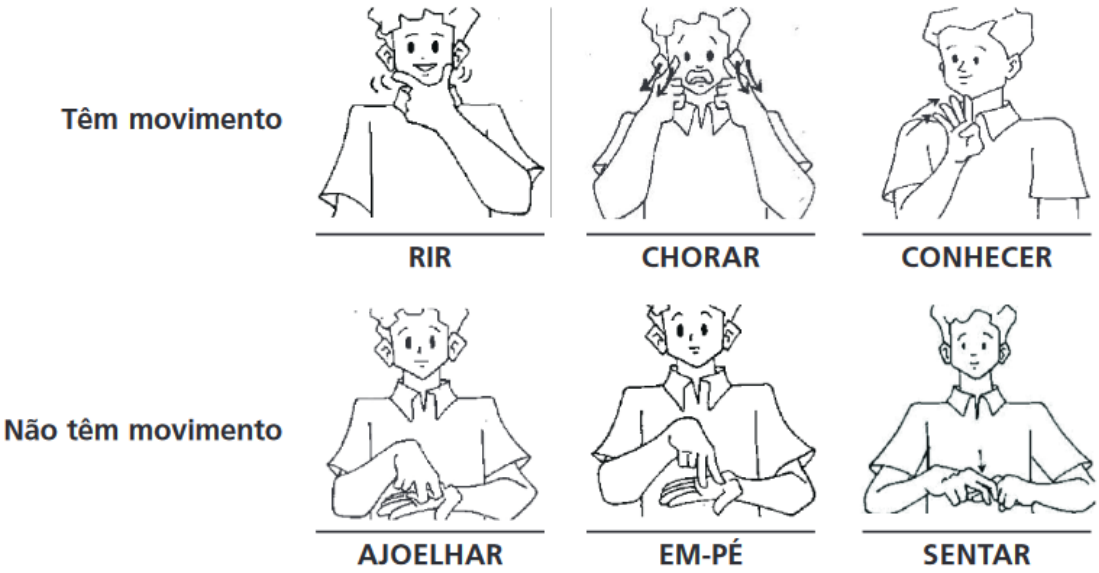
2.1.1.4 Orientação ou direcionalidade

Os sinais também podem possuir orientação ou direcionalidade definida com relação aos parâmetros (Stokoe, 1960). Na Figura 2.5 são mostrados como exemplo os verbos IR e VIR, SUBIR e DESCER, ACENDER e APAGAR, ABRIR-PORTA e FECHAR-PORTA, que possuem a mesma configuração de mãos, mas diferem na direcionalidade (Felipe e Monteiro, 2007).

2.1.1.5 Expressões

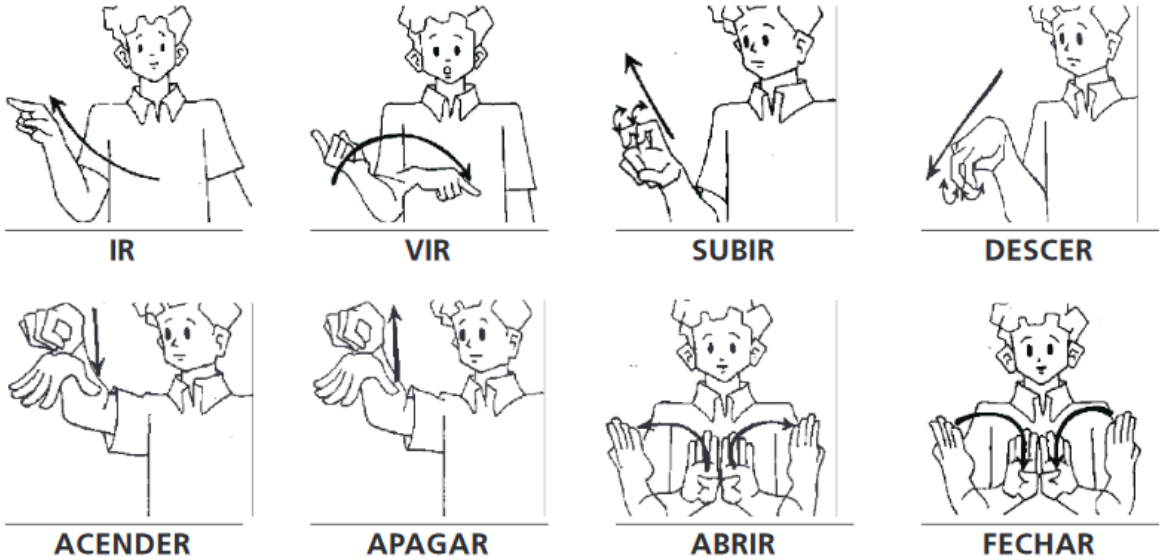
Além dos parâmetros relacionados ao aspecto manual, muitos sinais possuem como traço diferenciador a expressão facial e/ou a expressão corporal (Klima e Bellugi, 1979).

Figura 2.4: Representação do parâmetro “movimento”.



Fonte: Felipe e Monteiro (2007).

Figura 2.5: Representação do parâmetro “orientação”.

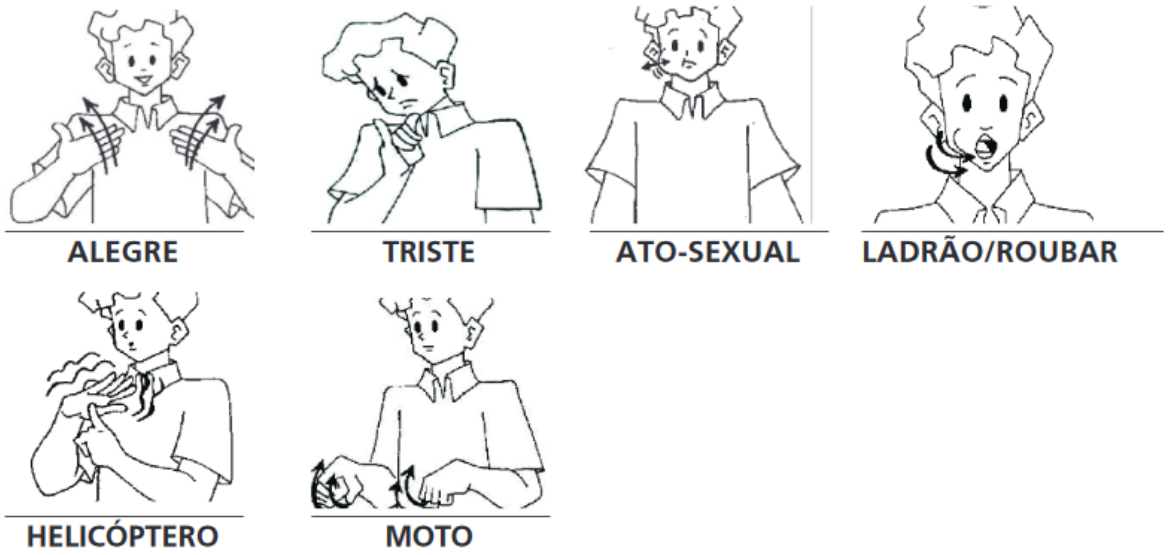


Fonte: Felipe e Monteiro (2007).

Apesar de esse parâmetro não ser comum à todas as línguas de sinais, é parte integrante da Libras. Na Figura 2.6 pode ser observado que os sinais da Libras ALE-

GRE e TRISTE possuem tanto expressão facial como corporal; os sinais LADRÃO, RELAÇÃO-SEXUAL são executados somente com a bochecha; há também sinais em que sons e expressões faciais são complementos dos traços manuais, como os sinais HELICÓPTERO e MOTO.

Figura 2.6: Representação do parâmetro "expressão".



Fonte: Felipe e Monteiro (2007).

2.1.2 Tipos de Sinais

É importante frisar que a Libras possui duas categorias de sinais, que se diferenciam pelo parâmetro "movimento":

- os sinais estáticos: os quais são independentes dos parâmetros que envolvem movimentação. A maioria deles engloba as configurações de mãos, letras, alguns números, e poucas palavras;
- os sinais dinâmicos: que são executados com a movimentação das mãos, sendo a maioria dependente de todos os parâmetros. A Libras é majoritariamente composta por sinais pertencentes a essa categoria;

No âmbito técnico relacionado ao RLS, em uma base de dados, para os sinais estáticos da Libras, a representação pode ser feita por imagens individuais, enquanto que para os sinais dinâmicos, a representação deve ser feita por vídeos ou sequências de imagens que representem a trajetória do sinal.

Pesquisadores da área de linguística da Libras reforçam a necessidade da clara identificação e entendimento dos parâmetros das línguas de sinais, pois estão diretamente relacionados ao significado do sinal. Alguns sinais são muito semelhantes, entretanto, com pequenas diferenças, chamadas de "contraste semântico" (Xavier e Barbosa, 2014), que dependem da correta execução e interpretação dos parâmetros.

Portanto, um sistema de RLS deve ser robusto em duas tarefas principais: a de aquisição dos dados, sendo capaz de capturar todos, ou a maioria, dos parâmetros que compõem o sinal; e de reconhecimento e classificação dos sinais, tratando corretamente as similaridades interclasses, ou seja, as semelhanças que podem existir entre sinais com diferentes significados.

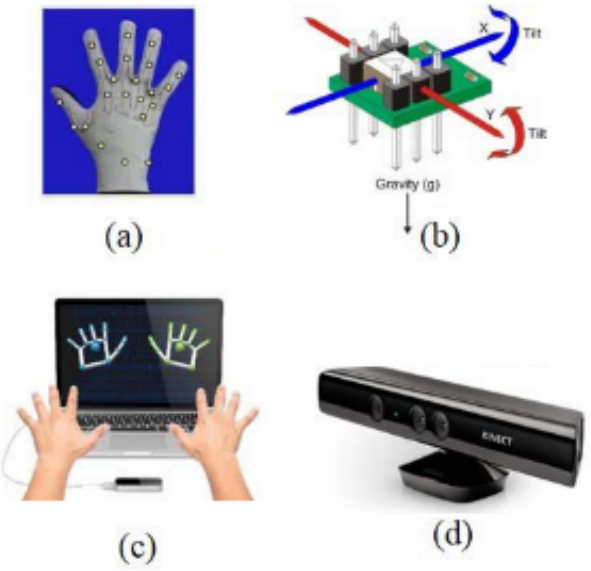
2.2 Dispositivos para Aquisição de Dados

Em sistemas de RLS, dois tipos de abordagens podem ser utilizados para aquisição das imagens ou vídeos: as baseadas em sensores portáteis e as baseadas em visão. A Figura 2.7 apresenta alguns exemplos de dispositivos. Na primeira categoria, dispositivos com sensores magnéticos, óticos, ou acústicos são atrelados ao corpo do usuário para capturar informações como movimento, posição e velocidade das mãos e/ou braços. A principal vantagem dessa abordagem é que esses dispositivos são mais precisos na captura dos dados, podendo dispensar tarefas como segmentação e rastreamento das mãos e corpo, facilitando o processo de reconhecimento do sinal.

Entretanto, pelo fato desses sensores serem, geralmente, embarcados em luvas (Mohandes et al., 2017) ou pulseiras (Shin et al., 2017), é exigido que o usuário porte o dispositivo, o que caracteriza o método como intrusivo. Consequentemente, pode haver certa dificuldade em extrair do usuário uma interação de maneira natural, além de que é frequentemente necessária a calibração do equipamento. Em suma, além do alto custo, são necessárias muitas intervenções do usuário, gerando uma grande desvantagem desses modelos (Zheng et al., 2017).

A segunda categoria, em contrapartida, realiza a aquisição por meio de câmeras de vídeo, tendo despontado após o lançamento de dispositivos que fornecem mapas de profundidade, como o *Kinect* e o *Leap Motion Controller* (LMC), embora câmeras RGB comuns ainda sejam utilizadas. As aplicações baseadas somente em visão computacional não são intrusivas, de forma que qualquer processo de rastreamento de movimento e posição das mãos e do corpo é feito via algoritmo, sem a necessidade da mediação do usuário: o sistema assume que o emissor inicia o gesto posicionado em frente ao dispositivo.

Figura 2.7: Dispositivos usados em reconhecimento de língua de sinais: (a) luvas, (b) acelerômetros, (c) *leap motion controller*, (d) *Kinect*.



Fonte: Adaptado de Zheng et al. (2017).

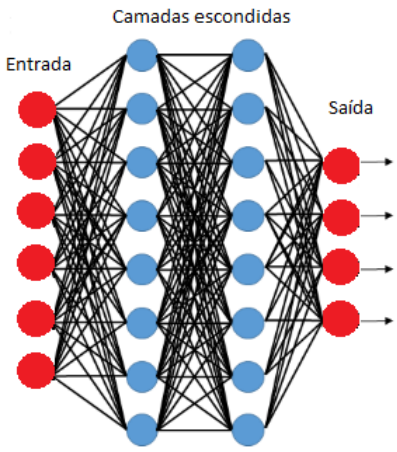
Antes de 2010, as principais desvantagens dos métodos baseados em visão eram relacionadas ao processamento digital de imagem, sobretudo nas fases iniciais de pré-processamento e segmentação da imagem (Gonzalez et al., 2004). Após 2010, entretanto, os avanços dos estudos em RNA e aprendizado profundo (LeCun et al., 2015) trouxeram muitos benefícios para a área de reconhecimento de imagem e vídeo, com as redes convolutivas, e de texto e fala, com as redes recorrentes (Goodfellow et al., 2016). Consequentemente, despontaram pesquisas em reconhecimento de gestos e ações, e com elas, os trabalhos de sistemas de RLS. O aprendizado profundo, portanto, tornou-se dominante como estado da arte em reconhecimento de fala, de ações, de gestos, e em muitos outros domínios (LeCun et al., 2015).

2.3 Redes Neurais Artificiais

Pesquisas em RNA iniciaram devido ao reconhecimento de que o cérebro humano processa informações diferentemente dos computadores convencionais, podendo ser considerado como um computador mais complexo, não linear, paralelo, e mais rápido (Haykin, 2007). Desse modo, redes neurais artificiais são modelos computacionais ins-

pirados no sistema nervoso central animal, capazes de reconhecer padrões e aprender por meio de dados e experiência (Goodfellow et al., 2016). Esses modelos seguem uma hierarquia, apresentada na Figura 2.8, composta por várias camadas de neurônios (também conhecidos como unidades ou elementos) conectados. Os nós representam os neurônios e as linhas representam os pesos das conexões. Por convenção, a camada que recebe os dados é chamada camada de entrada, a final é chamada camada de saída, e as internas são chamadas de camadas escondidas. Redes neurais que possuem muitas camadas escondidas são chamadas de redes neurais profundas. Nesta seção, serão apresentados os principais modelos de redes neurais profundas utilizados para detecção e reconhecimento em imagem e vídeo.

Figura 2.8: Representação de uma rede neural artificial.



Fonte: Próprio Autor.

2.3.1 Redes Neurais Profundas

Aprendizado profundo é a técnica utilizada em redes neurais profundas que permite que esses modelos, compostos por múltiplas camadas de processamento, aprendam representações de dados com vários níveis de abstração. De acordo com Goodfellow (2016), essa hierarquia de conceitos permite que o computador aprenda conceitos complexos, que são construídos a partir de conceitos mais simples. Isso significa que o conhecimento é adquirido por meio da experiência, não sendo necessário que um ser humano especifique explicitamente toda a informação para que a máquina atinja o objetivo.

Uma rede neural *feedfoward* é um tipo de rede, cujo objetivo é aproximar uma função, em que o processamento da informação flui em um único sentido ao avaliar a entrada, ou seja, dada uma camada, não há conexões de retroalimentação de suas saídas com as camadas anteriores (Goodfellow et al., 2016).

Componentes deste vasto arcabouço de métodos, as redes convolutivas profundas, descritas com mais detalhes na próxima subseção, trouxeram avanços no processamento de imagens, vídeo, e áudio, enquanto as redes recorrentes têm contribuído nas áreas de processamento de séries de tempo, como texto e fala (Goodfellow et al., 2016).

2.3.1.1 Redes Neurais Convolutivas

As redes neurais convolutivas, também conhecidas como redes neurais convolucionais, *ConvNets* ou CNN, são um tipo especializado de rede neural *feedfoward* para processamento de dados que possuem uma topologia semelhante a uma grade (Goodfellow et al., 2016). LeCun et al. (2015) citam alguns exemplos de representação desses tipos de dados: as imagens, que são organizadas em topologia 2D; os vídeos, que apresentam topologia 3D com imagens ao longo do tempo; e sequências de sinais de áudio, organizadas em topologia 1D.

A convolução é uma das principais técnicas de processamento digital de imagens, baseada na Transformada de Fourier. Dadas as funções $f(x)$ e $g(x)$, denotamos a convolução $f(x) * g(x)$ como a nova função $h(x)$, conforme a equação 2.1 (Gonzalez et al., 2004), onde a convolução no instante α é equivalente à área da intersecção entre $f(x)$ e $g(\alpha - x)$.

$$h(x) = f \otimes g = \int_{-\infty}^{\infty} f(\alpha)g(x - \alpha)d(\alpha), \tag{2.1}$$

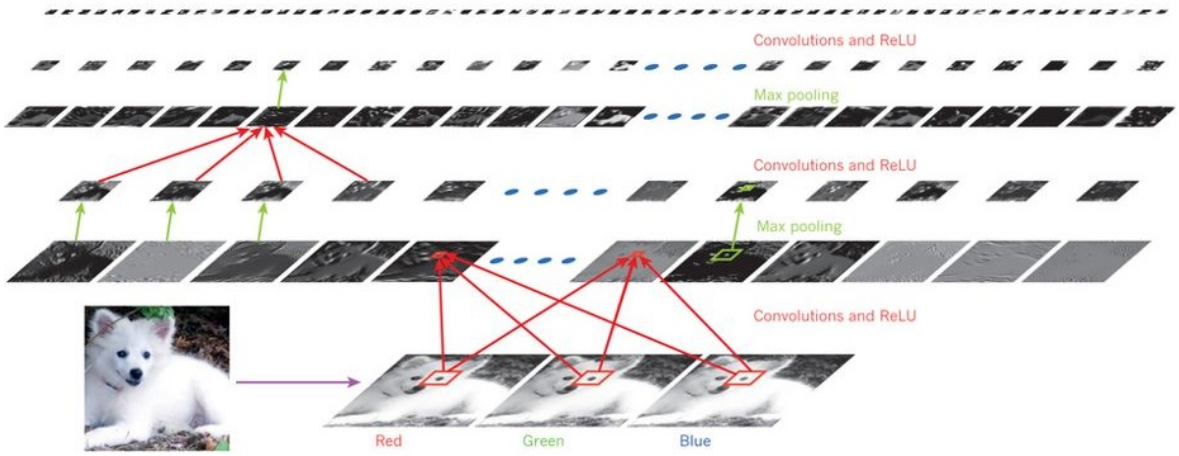
Em termos matemáticos, a convolução é um operador linear que, a partir de duas funções dadas, resulta em uma terceira que mede a soma do produto dessas funções ao longo da região subentendida pela superposição delas em função do deslocamento existente entre elas (Bracewell, 1986). As CNN são simplesmente redes neurais que, ao invés de realizar multiplicações de matrizes gerais, usam a operação matemática de convolução da matriz de entrada com o filtro (Goodfellow et al., 2016).

Segundo LeCun et al. (2015), uma rede convolutiva clássica é estruturada como uma pilha de camadas. A Figura 2.9 ilustra essa estrutura, na qual pode ser visto que, inicialmente, são empilhadas as camadas convolucionais, seguidas por camadas de agrupamento, mais conhecidas como camadas de *pooling*. Duas operações comumente aplicadas na camada de *pooling* são o cálculo do maior conjunto local e o cálculo da média dos valores do conjunto local. Isso faz com que a quantidade de parâmetros que

percorrem a rede seja reduzida, melhorando a eficiência computacional da arquitetura (Goodfellow et al., 2016). Na maioria das vezes, são adicionadas camadas densamente conectadas ao final da rede.

Como pode ser observado na ilustração da Figura 2.9, a primeira camada de uma rede convolutiva recebe como entrada as imagens, que podem ter tamanhos diferentes, mas aproximados. Cada elemento de uma camada recebe o resultado de uma operação realizada sobre um conjunto de unidades vizinhas da camada anterior. Como a segunda camada, a de convolução, é composta por diferentes mapas de características, com múltiplos tensores de pesos, várias características podem ser extraídas em cada coordenada. Conforme explanado por LeCun et al. (1995), a amostragem que segue as camadas de convolução e a média local reduz a resolução do mapa de características e mantém a consistência da saída, que se torna menos sensível às distorções.

Figura 2.9: Arquitetura clássica de redes convolutivas.



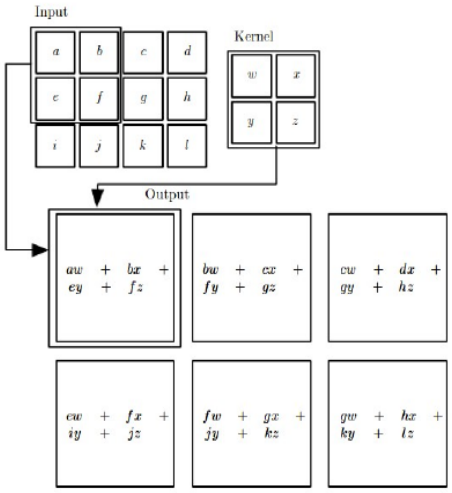
Fonte: Adaptado de LeCun et al. (2015).

Conforme pode ser visualizado na Figura 2.10, em redes convolutivas há, geralmente, dois tensores como entrada: a primeira é a matriz multidimensional de entrada, e a segunda, chamada de filtro, é uma matriz multidimensional que será adaptada pelo algoritmo. Matsugu et al. (2003) explicam que a estrutura geral das redes convolucionais é inspirada no funcionamento do córtex visual humano, que possui uma grande quantidade de células responsáveis pela detecção de luz em pequena escala. Tais células atuam como filtros locais sobre o espaço de entrada, que sobrepõem as regiões do campo visual, chamadas de campos receptivos.

As camadas convolutivas possuem os campos receptivos locais, que são regiões

menores da imagem que fornecem informações localizadas. Em uma camada completamente conectada, cada neurônio está conectado a todos os neurônios da camada anterior. No caso das redes convolutivas, cada elemento de uma camada está conectado somente a um conjunto de campos receptivos locais. Os campos receptivos de diferentes neurônios se sobrepõem parcialmente de forma a cobrir todo o campo de visão, permitindo que a rede aprenda características locais distintas. Esses elementos são organizados em mapas de filtros, de maneira que os elementos de cada mapa compartilham os mesmos pesos, consequentemente, a quantidade de pesos necessários por camada é reduzida. Essa organização permite que a rede aprenda características locais distintas ao longo dos campos receptivos, ao mesmo tempo que reduz a quantidade de pesos necessários por camada (LeCun et al., 2015).

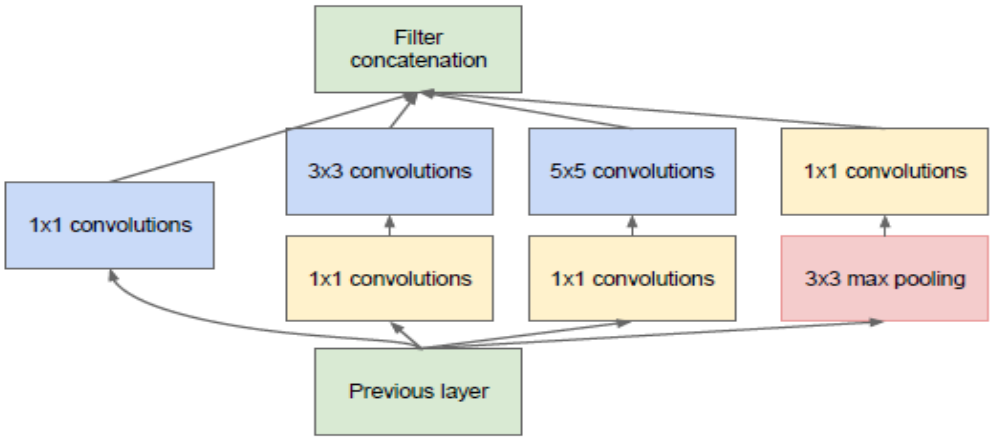
Figura 2.10: Operação de convolução.



Fonte: Goodfellow et al. (2016).

Outra grande vantagem de uma CNN é a independência de um conhecimento anterior e do esforço humano no desenvolvimento de suas funcionalidades básicas: o modelo exige um nível mínimo de esforço na fase de pré-processamento, quando comparado a algoritmos de classificação de imagens, visto que a rede aprende os filtros que em um algoritmo tradicional precisariam ser implementados manualmente. Consequentemente, o uso das CNN para reconhecimento de imagens elimina a necessidade de criar extratores de características manuais, assim como, na maioria dos casos, o trabalho de normalizar o tamanho das imagens e orientação (LeCun et al., 1995). Ademais, quando apropriadamente regularizadas, apresentam resultados muito satisfatórios em tarefas de reconhecimento de objetos (JI et al., 2013).

Figura 2.11: Estrutura do módulo *Inception* com redução de dimensão.



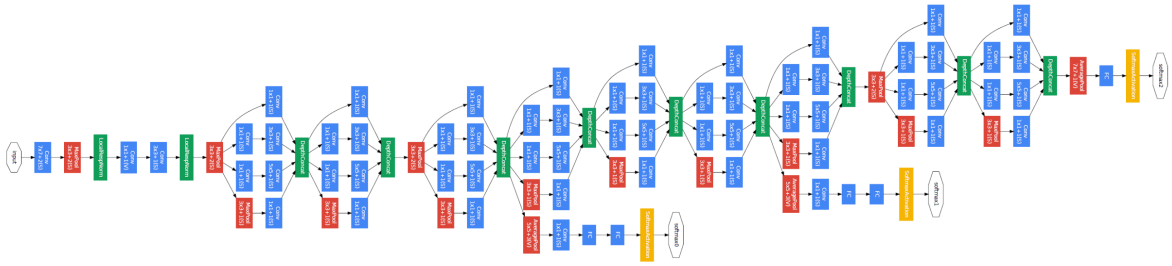
Fonte: Szegedy et al. (2015).

Diante da alta capacidade de desempenho das CNN, algumas arquiteturas foram projetadas para solucionar diversos tipos de problemas, sobretudo em tarefas relacionadas à área de visão computacional. Neste trabalho, pretende-se utilizar um modelo adaptado da arquitetura InceptionV1. Em 2015, a InceptionV1 (também conhecida como GoogLeNet) venceu o desafio *ImageNet Large-Scale Visual Recognition Challenge* (ILSVRC) de classificação de 1000 classes de imagens (Szegedy et al., 2015). O modelo, que utiliza técnicas introduzidas por Lin et al. (2013), é composto por convoluções 1x1, que são usadas para reduzir a dimensionalidade, fazendo com que a computação também seja reduzida. Ao final, há uma camada de agrupamento utilizando a média global. A arquitetura também introduz o módulo *Inception*, apresentado na Figura 2.11, que executa diferentes tamanhos de convolução em paralelo para a mesma entrada. Ao longo da arquitetura são empilhados vários destes módulos.

Em arquiteturas apresentadas anteriormente, cada camada utilizava um tamanho fixo para convolução. Na InceptionV1, são utilizados diferentes tamanhos de convoluções, seguidas pela camada de *pooling*. Os diferentes tipos de características são comprimidos em mapas que são concatenados e enviados para a entrada do próximo módulo. A arquitetura completa pode ser visualizada na Figura 2.12.

Como mencionado, em tarefas que envolvem reconhecimento em vídeo, podem ser utilizados os modelos de CNN tridimensionais (3D-CNN), arquiteturas capazes de realizar convoluções considerando entradas em que o contexto temporal ou volumétrico é importante (Maturana e Scherer, 2015). Portanto, neste trabalho, será utilizada uma versão tridimensional da Inception-V1, proposta por Carreira e Zisserman (2017), detalhada na seção 3.3.

Figura 2.12: Arquitetura completa da InceptionV1.



Fonte: Szegedy et al. (2015).

2.4 Fusão de Dados

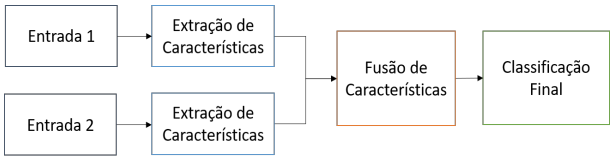
Outro ponto a ser destacado é que as CNN têm se apresentado eficientes quando são combinados diferentes tipos de entradas que podem, ou não, ser obtidos por meio de sensores diferentes. A integração de várias mídias, a associação de suas características ou das decisões intermediárias para executar uma tarefa de análise é chamada de fusão multimodal. A fusão de múltiplas modalidades pode fornecer informações complementares e aumentar a precisão do processo geral de tomada de decisão (Atrey et al., 2010). Podem ser utilizadas informações de áudio e vídeo, por exemplo. Nesses casos, entretanto, o benefício da fusão multimodal é acompanhado de um aumento da complexidade no processo de análise, pois envolve as diferentes características das modalidades, as quais podem requerer diferentes tipos de arquiteturas de rede.

Um alternativa que pode ser utilizada no processamento de sequências de imagens é a abordagem multicanal, que consiste na utilização de diferentes canais das imagens de entrada. Conforme destacado por Narayana et al. (2019), há na literatura uma forte tendência quanto à utilização de arquiteturas multicanais, em que duas ou mais CNN processam versões diferentes do mesmo vídeo em paralelo. Geralmente, na pesquisa de reconhecimento de gestos via visão computacional, são utilizados dados dos canais RGB, de fluxo ótico, de profundidade e infravermelho, por exemplo.

Ao se fazer uso de múltiplas modalidades ou múltiplos canais, deve ser levada em consideração a estratégia de fusão que será utilizada. A literatura apresenta diversas estratégias para fusão de dados, dentre as quais duas se destacam: a fusão de dados antecipada ou inicial (ou fusão em nível de características) e a fusão de dados tardia (ou fusão em nível de decisão) (Kopuklu et al., 2018). As abordagens que se baseiam na fusão antecipada seguem o esquema ilustrado na Figura 2.13: extraem separadamente as características de cada modalidade ou canal e, em seguida, os vetores de características

são combinados em uma única representação, antes da etapa de classificação. A maior desvantagem desse esquema é a dificuldade de combinar características diferentes em uma representação comum (Snoek et al., 2005).

Figura 2.13: Esquema de fusão em nível de característica.



Fonte: Adaptado de Snoek et al. (2005).

O processo que utiliza a fusão em nível de decisão é ilustrado na Figura 2.14, em que pode ser visto que o fluxo também inicia com a extração de características unimodais. Porém, essa estratégia é focada no potencial individual de cada modalidade ou canal: os modelos executam individualmente até a etapa de classificação, ou seja, as instâncias são associadas às classes diretamente pelo modelo, a partir de suas características individuais; em contraste com a fusão antecipada, em que as características são combinadas em uma única representação multimodal. Em seguida, as opiniões dos diferentes modelos gerados são combinadas (Kittler et al., 1998; Gunes e Piccardi, 2005), resultando em uma única opinião final. Uma desvantagem dessa abordagem é o custo relativo ao esforço de aprendizado, visto que todas as modalidades precisam passar pela etapa de classificação (Snoek et al., 2005).

Figura 2.14: Esquema de fusão em nível de decisão.

Fonte: Adaptado de Snoek et al. (2005).

Nos trabalhos que envolvem RLS, apesar de serem exploradas várias maneiras de processamento dos dados de entrada, geralmente, quatro tipos de entrada se destacam: as imagens RGB, dados de fluxo ótico, informações de profundidade e pontos das articulações do esqueleto. Os dois últimos, como detalhado na seção 2.2, exigem a utilização de dispositivos específicos de aquisição, enquanto que os dois primeiros podem ser obtidos a partir de dados adquiridos utilizando uma câmera simples, considerando que o fluxo ótico pode ser estimado a partir do RGB.

O fluxo ótico, ou fluxo de imagem, descreve o campo de deslocamento e de velocidade de cada *pixel* de sequências de imagens. Quando os objetos de uma cena se movem dentro de um campo de visão, ocorre uma mudança no padrão de brilho da sequência de imagens (Schunck, 1989). Por meio da estimativa de fluxo ótico, é possível mensurar a direção e a velocidade de objetos em um vídeo. Sendo assim, a técnica é importante quando o movimento no vídeo tem relevância. Há vários métodos para o cálculo do fluxo ótico, sendo todos baseados nos princípios estabelecidos por Horn e Schunck (1981) em seu algoritmo original.

Na literatura, há pouca alternância entre fontes de entrada, combinação de várias entradas, como RGB, profundidade, esqueleto ou articulações, e fluxo ótico, ou mesmo uma análise do impacto que cada tipo de entrada reflete no resultado final. Sendo assim, neste trabalho é empregada a abordagem multicanal, utilizando dados RGB e de fluxo ótico, comparando as estratégias de fusão antecipada e tardia.

2.5 Transferência de Aprendizado

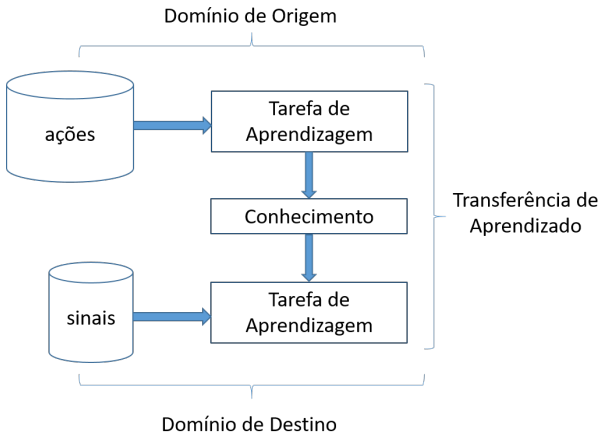
Um grande limitador em aprendizagem profunda é a necessidade de utilização de grandes conjuntos de dados para que o modelo consiga generalizar. Essa dependência de dados de treinamento massivos para entender os padrões latentes de dados é uma das principais desvantagens do aprendizado profundo em comparação com os métodos tradicionais de aprendizado de máquina (Tan et al., 2018). O problema é ainda mais agravado, pois, na maioria dos casos, as bases de dados reais rotuladas não são suficientemente grandes para realizar o treinamento supervisionado.

Assim, uma possibilidade é aplicar o conhecimento adquirido no treinamento (realizado em um conjunto suficientemente grande de dados) de uma outra tarefa, geralmente similar. Denota-se essa alternativa como transferência de aprendizado, em que os dados usados para realizar o treinamento não seguem a mesma distribuição da base de teste ou têm um espaço de características diferente (Pan e Yang, 2010), mas que, por meio de adaptações, podem contribuir para a tarefa alvo do modelo final. Para facilitar o entendimento, esse processo é exemplificado na Figura 2.15. Em termos matemáticos, Pan e Yang (2010) definem transferência de aprendizado da seguinte maneira: dado um domínio de origem D_s e uma tarefa de aprendizagem T_s, um domínio de destino D_t e uma tarefa de aprendizagem T_t, onde $D_s \neq D_t$ ou $T_s \neq T_t$, a transferência de aprendizado objetiva melhorar a aprendizagem da função de predição $f_t(\cdot)$ no domínio de destino D_t utilizando o conhecimento disponível D_s e T_s.

A transferência de aprendizado utilizando CNN tem sido aplicada com sucesso em

vários domínios e tarefas de visão computacional. A proposta deste trabalho é transferir aprendizado de modelos pré-treinados em tarefas de reconhecimento de gestos e ações para a tarefa de reconhecimento de sinais da Libras.

Figura 2.15: Exemplo do fluxo do processo em transferência de aprendizado.



Fonte: Adaptado de Tan et al. (2018).

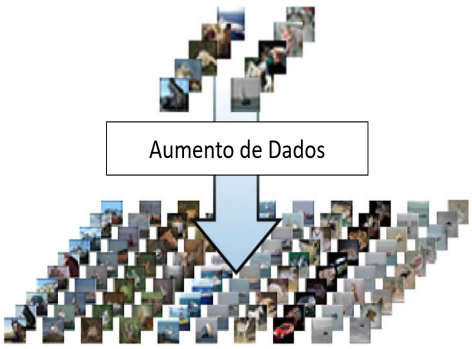
2.6 Aumento de Dados

Além da transferência de aprendizado, há na literatura outra técnica que reduz o superajuste de modelos de RNA, conhecida como aumento de dados. Em visão computacional, o aumento de dados infla artificialmente o conjunto de treinamento aplicando transformações nas imagens, podendo gerar o efeito exemplificado na Figura 2.16. Recentemente, tem havido amplo uso de aumento de dados para melhorar o desempenho das CNN (Taylor e Nitschke, 2017).

Conforme detalhado no trabalho de Mikołajczyk e Grochowski (2018), os métodos existentes para aumento de dados em imagens podem ser classificados em duas categorias gerais: métodos tradicionais de caixa branca ou métodos de caixa preta baseados em redes neurais profundas. A seguir, serão apresentados detalhes característicos da primeira categoria, que consiste em transformações geométricas e fotométricas (He et al., 2015; Simonyan e Zisserman, 2014), a qual será utilizada neste trabalho.

As transformações geométricas alteram a geometria da imagem com o objetivo de tornar a CNN robusta à mudança de posição e orientação. Como exemplo podem ser incluídas as transformações de reflexão, corte, cisalhamento, redimensionamento e rotação (Taylor e Nitschke, 2017). As transformações fotométricas alteram os canais

Figura 2.16: Exemplo de efeito do aumento de dados.



Fonte: Adaptado de Taylor e Nitschke (2017).

de cores com o objetivo de tornar a CNN robusta à mudança de iluminação e cor, como por exemplo, transformações como equalização do histograma, aprimoramento do contraste ou brilho, equilíbrio do branco, nitidez e desfoque da imagem (Mikołajczyk e Grochowski, 2018).

Neste trabalho, serão aplicados os métodos tradicionais de aumento de dados baseados na combinação de transformações de imagens afins e em modificação de cores, considerando que, conforme apresentado por Mikołajczyk e Grochowski (2018), são de rápida implementação e já provaram ser uma boa estratégia para aumentar o conjunto de dados de treinamento em bases compostas por imagens, ou por vídeos representados por sequências de imagens.

2.7 Métricas de Avaliação

Existem várias métricas que visam avaliar os resultados apresentados após o treinamento dos modelos. A seguir, serão descritas algumas métricas utilizadas para análise de resultados, sob um ponto de vista matemático, em problemas de classificação. As métricas mais comuns geralmente são baseadas na matriz de confusão: uma tabela que mostra as frequências de classificação para cada classe do modelo. Um exemplo de matriz de confusão pode ser visualizado na Figura 2.17:

- Verdadeiros Positivos (VP): representam a quantidade de instâncias corretamente classificadas. Tome-se como exemplo a classe “maçã” da Figura 2.17b: o modelo classificou corretamente todas as instâncias da classe “maçã”, pois todas as previsões condizem com o que se espera.

- Verdadeiros Negativos (VN): ocorrem quando a negação de uma classe é feita corretamente, como na 2.17b, em que 4 instâncias de laranja foram classificadas como "não maçã".

- Falsos Positivos (FP): acontecem quando uma instância é positivamente classificada para uma classe à qual não pertence: na Figura 2.17b, todas as instâncias de maçã foram classificadas corretamente, entretanto, há muitos FP para a classe laranja.

- Falsos Negativos (FN): são representados por instâncias que foram classificadas como não pertencentes à uma classe, quando na realidade elas pertencem. Nesse exemplo, os FN para a classe "laranja" pertencem ao mesmo conjunto dos FP da classe "maçã". A percepção da diferença entre os conceitos se torna mais clara quando se trata de um problema multiclasse.

A acurácia é a métrica que avalia o desempenho do modelo no geral. Isso significa que, por não focar em uma classe específica, trata-se de uma métrica muito suscetível à distribuição das classes, ou seja, o desbalanceamento do número de instâncias por classe pode facilmente influenciar de maneira negativa o resultado. A acurácia é calculada da seguinte maneira:

$$acurácia = \frac{VP + VN}{VP + FP + FN + VN} \tag{2.2}$$

Portanto, a acurácia corresponde ao quociente entre soma de predições corretas e o número total de instâncias do conjunto. Sendo assim, no exemplo da Figura 2.17b, a acurácia é de 73.68%. Como pode ser observado, apesar de ser aparentemente um bom resultado de classificação, o desbalanceamento do conjunto de dados fez com que o modelo não aprendesse bem as características da classe "laranja". Isso significa que o modelo pode estar apenas arriscando ao classificar uma instância.

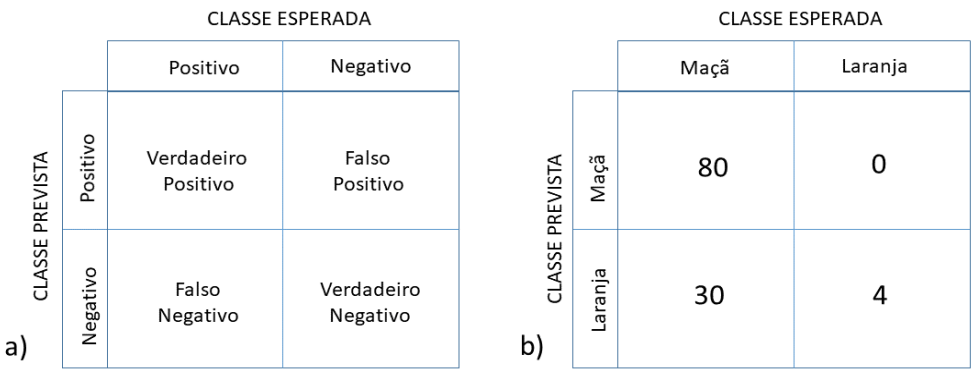
Diante disso, uma alternativa é avaliar o desempenho de uma classe específica. Nesse caso, dois cenários podem ser considerados: o total de instâncias classificadas corretamente dentre todas as que foram classificadas como pertencentes a uma classe; e o total de instâncias de uma classe que foram corretamente classificadas. Para o primeiro cenário, utiliza-se a métrica chamada precisão, que se dá da seguinte maneira:

$$precisão = \frac{VP}{VP + FP} \tag{2.3}$$

Para o segundo cenário, utiliza-se a métrica revocação, calculada conforme a Equação 2.4:

$$revocacão = \frac{VP}{VP + FN} \tag{2.4}$$

Figura 2.17: Matriz de Confusão



Fonte: Próprio Autor.

2.8 Considerações Finais

Neste capítulo foram apresentados os conceitos teóricos fundamentais para o entendimento, direcionamento e execução deste trabalho. A apresentação de conceitos relacionados à estrutura e aos aspectos que compõem os sinais da Libras possibilitou a percepção de que a tarefa de reconhecimento de sinais é complexa.

Também foi ressaltado que os pesquisadores da área de linguística da Libras reforçam a necessidade de identificação e compreensão dos parâmetros, pois, como alguns sinais são muito semelhantes, o contraste semântico pode passar desapercebido e a tradução pode ser incorreta. Portanto, um sistema de RLS deve ser robusto nas tarefas de aquisição dos dados, sendo capaz de capturar todos os parâmetros do sinal, e de reconhecimento dos sinais, tratando corretamente as similaridades interclasses, ou seja, a semelhança que pode existir entre sinais com diferentes significados.

Como apresentado, as redes neurais tornaram-se dominantes como estado da arte em reconhecimento de fala, de ações, e de gestos. Além disso, estratégias de fusão de dados, associadas a técnicas de transferência de aprendizado e aumento de dados têm contribuído para esse avanço. Sendo assim, o próximo capítulo apresenta os trabalhos de RLS mais recentes, baseados em RNA profundas.

Capítulo 3

Trabalhos Relacionados

Neste capítulo são primeiramente apresentados os trabalhos que envolvem reconhecimento de sinais da Libras, e uma síntese dos trabalhos, na Seção 3.1. Em seguida, é apresentada uma taxonomia das soluções utilizadas recentemente em RLS de outros países, bem como uma discussão sobre os trabalhos estudados (Seção 3.2). Além disso, é apresentado brevemente o cenário atual em reconhecimento de ações em vídeo, e as contribuições que podem ser obtidas dos avanços nesse cenário e aplicadas ao RLS. Para finalizar, a Seção 3.4 apresenta as considerações finais deste capítulo.

3.1 Reconhecimento de Sinais da Libras

Há na literatura algumas aplicações que foram desenvolvidas para o reconhecimento de sinais da Libras. A seguir, são apresentados trabalhos importantes publicados entre os anos de 2016 e 2019.

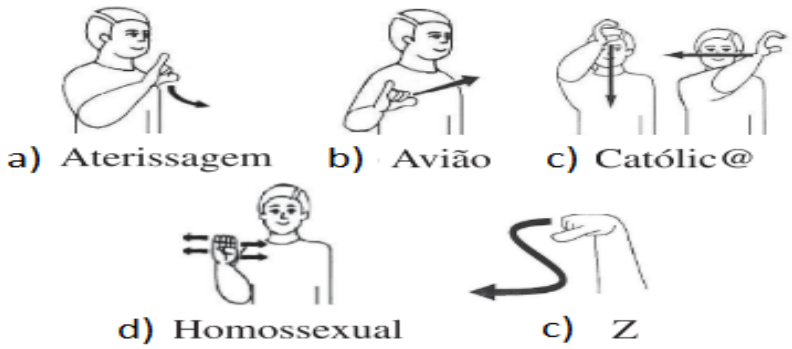
No trabalho de Barros Junior (2016) foi feito o reconhecimento das 16 configurações de mãos da Libras, executadas por cinco pessoas. As imagens, capturadas pelo sensor de profundidade do *Kinect* foram convertidas para escala de cinza. Na etapa de pré-processamento foram extraídas apenas as regiões onde se localizavam as mãos. Uma CNN foi utilizada para extração de características e classificação, alcançando acurácia de 87.5%.

Semelhantemente, Gonçalves et al. (2016) desenvolveram uma aplicação para tradução de sinais estáticos da Libras, com imagens capturadas pelo *Kinect*. Informações de bordas e formas da mão, representadas por matrizes binárias geradas na fase de pré-processamento, foram enviadas para uma rede neural para o reconhecimento dos gestos. A base de dados gerada é composta por letras do alfabeto, eliminando as letras cujos sinais possuem algum movimento, e as letras U e T, por apresentarem configu-

rações semelhantes às das letras R e F, respectivamente. Os 17 sinais selecionados foram executados utilizando somente as mãos, desconsiderando outras partes do corpo ou expressões faciais. Os resultados foram de 88% para a acurácia.

Leal (2018) apresenta um modelo de reconhecimento de dois dos parâmetros da Libras: configurações de mão e movimento. A captura da estrutura da mão é realizada por meio do LMC, de modo que a estrutura foi reconstruída e codificada em um espaço tridimensional, e enviada para uma RNA Perceptron Multicamadas (MLP) para extrair características e realizar a classificação. O modelo realizou a classificação de 21 sinais estáticos e 5 sinais dinâmicos, ilustrados na Figura 3.1, obtendo uma precisão de 99.8% e 86.7%, respectivamente. Considerando que foi utilizado o LMC para aquisição dos dados, parâmetros como expressões faciais e corporais não foram considerados.

Figura 3.1: Conjunto de sinais utilizados por Leal (2018).

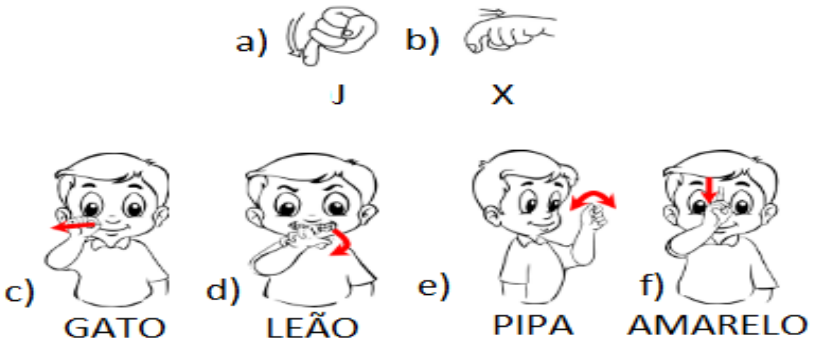


Fonte: Adaptado de Leal (2018).

No trabalho de Silva (2018), o objetivo foi desenvolver quatro dispositivos capazes de identificar configuração, orientação e movimento das mãos, verificando qual possui melhor desempenho para reconhecimento de sinais da Libras. Foi desenvolvida uma luva para receber sinais de sensores de flexão, acelerômetros e giroscópios. O reconhecimento dos padrões de cada sinal é realizado utilizando RNA. Após treinada, validada e testada, a rede neural interligada aos dispositivos obteve média de acerto de até 96,8%. O autor menciona que foram treinadas 36 classes, apesar de terem sido apresentados resultados para somente 16 classes, sendo 10 classes de sinais estáticos e 6 classes de sinais dinâmicos, ilustrados na Figura 3.2. Os resultados relatados alcançaram uma média de 99.68% de acurácia para sinais estáticos e 97.48% para sinais dinâmicos.

Voigt (2018) apresentou algumas estratégias para RLS utilizando aprendizado profundo para o reconhecimento das 26 letras do alfabeto, considerando 19 sinais estáticos e 7 sinais dinâmicos, ilustrados na Figura 3.3. Por meio de dados adquiridos

Figura 3.2: Conjunto de sinais utilizados por Silva (2018).



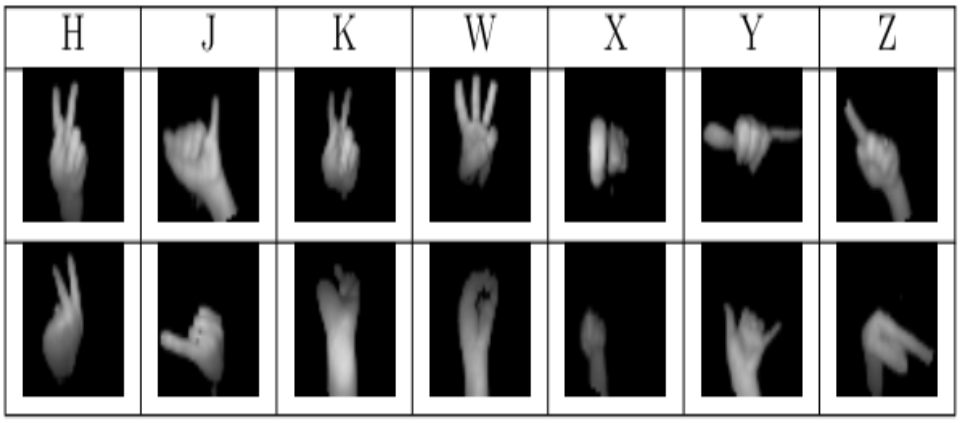
Fonte: Adaptado de Silva (2018).

pelo LMC (imagens RGB e informações de esqueleto), foram avaliados modelos de RNA profundas para RLS. Nesse trabalho é feita a fusão em nível de características: é realizado o reconhecimento utilizando RNA Multicamadas para os dados do esqueleto, CNN para as imagens RGB e uma terceira rede, combinando os dois tipos de informação. Em seguida, para a classificação de sinais dinâmicos, foram incluídas camadas LSTM. Também foi realizada a transferência de aprendizado nos blocos de convolução, reutilizando o conhecimento adquirido durante o treinamento com sinais estáticos para o modelo de classificação de sinais dinâmicos. Foram realizados experimentos com os seguintes cenários: variando o tipo de entrada; combinando tipos diferentes de entrada; para sinais dinâmicos, com transferência de aprendizado e sem transferência. Para os 19 sinais estáticos, foram obtidos os seguintes resultados de acurácia: 92%, com apenas imagens RGB como entrada; 78%, com apenas dados do esqueleto na entrada; 99%, combinando ambos os tipos de dado. Para os 7 sinais dinâmicos, foram obtidos os seguintes resultados de acurácia: 79% (sem transferência de aprendizado) e 92% (com transferência de aprendizado), com apenas imagens RGB como entrada; 55%, com apenas dados do esqueleto na entrada; 60% (sem transferência de aprendizado) e 63% (com transferência de aprendizado), combinando ambos os tipos de dado.

Magalhaes (2018) também emprega uma abordagem que utiliza aprendizado profundo. Por meio de imagens RGB é realizado o reconhecimento de 30 sinais estáticos, utilizando uma CNN residual. Também é feita a classificação de 2000 vídeos de frases que foram aleatoriamente construídas por um gerador de sentenças, sem um contexto específico, a partir de 72 sinais da Libras. Para essa base de dados, foram adicionadas ao modelo células LSTM no topo das camadas convolutivas.

Semelhantemente, Machado (2018) implementa um modelo de RNA profunda

Figura 3.3: Conjunto de sinais utilizados por Voigt (2018).



Fonte: Voigt (2018).

que, com fusão a nível de características, por meio de dados RGB-D e transferência de aprendizado, realiza a classificação de 84 sinais dinâmicos da base LIBRAS_APOEMA. Vale destacar que foram filmados 560 sinais, mas devido ao tempo necessário para finalização da base, somente 84 palavras foram utilizadas. O conjunto contém sinais que são frequentemente empregados na alfabetização de surdos e deficientes auditivos. É feita a combinação de três modelos: uma 3D-CNN, para extração de características de curto prazo do vídeo de entrada; um modelo de CNN Bidirecional LSTM (ConvLSTM), para aprender características de longo prazo e globais; e uma CNN-2D, para aprender características de alto nível. A etapa de classificação é realizada pelo classificador SVM. O modelo é inicialmente treinado no *Isolated Gesture Dataset*[1] (IsoGD) e em seguida, o conhecimento é utilizado no treinamento da amostra de 84 sinais.

3.1.1 Síntese dos trabalhos

Dentre os trabalhos estudados, as propostas apresentadas por Barros Junior (2016), Gonçalves et al. (2016) e Magalhaes (2018) não realizam o reconhecimento de sinais dinâmicos isolados. Entretanto, no trabalho apresentado por Silva (2018), apesar de ser feita a classificação de sinais dinâmicos, a aquisição dos dados é realizada por meio de uma luva, o que pode ser considerada uma desvantagem em relação aos trabalhos que utilizam apenas visão computacional.

Por outro lado, todos os trabalhos apresentam bases de dados que possuem uma quantidade inferior a 10 sinais dinâmicos, com exceção do conjunto utilizado por Ma-

[1]Disponível em *http://www.cbsr.ia.ac.cn/users/jwan/database/isogd.html*

Tabela 3.1: Resumo Comparativo dos Trabalhos de Libras

Autor	Acurácia	Sinais Estáticos (SE)	Sinais Dinâmicos Isolados (SD)	CNN/ LSTM
Barros (2016)	87.5%	16	-	CNN
Golçalves (2016)	88%	17	-	-
Leal (2018)	99.8% (SE) 86.7% (SD)	21	5	-
Silva (2018)	99.6% (SE) 97.4% (SD)	10	6	-
Voigt (2018)	99% (SE) 92% (SD)	19	7	CNN e LSTM
Magalhaes (2018)	99.83%	30	*	CNN e LSTM
Machado (2018)	79.8%	**	84	CNN e LSTM

* Reconhecimento dinâmico de sentenças.
** Base composta por sinais estáticos e dinâmicos, todos representados em vídeo.

chado (2018), em que essa quantidade é muito superior a todos os outros. Apesar dos avanços trazidos por esse autor, o tamanho dessa base ainda não é suficiente para que conclusões mais precisas sejam obtidas quanto à possibilidade concreta de classificação de sinais da Libras. Uma possível razão para a insuficiência de resultados abrangentes e mais conclusivos seja a falta de bases de dados disponíveis publicamente, fazendo com que cada trabalho seja responsável pela geração de um conjunto próprio, tarefa que demanda bastante tempo e recurso, conforme detalhado por Machado (2018).

Outra questão a ser destacada é que a maioria dos trabalhos utiliza técnicas que não fazem parte do atual estado da arte em reconhecimento de gestos. Dos oito trabalhos apresentados, apenas três dos mais recentes aplicam métodos de modelagem temporal explícita. Além disso, há pouca variação entre fontes de entrada, ou combinação de várias entradas, como RGB, profundidade, fluxo ótico e esqueleto. Também são pouco exploradas as abordagens multimodal e multicanal.

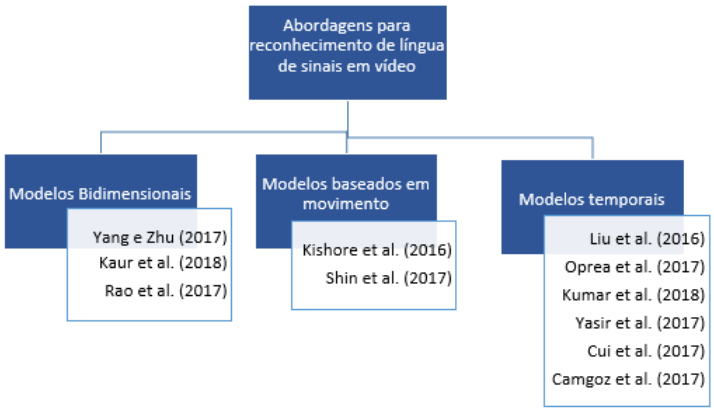
Um resumo comparativo entre os trabalhos pode ser visto na Tabela 3.1, considerando os resultados apresentados, se o trabalho realiza reconhecimento de sinais dinâmicos, o tamanho das bases de dados e se o trabalho utiliza modelagem temporal explícita. Dados esses fatores, constatou-se a necessidade de uma pesquisa mais extensa na literatura. Por esse motivo, foi realizada uma revisão para identificar os avanços alcançados no RLS de outros países.

3.2 Reconhecimento de Línguas de Sinais de Outros Países

A revisão da literatura foi conduzida com base na estrutura proposta por Kitchenham e Charters (2007), dividida em três fases: planejamento, condução e documentação dos resultados. Como a evolução das pesquisas na área de aprendizado profundo é muito rápida, os métodos se tornam obsoletos após um curto período de tempo. Sendo assim, foi delimitado um período de tempo a ser estudado, compreendendo os anos de 2016, 2017 e 2018, por ser entendido que dentro desse intervalo devem estar registradas as abordagens consideradas estado da arte no contexto da utilização de redes neurais profundas na tarefa de RLS.

Asadi-Aghbolaghi et al. (2017) introduziram uma taxonomia que resume abordagens que utilizam redes neurais para reconhecimento de gestos e ações. Para estruturar os trabalhos estudados nesta pesquisa, foi feita uma adaptação da taxonomia apresentada por esses autores. A Figura 3.4 apresenta tal adaptação, que abrange os principais modelos empregados em trabalhos de RLS que utilizam redes neurais profundas. A seguir, serão descritas as principais características individuais de cada abordagem, bem como a sua aplicação em conjunto, além de uma análise dos resultados.

Figura 3.4: Abordagens para reconhecimento de língua de sinais: taxonomia.



Fonte: Próprio Autor.

Como já destacado, um dos maiores desafios no RLS é tratar o aspecto dinâmico do gesto, isto é, a dimensão temporal. Sendo assim, os trabalhos foram agrupados em três categorias, de acordo com a forma como enfrentam esse problema:

- Usando redes convolutivas bidimensionais: exploram a informação espacial das

imagens que compõem a sequência e aplicam modelos pré-treinados em bases de dados maiores (Yang e Zhu, 2017; Kaur et al., 2018; Rao et al., 2018)

- Extraindo previamente características do movimento: utilizam técnicas que primeiro descrevem o movimento da sequência de imagens e depois adaptam tais descrições para que alimentem uma CNN (Kishore et al., 2016; Shin et al., 2017).

- Realizando modelagem da sequência temporal: combinam modelos de redes 2D com métodos de modelagem de sequência temporal, como o modelo oculto de Markov (Eddy, 1998), do inglês *Hidden Markov Model* (HMM), ou redes neurais recorrentes e suas variações, como LSTM e LSTM bidirecionais (BiLSTM) (Liu et al., 2016; Oprea et al., 2017; Kumar et al., 2018; Yasir et al., 2017; Cui et al., 2017; Camgoz et al., 2017).

3.2.1 Modelos Bidimensionais

Os modelos denominados bidimensionais são os modelos baseados em convoluções 2D, que exploram a informação espacial dos dados de entrada, sendo os mais utilizados atualmente em processamento de imagens. Geralmente, é realizada a transferência de aprendizado de modelos pré-treinados em bases de dados maiores.

No trabalho apresentado por Yang e Zhu (2017) é feito o reconhecimento de 40 sinais da língua chinesa de sinais. Primeiramente é realizada a segmentação da imagem utilizando o método de classificação de características Harr, da biblioteca *OpenCV*[2]. As imagens da parte superior do corpo, centradas na mão do emissor, são extraídas dos vídeos por meio das seguintes etapas: remoção da região da face; detecção do contorno da mão; detecção do centro da mão; e remoção da parte superior do corpo ao redor do centro da mão. As imagens, originalmente no espaço de cores RGB, passam por duas conversões: primeiramente para o espaço YCbCr, para facilitar a detecção da mão por meio do tom de pele, e posteriormente, para o espaço HSV. Um modelo de CNN pré-treinado é utilizado como extrator e classificador.

Kaur et al. (2018) apresentam um modelo que segue a arquitetura de uma rede convolutiva bidirecional, treinado com uma base de dados composta por 35 sinais estáticos e dinâmicos da língua americana de sinais. As imagens RGB alimentam diretamente o modelo, sem passarem por qualquer etapa mais complexa de segmentação ou pré-processamento. O modelo é treinado com transferência de aprendizado da *Inception-V3* pré-treinada na base ImageNet. Semelhantemente, Rao et al. (2018) também apresentam uma arquitetura que utiliza 2D-CNN, sem etapa de segmentação. O foco no

[2]https://opencv.org/

trabalho é reconhecer os sinais da língua indiana de sinais usando uma câmera frontal móvel, com o objetivo de simular o comportamento do modelo ao ser utilizado em plataformas móveis. O conjunto de dados gerado contém 200 sinais. O modelo também foi pré-treinado na base de dados ImageNet.

3.2.2 Modelos Baseados em Movimento

Como já foi mencionado, as redes convolutivas representam uma estimativa baseada em informações espaciais. Entretanto, para que seja obtido melhor desempenho no reconhecimento de gestos dinâmicos, técnicas de extração prévia de características de movimento têm sido largamente aplicadas, visto que essa estratégia pode reduzir o tempo de treinamento. Diante disso e devido ao alto custo computacional necessário durante a fase de treinamento de CNN, alguns trabalhos recentes tentam explorar a capacidade das redes neurais profundas seguindo a arquitetura tradicional no reconhecimento de sinais, enviando como entrada dados de movimento previamente quantificado.

Na literatura podem ser encontradas diversas abordagens de quantificação de movimento, dentre elas, a abordagem baseada em gradiente, conhecida como fluxo ótico. Kishore et al. (2016) propuseram um sistema de reconhecimento de frases da língua indiana de sinais com rastreamento de mão utilizando *Horn Shunck Optical Flow* (HSOF) e análise da forma da mão. Essas informações compõem um vetor de características que alimenta uma RNA profunda clássica. É realizado o reconhecimento de 10 frases, compostas por 58 palavras.

Outros autores fazem uso de sensores para extrair informações de movimento mais precisas, com a desvantagem da utilização de dispositivos considerados intrusivos. Shin et al. (2017) utilizam dois sensores diferentes que podem produzir informação redundante para a mesma variável física. O método proposto é para reconhecimento automático da língua coreana de sinais baseado em uma tecnologia de fusão de sensores e redes neurais tradicionais. Os sensores de eletromiografia (EMG) e de unidade de medição inercial (IMU) são embutidos em uma pulseira, que deve ser utilizada pelo emissor do sinal. As informações são armazenadas em um vetor de características e passadas como entrada para o modelo de RNA profunda. Não foram encontradas informações referentes ao conjunto de dados utilizado.

3.2.3 Modelos Temporais

Métodos para modelagem da sequência temporal que utilizam aprendizagem profunda têm sido amplamente utilizados para o reconhecimento de sinais. Redes neurais recor-

rentes e suas variações (as LSTM e as LSTM bidirecionais (BLSTM)), e os modelos de sequência temporal, como o modelo oculto de Markov (HMM), são muitas vezes associados a modelos espaciais.

Liu et al. (2016) realizam a classificação baseada em informações sobre as localizações das articulações do esqueleto capturadas pelo sensor do *Kinect 2.0*: mão esquerda, mão direita, cotovelo esquerdo e cotovelo direito. São classificados 100 sinais da língua chinesa de sinais. Diferentemente dos métodos baseados em trajetórias tradicionais, é projetado um modelo de ponta a ponta baseado em LSTM. O vetor de características espaço-temporal, composto pelas localizações das articulações do esqueleto, é a entrada da LSTM. Similarmente, no trabalho de Oprea et al. (2017), a rede é alimentada com um vetor de características oriundo da câmera RGB-D do *Kinect*, contendo informações sobre distância e ângulo entre ombros, coluna vertebral, cotovelos e pulsos. Nesse trabalho, são implementados os modelos RNN, LSTM e GRU (*Gated Recurrent Unit*). O trabalho é voltado para reconhecimento de 25 sinais da *Schaeffer*, uma língua de sinais criada especificamente para crianças com distúrbios comunicativos, como o autismo. A proposta de Kumar et al. (2018) representa as informações de distâncias e os ângulos entre pares de articulações com mapas de deslocamento angular (JADMs). Os JADMs são extraídos de vídeos 3D *motion-capture* e codificados em imagens RGB que alimentam a CNN. O reconhecimento é feito para 200 sinais da língua indiana.

Diferentemente dos modelos temporais apresentados até agora, a solução proposta por Yasir et al. (2017) associa redes recorrentes a redes convolutivas para reconhecimento da língua bangla de sinais. O LMC captura a sequência das imagens. O HMM é usado para separar os sinais contínuos usando estados de transição. Depois de segmentar o sinal do início ao fim, uma CNN é alimentada com o vetor de características. O LMC calcula a rotação, orientação e texturas das mãos para determinar e extrair o gesto da mão. Para cada sinal, há um ponto inicial e um ponto final de estado, cujas transições de estado são segmentadas em HMM. Caso haja uma diferença de histograma em qualquer estado, o estado de transição é movido para o novo quadro para obter uma nova representação de sinal. Um vetor de características é recebido na entrada da CNN e conectado por meio de uma série de camadas ocultas. Não foram encontradas informações sobre a quantidade de sinais classificados.

A proposta de Cui et al. (2017) emprega CNN com convolução temporal e agrupamento para aprendizagem de representação espaço-temporal a partir de vídeos de frases, com módulo LSTM para aprender o mapeamento de sequências de glosas do alemão. É utilizada a base pública RWTH-PHOENIX-Weather 2014, que contém 5.672 frases, compostas por 65.227 palavras. Para garantir a consistência temporal, os autores utilizaram a função CTC (Classificação Temporal Conexionista). Esse compo-

nente temporal foi introduzido por Graves et al. (2006), sendo inicialmente aplicado ao problema de reconhecimento de escrita. O CTC é responsável por garantir que o encadeamento dos sinais da sequência faça sentido.

Camgoz et al. (2017) apresentaram um modelo de arquitetura profunda para sequenciar problemas de aprendizado em vídeos genéricos, empregando sub-redes especializadas, as quais são treinadas para modelar subunidades de uma determinada tarefa. Essa abordagem força a rede a modelar explicitamente o conhecimento específico do domínio, restringindo melhor o problema de reconhecimento geral. Cada subconjunto consiste em três camadas de rede neural. Primeiro, as CNN captam as imagens de entrada e extraem características espaciais. Depois, as BiLSTM realizam a modelagem temporal das características espaciais extraídas pelas CNN. Ao final, há uma camada para a função CTC, que permite que as redes sejam treinadas com vídeos de duração diferente. O modelo foi treinado com o conjunto de dados *One Million Hands* (Koller et al., 2016), que possui sinais em RGB das línguas dinamarquesa, alemã (RWTH-PHOENIX-Weather 2014) e da Nova Zelândia.

3.2.4 Síntese dos trabalhos

Um comparativo entre os trabalhos pode ser visualizado na Tabela 3.2, que apresenta uma síntese dos trabalhos, ordenados pelo ano de publicação, incluindo a forma de aquisição dos dados. Além das métricas comuns em avaliação de métodos de aprendizagem de máquina, como acurácia, revocação e precisão, os trabalhos apresentaram as métricas WER (a taxa de erro de palavra) e a WMS (pontuação de correspondência de palavra).

Como discorrido, os métodos da primeira categoria extraem características apenas da informação espacial, ou seja, não há processamento envolvendo a dimensão temporal. Sendo assim, a tarefa de reconhecimento de gestos dinâmicos apresenta resultados aquém do esperado. Por outro lado, devido à disponibilidade de grandes conjuntos de dados rotulados, o ajuste do modelo por transferência de aprendizado se torna mais fácil. Para que seja obtido melhor desempenho no reconhecimento de gestos dinâmicos, trabalhos pertencentes à segunda categoria aplicam técnicas de extração de características de movimento. Essas características são calculadas previamente e depois alimentadas no modelo, mantendo-se o controle sobre as características de movimento. Entretanto, esses modelos se limitam a explorar apenas informações temporais locais. Em contraste, a grande vantagem das abordagens da terceira categoria, a saber os modelos temporais como RNN e LSTM, é a capacidade de capturar relações temporais de longo alcance. Como a maioria desses trabalhos utiliza dados do esqueleto humano,

que não possuem muitas dimensões, as redes possuem menos pesos.

3.3 Reconhecimento de Ações Utilizando 3D-CNN

Além dos trabalhos selecionados a partir da revisão conduzida para identificar os avanços em reconhecimento de línguas de sinais, outros trabalhos foram estudados no âmbito de reconhecimento de gestos e ações, com o objetivo de analisar o cenário atual nessa área e verificar as possíveis contribuições que poderiam ser agregadas a este trabalho.

Recentemente, os índices de desempenho de 3D-CNN no campo do reconhecimento de ações melhoraram significativamente. O trabalho de Carreira e Zisserman (2017) se tornou um dos mais relevantes do ramo por ter apresentado muitos avanços para esse tipo de aplicação. Os autores introduziram o *Inflated* 3D Convnet (I3D), modelo para representação e classificação de ações a partir de imagens RGB e de fluxo ótico, calculado com o algoritmo TV-L1 (Zach et al., 2007).

O I3D é baseado na Inception-V1. Visto que a arquitetura original envolve um modelo bidimensional, os autores propuseram a conversão para uma arquitetura 3D "inflando" todos os filtros em uma dimensão temporal adicional: os pesos dos filtros 2D são repetidos ao longo de toda a terceira dimensão. Essa estratégia permite que a rede reutilize filtros 2D pré-treinados em bases de imagens, aumentando o poder de generalização das 3D-CNN. Diante do sucesso do I3D, esse modelo foi selecionado para os experimentos deste trabalho.

No trabalho de Carreira e Zisserman (2017) foram utilizadas duas bases de dados consideradas como referências em reconhecimento de ações: a UCF-101, que contém 101 classes; e a HMDB-51, composta por 51 classes. Além disso, foi gerada a base de dados Kinetics, conjunto composto por várias categorias de ações humanas, grande o suficiente para treinar com sucesso as CNN tridimensionais. Para os experimentos, foram realizados treinamentos nas bases aplicando transferência de aprendizado do modelo pré-treinado na base Kinetics, que também foi publicamente disponibilizado. O modelo multimodal alcançou 97.9% e 80.2% de acurácia para as bases UCF-101 e HMDB-51, respectivamente. A partir desse modelo, muitos trabalhos vêm sendo desenvolvidos e há um alto potencial de progresso (Hara et al., 2018).

3.4 Considerações Finais

Após analisar os trabalhos pode-se concluir que, apesar dos avanços adquiridos em pesquisas de RLS de outros países, não é observada essa mesma evolução no contexto

de reconhecimento de sinais da Libras. Os melhores resultados encontrados são para reconhecimento de sinais estáticos, ou de uma quantidade muito reduzida de sinais dinâmicos. Outro ponto a ser destacado é que existem poucos conjuntos de imagens de sinais Libras disponíveis: a maioria dos trabalhos cria sua própria base para realizar os experimentos, ou não apresenta a base de imagens utilizada.

Em contrapartida, os trabalhos de RLS de outros países ainda não refletem os avanços alcançados no campo de reconhecimento de ações. A nossa revisão de literatura mostra o trabalho de Machado (2018) como uma exceção, por usar redes convolutivas 3D para reconhecimento de sinais dinâmicos, ainda que associadas às redes recorrentes. Diante desse cenário, desponta a oportunidade de reproduzir os avanços obtidos em reconhecimento de ações em reconhecimento de sinais da Libras.

Tabela 3.2: Resumo dos Trabalhos Apresentados

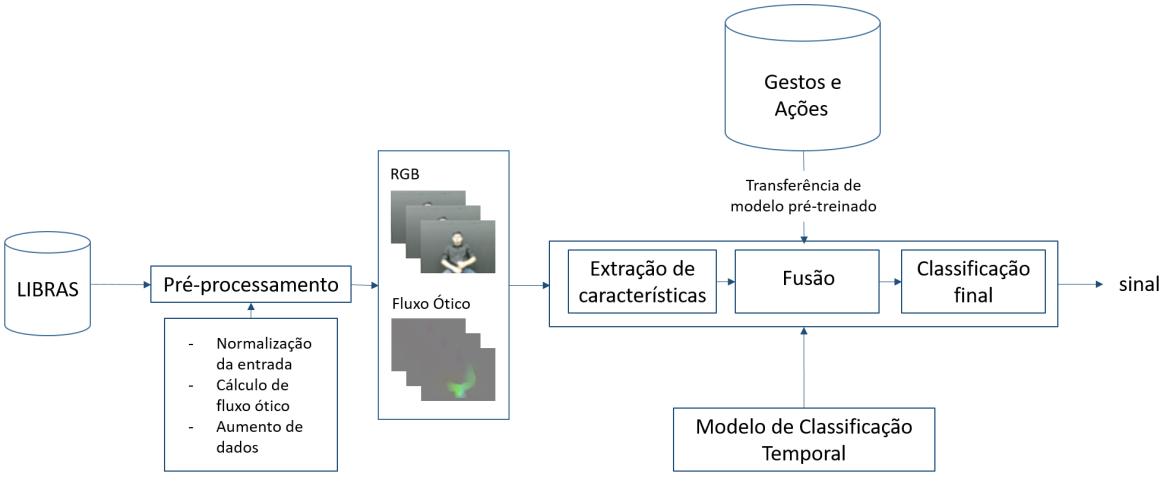
Autor	Resultado		Modelo			Aquisição
	Métrica	%	Bidimensional	Movimento	Temporal	
Kishore et al. (2016)	WMS	90	-	X	-	RGB
Liu et al. (2016)	Acurácia	86	-	-	X	Kinect
Camgoz et al. (2017)	WER	40.8	X	-	X	RGB
Cui et al. (2017)	WER	38.7	X	-	X	RGB
Oprea et al. (2017)	Acurácia	93.13	-	-	X	Kinect
Rao et al. (2017)	Acurácia	74	X	-	-	RGB
Shin et al. (2017)	Acurácia	99.13	X	X	X	RGB, EMG, IMU
Yang e Zhu (2017)	Acurácia	99.84	X	-	-	RGB
Yasir et al. (2017)	WER	3	X	X	-	LMC
Kaur et al. (2018)	Acurácia	92.3	X	-	-	RGB
Kumar et al. (2018)	Reconhecimento	89.15		-	X	RGB
	Precisão	89.22				
	Revocação	92.14				

Capítulo 4

Abordagem Proposta

Neste capítulo é descrita a abordagem proposta para reconhecimento de sinais da Libras. A Figura 4.1 apresenta a arquitetura geral da abordagem, comportando as principais fases necessárias para a realização do reconhecimento do sinal, que são: pré-processamento, extração de características e classificação do sinal. Durante a fase de pré-processamento, serão realizadas adaptações nos dados da base para que sejam enviados como entrada para o modelo de rede neural. O modelo é responsável por realizar tanto a fase de extração de características quanto a fase de classificação.

Figura 4.1: Visão geral da estratégia proposta.



Fonte: Próprio Autor.

4.1 Metodologia

O objetivo deste trabalho é identificar cenários para o reconhecimento de sinais dinâmicos da Libras, considerando a relação existente entre três aspectos:

- Canais de dados: utilizando dados RGB e de fluxo ótico, separadamente e em conjunto, caraterizando um modelo multicanal;
- Transferência de aprendizado: executando diferentes estratégias de transferência de aprendizado e ajuste fino, considerando transferir aprendizado de modelos pré-treinados em tarefas como reconhecimento de gestos, de ações ou de outras línguas de sinais;
- Estratégia de fusão de dados: implementando as estratégias de fusão de dados em nível de características e em nível de decisão.

Portanto, esta abordagem proposta trata o problema a partir de uma perspectiva multicanal e envolve um estudo sobre a influência do aumento de dados e da transferência de aprendizado de diferentes tarefas, variando o método de fusão de dados, para a definição de uma estratégia que proporcione a elevação da taxas de reconhecimento de sinais da Libras. Antes da discussão sobre as principais etapas da abordagem proposta, a próxima seção descreve a base de dados utilizada e organizada no contexto deste trabalho.

4.2 Base de dados LIBRAS_APOEMA

A base da Libras utilizada neste trabalho foi filmada por Machado (2018) com a colaboração de profissionais da área de educação especial; profissionais surdos e ouvintes intérpretes da Libras; e professores com experiência em trabalhos de transcrição de livros para surdos e traduções em geral. Os participantes colaboraram com o processo de seleção e definição de quais seriam os sinais mais relevantes a serem adicionados ao conjunto de dados. Foram selecionados alguns sinais considerados importantes para o processo de alfabetização de surdos e deficientes auditivos, para redução do escopo dos sinais incluídos na base.

A base é composta por 560 sinais isolados da Libras, conforme pode ser observado na Tabela 4.1. Os sinais foram executados por 7 pessoas, sendo três homens (um surdo) e quatro mulheres (duas surdas), dos quais 4 executaram os sinais da partição de treino, 2 da partição de validação, e 1 da partição de teste. Todos os sinais foram repetidos 6

Tabela 4.1: Estrutura da base LIBRAS_APOEMA

Base	Classes	Intérpretes	Repetições por Intérprete	Instâncias por Classe	Total de Instâncias
Treino	560	4	6	24	13.440
Validação	560	2	6	12	6.720
Teste	560	1	6	6	3.360
Total	**560**	**7**	**6**	**42**	**23.520**

vezes por cada pessoa, totalizando 23.520 vídeos. O tempo de execução de cada sinal pode variar de 2 a 5 segundos, de forma que a quantidade de quadros por sinal não é padronizada, o que requer uma normalização posterior.

4.3 Detalhamento do Método

A estratégia é dividida em três etapas principais: pré-processamento dos vídeos, que envolve principalmente a padronização dos dados, aumento de dados e geração da base de dados em fluxo ótico; extração de características, realizada pelo modelo de 3D-CNN, aproveitando os padrões aprendidos previamente via transferência de aprendizado; e classificação, cujo resultado final é calculado a partir da fusão dos resultados obtidos por cada modelo individual. Essas etapas serão detalhadas a seguir.

4.3.1 Pré-processamento dos dados

Originalmente, cada sinal da base LIBRAS_APOEMA é representado por uma sequência de quadros somente em RGB, conforme exemplificado na Figura 4.2. Logo, a primeira etapa da estratégia proposta neste trabalho envolve o pré-processamento dos dados e a geração de informações para fusão multicanal. A primeira tarefa realizada durante essa etapa é a estimação do fluxo ótico. Como a intenção inicial é enviar para o modelo os dados de fluxo ótico previamente calculados, foi implementado um módulo de geração das imagens para representar o fluxo ótico, composto por duas etapas:

- amostragem de n quadros RGB, onde n é o tamanho da sequência temporal esperado pelo modelo. No caso de não haver n quadros, a diferença é preenchida com a replicação de outros quadros;
- estimativa do fluxo ótico, que, similarmente ao trabalho de Carreira e Zisserman (2017), é calculado com o algoritmo TV-L1, introduzido por Zach et al. (2007).

A Figura 4.3 apresenta uma amostra da representação da sequência de imagens com estimativas de fluxo ótico.

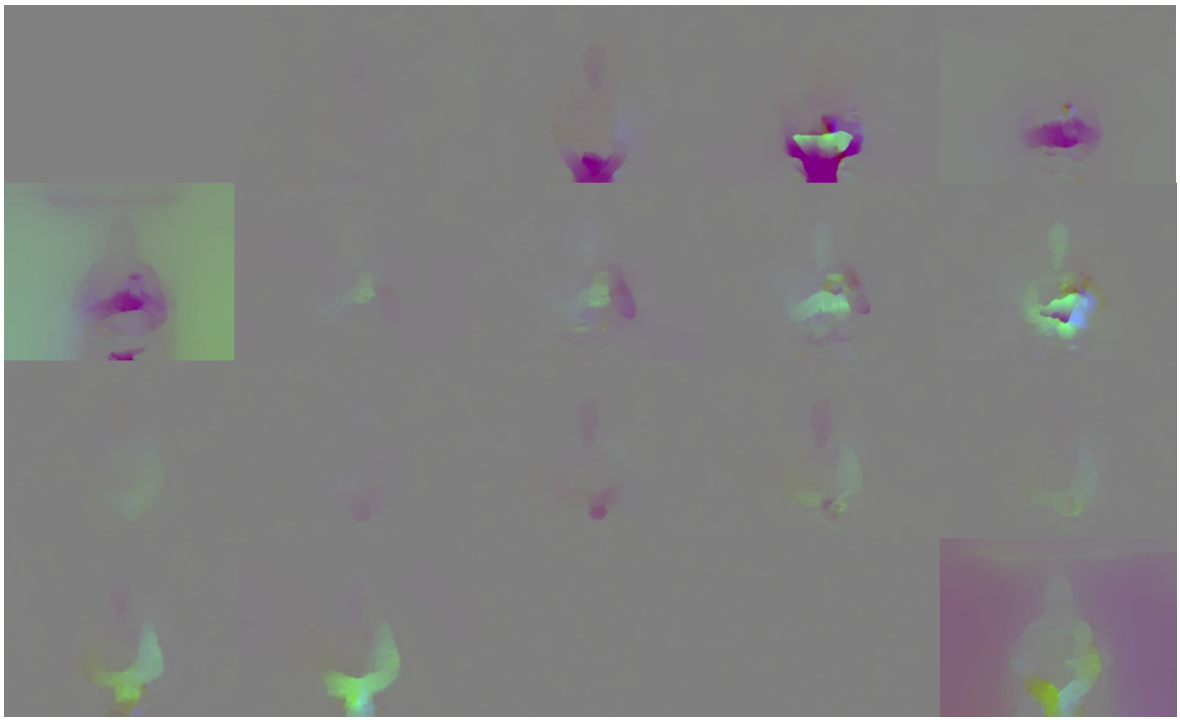
Figura 4.2: Exemplo de instância da base de dados: sequência em RGB do sinal À_FORÇA.

Fonte: Próprio Autor.

Outra atividade realizada durante essa etapa é a aplicação do aumento de dados na base. Portanto, foi implementado um módulo para geração de novas imagens aplicando transformações geométricas e fotométricas, utilizando os aumentadores disponibilizados por Jung et al. (2020). No total, foram utilizados 11 padrões de transformações, descritos a seguir:

- Adição: adiciona os *pixels* de uma imagem a um valor específico, causando um clareamento ou escurecimento na imagem;
- Multiplicação: multiplica os *pixels* de uma imagem por um valor específico, causando um clareamento ou escurecimento da imagem;
- Ruído de Sal: substitui uma porcentagem específica de *pixels* da imagem por ruído de *pixels* brancos.

Figura 4.3: Exemplo de instância da base de dados: sequência em fluxo ótico do sinal À_FORÇA.



Fonte: Próprio Autor.

- Ruído de Pimenta: substitui uma porcentagem específica de *pixels* da imagem por ruído de *pixels* pretos.
- Ruído aditivo gaussiano: adiciona à imagem ruídos amostrados por distribuição gaussiana.
- Transformação afim: aplica na imagem uma transformação linear, seguida por uma translação, podendo resultar em uma rotação ou cisalhamento da imagem.
- Transformação afim por partes: aplica uma grade regular de pontos em uma imagem e move aleatoriamente a vizinhança desses pontos por meio de transformações afins, gerando distorções locais.
- Corte: remove uma quantidade específica de colunas ou linhas de *pixels* das laterais da imagem.
- Espelhamento: espelha horizontalmente a imagem.

- Agrupamento: a partir de um núcleo de tamanho específico, aplica agrupamento nas imagens utilizando a média global.
- Desfoque: desfoca a imagem a partir de um núcleo gaussiano de tamanho específico.

Para cada instância da base foram geradas transformações aleatórias fixadas por classe, isto é, instâncias pertencentes à mesma classe receberam o mesmo tipo de transformação. Todos os padrões de transformações são parametrizados, ou seja, podem ser geradas várias versões da mesma imagem utilizando um único padrão.

4.3.2 Transferência de aprendizado

Apesar do conjunto de dados a ser utilizado neste trabalho ser maior do que qualquer outro trabalho encontrado na literatura que trata de reconhecimento de sinais da Libras, essa base ainda é considerada insuficiente para o treinamento de redes profundas do zero, visto que uma base suficientemente grande ultrapassa milhões de instâncias. Por esse motivo, foram utilizadas estratégias de transferência de aprendizado e ajuste fino de bases de imagens e vídeos consideradas como referência em reconhecimento de gestos e ações.

4.3.2.1 Bases de dados para transferência de Aprendizado

As duas bases utilizadas para pré-treinamento dos modelos são:

- ImageNet: os dados de treinamento desta base totalizam cerca de 1 milhão e 200 mil imagens, distribuídas entre 1.000 categorias de objetos (Krizhevsky et al., 2012). Apesar de ser composta por imagens, o modelo utilizado emprega uma estratégia para transferir o aprendizado para classificação de vídeos, que será descrita na próxima seção.
- Kinetics: a base de dados é focada em ações humanas. As classes são distribuídas nas seguintes categorias: ações de uma pessoa, como por exemplo, desenhar, beber, rir, socar; ações entre pessoas, como apertar as mãos, abraçar, beijar; e ações pessoa-objeto, como por exemplo abrir presentes, cortar a grama, lavar a louça. Além disso, algumas ações possuem ramificações, como os diferentes tipos de natação, portanto, requerem uma análise temporal mais sofisticada. Outras ações requerem mais ênfase no objeto para classificação, como é o caso das cenas que apresentam pessoas tocando tipos de instrumentos de sopro. O conjunto de

dados possui 400 classes de ação humana, com no mínimo 400 vídeos por classe, totalizando 240 mil vídeos, com duração de 10 segundos (Kay et al., 2017).

4.3.3 Extração de características e classificação

O objetivo desta etapa é explorar a capacidade da RNA investigada como extratores de característica: a rede reduzirá a informação dos quadros RGB e de fluxo ótico, gerando vetores de descrição discriminativa que codificam as principais características do sinal recebido na entrada. Para todas as classes é gerado um vetor, contendo a representação de uma ou mais características; cada classe é representada por um sinal da Libras.

Como mencionado na Seção 3.3, este trabalho utiliza a adaptação da arquitetura da Inception-V1, proposta por Carreira e Zisserman (2017). O modelo adaptado foi pré-treinado no contexto de reconhecimento de ações. Com a transferência de aprendizado, a fase de extração de características pode aproveitar os pesos aprendidos por modelos pré-treinados, reduzindo o tempo de treinamento. O treinamento é realizado seguindo uma estratégia de ajuste fino. Os vetores de características são repassados para o classificador e então é iniciada a fase de classificação do sinal.

4.3.3.1 Fusão de canais

Com relação à estratégia de fusão de dados, optou-se por realizar experimentos aplicando tanto fusão em nível de características quanto fusão em nível de decisão.

Para a fusão em nível de características, são primeiramente treinados os modelos separadamente. A etapa de fusão é realizada em seguida, quando os espaços de características de ambos os modelos são integrados e enviados para uma MLP. O treinamento é realizado novamente, associando as características aprendidas individualmente pelos modelos, seguindo até o processo de classificação.

Para a fusão em nível de decisão, os modelos também são treinados separadamente, de modo que cada modelo aprende a classificar os sinais, de acordo com os dados de entrada específicos. A etapa de fusão consiste na associação, por meio de determinadas regras de decisão, dos vetores de probabilidade resultantes de cada modelo.

Com base nas definições apresentadas por Kittler et al. (1998), considere-se um problema de reconhecimento de sinais, onde a instância Z deve ser atribuída a uma das classes Y, onde $Y = \{y_1, ..., y_m\}$. Assume-se que há R classificadores, cada um identificando uma instância, fornecida por um vetor de probabilidade independente. Denota-se por x_i o vetor de probabilidade usado pelo *iésimo* classificador.

Foram investigadas as seguintes regras de decisão: o valor máximo, a regra da média e a regra do produto das probabilidades *a posteriori* entre classificadores. Tais

regras são populares na área de aprendizado de máquina, mais especificamente no contexto de fusão de conjuntos/comitês de classificadores. As definições a seguir são baseadas no trabalho de (Kittler et al., 1998).

- Valor Máximo: dadas as medidas R e x_i, a instância Z deve ser atribuída à classe y_j, desde que a probabilidade *a posteriori* dessa interpretação seja a máxima, ou seja, atribui-se $Z \rightarrow \mathrm{y}_j$ se

$$P(y_j|x_1, ..., x_R) = \max_{k=1}^{m} P(y_k|x_1, ..., x_R) \tag{4.1}$$

- Regra da Média: é calculada a média das probabilidades *a posteriori* para cada classe, da saída de todos os classificadores. Dadas as medidas R e x_i, a instância Z deve ser atribuída à classe y_j, desde que o resultado da média das probabilidades *a posteriori* dessa interpretação seja o máximo, ou seja, atribui-se $Z \rightarrow \mathrm{y}_j$ se

$$\frac{1}{R}\sum_{i=1}^{R} P(y_j|x_i) = \max_{k=1}^{m} \frac{1}{R}\sum_{i=1}^{R} P(y_k|x_i) \tag{4.2}$$

- Regra do Produto: é calculado o produto das probabilidades *a posteriori* para cada classe, da saída de todos os classificadores. Dadas as medidas R e x_i, a instância Z deve ser atribuída à classe y_j, desde que o resultado do produto das probabilidades *a posteriori* dessa interpretação seja o máximo, ou seja, atribui-se $Z \rightarrow \mathrm{y}_j$ se

$$\prod_{i=1}^{R} P(y_j|x_i) = \max_{k=1}^{m} \prod_{i=1}^{R} P(y_k|x_i) \tag{4.3}$$

4.4 Considerações Finais

Neste capítulo foi apresentada a solução proposta para atingir os objetivos deste trabalho, visando aprimorar o reconhecimento de sinais da Libras. O modelo proposto pretende incorporar a modelagem temporal disponível na sequência de imagens da execução do sinal. Devido à quantidade limitada de imagens da base de dados de sinais da Libras a ser utilizada neste trabalho, este trabalho envolve um estudo sobre: a influência da transferência de aprendizado de diferentes bases de vídeos, representados por sequência de imagens, de reconhecimento de gestos e ações; a importância de se explorar diversos canais da dados; e as formas de fundir os resultados obtidos. No próximo capítulo serão apresentados os resultados obtidos.

Capítulo 5

Resultados

Neste capítulo, será detalhado o procedimento experimental executado e os resultados obtidos. Primeiramente, na Seção 5.1 é apresentado o protocolo experimental seguido nesta pesquisa, com informações sobre o modelo utilizado e os hiperparâmetros definidos. Em seguida, na 5.2 são apresentados os resultados dos experimentos, divididos em séries. A primeira série, que é descrita na Seção 5.2.1, foi executada em uma pequena amostra da base para definição de hiperparâmetros. Na na segunda série de experimentos (Seção 5.2.2), é feita uma comparação ao trabalho *baseline*, de Machado (2018). Paralelamente, é realizada uma análise sobre o impacto da transferência de aprendizado quando realizada a partir de bases de dados de diferentes contextos. Na Seção 5.2.3 é descrita a terceira série, quando são iniciados os experimentos para avaliar o impacto da utilização de técnicas de aumento de dados e de fusão de canais.

5.1 Protocolo Experimental

Como mencionado anteriormente, o I3D foi utilizado como modelo base para execução dos experimentos. O modelo original foi implementado em Python, utilizando a biblioteca PyTorch. Para os experimentos desta proposta, foi utilizada uma adaptação[1], que emprega a biblioteca Keras. O modelo foi pré-treinado nas bases ImageNet e Kinetics.

Considerando que os experimentos utilizando redes neurais profundas para classificação de vídeos podem ser muito demorados, os primeiros testes foram realizados com partições das bases. Inicialmente, a amostragem por instância foi mantida em n quadros, onde $n = 20$ – ou seja, todos os vídeos têm o tamanho fixo de 20 quadros. O processo de amostragem é descrito na seção 4.3.

[1]Disponibilizada publicamente em *https://github.com/dlpbc/keras-kinetics-i3d*

As tentativas de ajustes de parâmetros incluíram a variação entre os otimizadores SGD e Adam, de modo que o segundo, associado às taxas de aprendizado definidas, foi o que proporcionou melhores resultados. Foi utilizado o otimizador Adam tanto na fase de inicialização de pesos quanto na fase de ajuste fino, com as taxas de aprendizagem 10^{-3} e 10^{-4}, respectivamente. A primeira fase foi executada durante 3 épocas em todos os experimentos, enquanto a segunda variou conforme o desempenho do modelo, sendo, em alguns casos, alterada de acordo com o resultado obtido nos experimentos anteriores, utilizando monitoradores da função de perda entropia cruzada categórica. O tamanho do lote foi mantido em 4 instâncias em todos os experimentos. Para todos os experimentos, aplicou-se *dropout* de 50%.

5.2 Experimentos

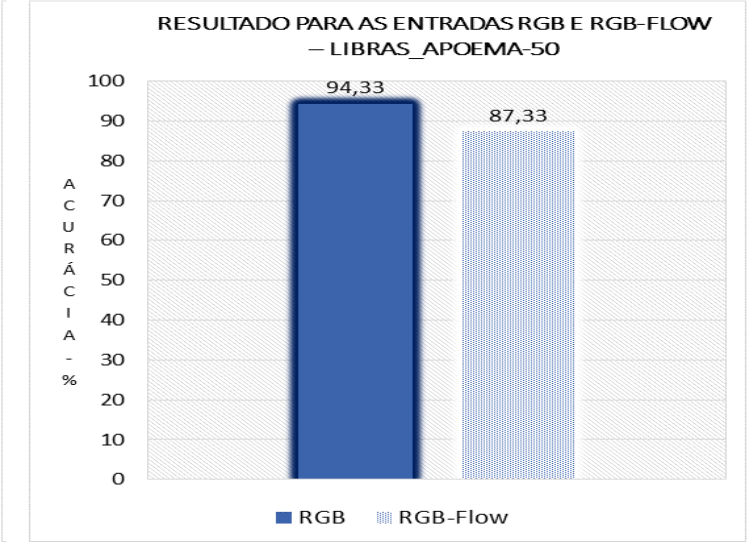
Os experimentos estão divididos em três séries. Na primeira série de experimentos, foi utilizada uma amostra com 50 sinais da base LIBRAS_APOEMA (LIBRAS_APOEMA-50). Os cenários de treinamento para os quais o modelo obteve melhor desempenho para essa amostra foram posteriormente treinados com a base completa, na terceira série de experimentos. Além disso, com o intuito de validar o modelo e compará-lo ao *baseline* Machado (2018), foram realizados experimentos utilizando a amostra de 84 sinais (LIBRAS_APOEMA-84) utilizada pelo referido autor. É importante observar que, conforme destacado no final no Capítulo 3, esse foi o único trabalho retornado em nossa revisão de literatura que utiliza 3D-CNN para classificação de sinais de Libras.

5.2.1 Primeira Série de Experimentos: 50 classes

Conforme descrito anteriormente, nesta primeira série de experimentos foi utilizada uma amostra composta pelas 50 primeiras classes da base LIBRAS_APOEMA, chamada aqui de LIBRAS_APOEMA-50. O modelo I3D, pré-treinado na ImageNet e na Kinetics, foi inicialmente utilizado para aprender os dados representados por fluxo ótico estimado dos dados RGB, chamado aqui de RGB-Flow.

A Figura 5.1 apresenta uma síntese dos resultados obtidos na amostra de teste. A partir dessa figura, é possível perceber que o modelo alcançou 87.33% de acurácia na base de teste, com treinamento executado durante 50 épocas. Em seguida, o modelo foi treinado a partir das imagens RGB. Nesse caso, o modelo que recebeu diretamente as imagens RGB obteve melhor resultado, alcançando 94.33%.

Figura 5.1: Resultado dos treinamentos do modelo com as entradas RGB-Flow e RGB — LIBRAS_APOEMA-50



Fonte: Próprio Autor.

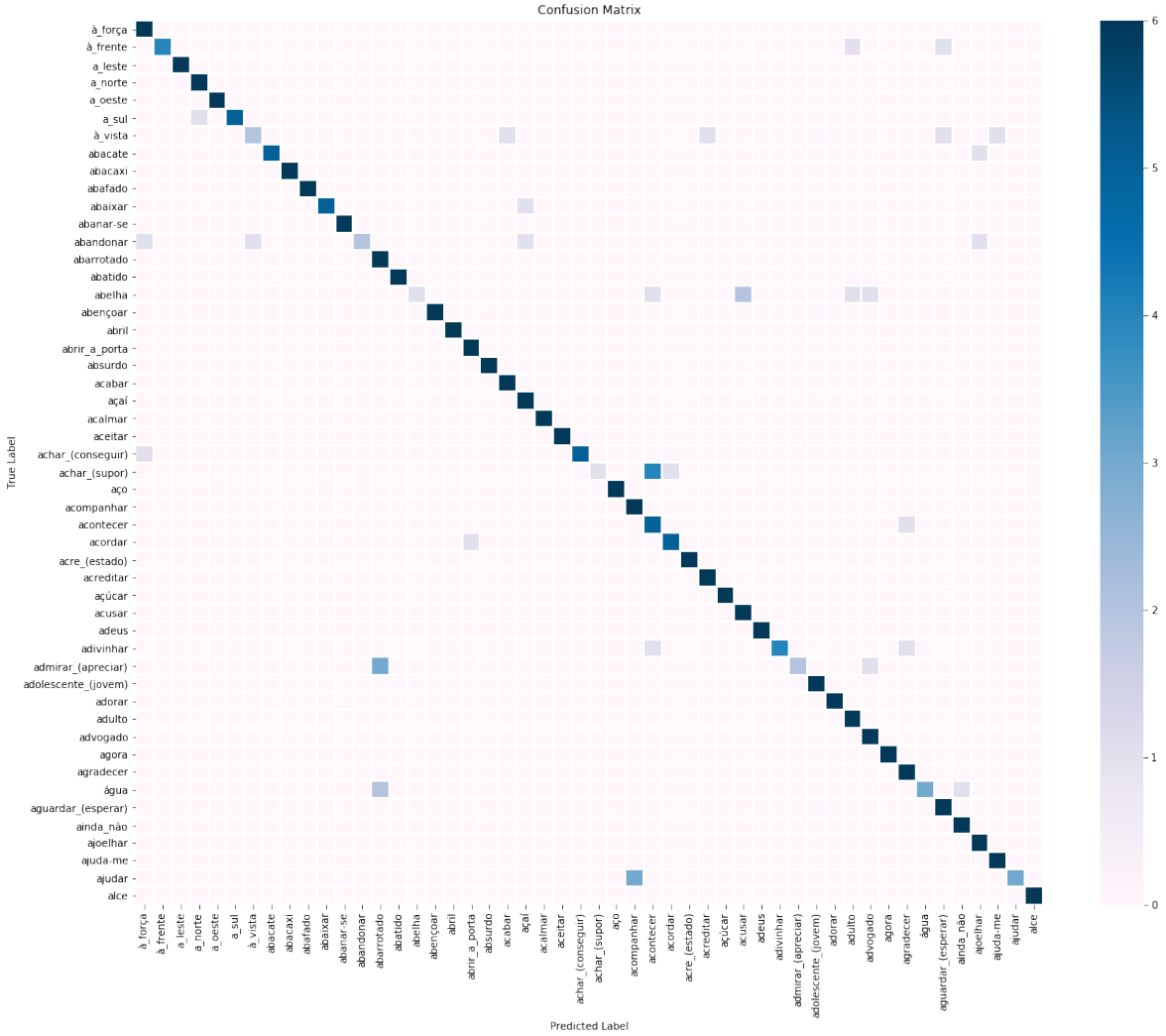
A matriz de confusão do modelo que apresentou melhor resultado é apresentada na Figura 5.2 em um mapa de calor, onde a intensidade da cor é diretamente proporcional à porcentagem de instâncias atribuídas à determinada classe. Ao analisar-se a matriz é possível perceber que o modelo se equivocou ao classificar, principalmente, as seguintes classes:

- o sinal ACHAR_(SUPOR) confundindo com o sinal ACONTECER;
- o sinal ADMIRAR_(APRECIAR) confundindo com o sinal ABARROTADO;
- e o sinal AJUDAR confundindo com o sinal ACOMPANHAR.

Ao analisar esses sinais, é possível observar que são sinais realmente muito similares, possuindo diferenças que uma sequência de imagens no espectro RGB não consegue captar com exatidão. As Figuras 5.3, 5.4 e 5.5 apresentam ilustrações que mostram essas similaridades. Uma hipótese é que talvez essa confusão possa ser diminuída ao associar-se os dados RGB aos dados de fluxo ótico.

A Tabela 5.1 apresenta uma relação dos resultados obtidos pelos trabalhos que envolvem reconhecimento de sinais dinâmicos da Libras e o resultado obtido neste trabalho para a amostra da base de 50 classes. É importante destacar que trata-se de uma

Figura 5.2: Matriz de confusão da amostra composta por 50 classes: dados RGB.

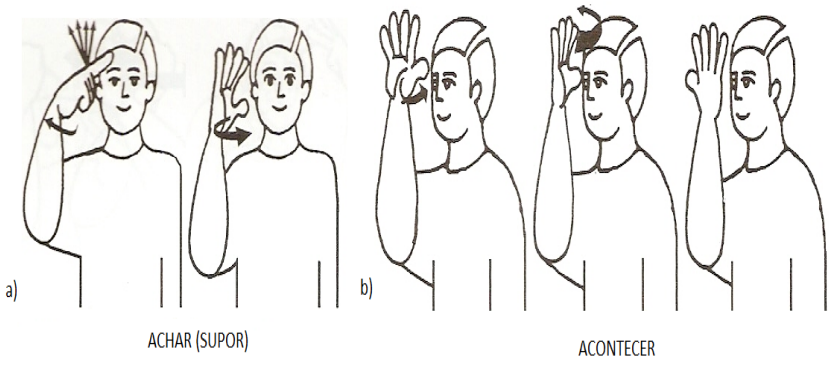


Fonte: Próprio Autor.

listagem apenas, dada a inviabilidade de replicação dos experimentos[2], para se obter um referencial das taxas encontradas na literatura. Apesar dos resultados de acurácia apresentados por Silva (2018) terem sido superiores aos obtidos pelo modelo I3D treinado com dados de RGB utilizado neste trabalho, é importante ressaltar que os dados foram adquiridos por dispositivos considerados intrusivos. Além disso, a base de dados utilizada pelo autor, que possui 6 sinais, contém uma quantidade de sinais muito menor do que a quantidade de sinais existentes na partição da base LIBRAS_APOEMA

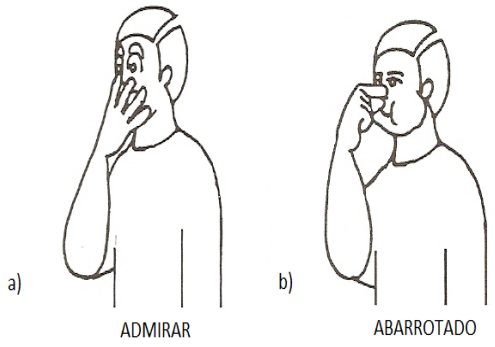
[2]Não foram encontradas publicamente as bases de dados, nem os modelos utilizados nesses trabalhos.

Figura 5.3: Execução dos sinais (a) ACHAR_(SUPOR) e (b) ACONTECER.



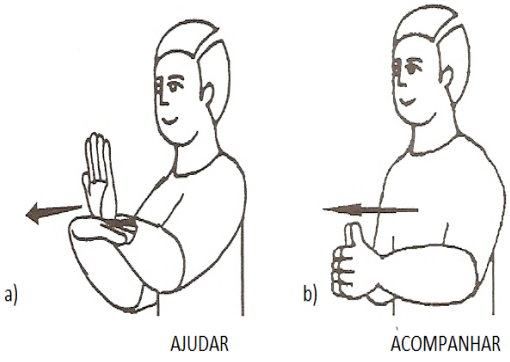
Fonte: Capovilla e Raphael (2001).

Figura 5.4: Execução dos sinais (a) ADMIRAR_(APRECIAR) e (b) ABARROTADO.



Fonte: Capovilla e Raphael (2001).

Figura 5.5: Execução dos sinais (a) AJUDAR e (b) ACOMPANHAR.



Fonte: Capovilla e Raphael (2001).

Tabela 5.1: Relação dos Trabalhos de Reconhecimento de Sinais Dinâmicos da Libras — Bases de até 50 Classes.

Autor	Quantidade de Sinais Dinâmicos	Aquisição	Modelo	Acurácia
Leal (2018)	5	LMC	MLP	86.7%
Silva (2018)	**6**	Sensores em luva	MLP	**97.4%**
Voigt (2018)	7	LMC	CNN+LSTM	92%
Proposta	**50**	Câmera simples (RGB)	3D-CNN	**94.33%**

utilizada nesta primeira série de experimentos. Ademais, a base LIBRAS_APOEMA possui 5 dentre os 6 sinais existentes na base de Silva (2018), para os quais o modelo treinado apenas com os dados RGB alcançou 100% de revocação.

5.2.2 Segunda Série de Experimentos: 84 classes

Para validação do modelo e a fim de comparar os resultados alcançados aos resultados apresentados no trabalho de Machado (2018), também foram realizados experimentos utilizando a amostra de 84 classes usada por esse autor, composta por 58 das primeiras palavras da base de LIBRAS_APOEMA, que iniciam com as letras entre A e D, além das 26 letras do alfabeto. É importante destacar que, mesmo que a amostra de dados seja a mesma, a divisão das partições de treino, validação e teste não é igual, o que significa que um intérprete que foi visto no treino pelo modelo apresentado por Machado (2018) pode estar na partição de validação deste trabalho ou o contrário. Ambos os trabalhos tomaram precauções para que todas as execuções de um intérprete compusessem apenas uma das partições. A escolha de intérpretes por partição foi aleatória, com a restrição de que os dois intérpretes surdos não pertencessem à mesma partição. Para essa etapa, considerou-se o cenário do experimento em que o modelo obteve melhor desempenho com a amostra de 50 classes.

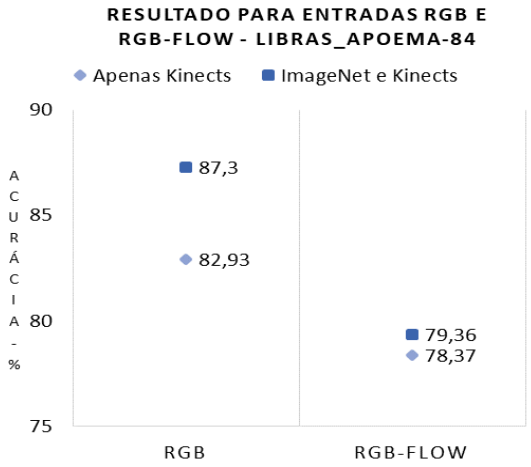
Além do conjunto de sinais da Libras, Machado (2018) realizou transferência de aprendizado, treinando previamente o modelo na base de dados IsoGD. Essa base de dados, considerada como referência em reconhecimento de gestos, é composta por vídeos RGB e de profundidade para 249 classes de gestos executados por 21 indivíduos (Wan et al., 2016). No total, há cerca de 48 mil vídeos de gestos dinâmicos isolados[3], que possuem uma média de 5 segundos de duração.

Nessa série de experimentos foram testadas duas estratégias de transferência de aprendizado: tal como Machado (2018) realizou transferência de aprendizado apenas de uma base de dados, composta por gestos, optou-se por realizar um experimento com

[3]Ou seja, não contempla gestos dinâmicos em frases.

transferência somente da base Kinetics, que contém ações, sem os pesos aprendidos na base ImageNet; em seguida, foram realizados experimentos com as mesmas entradas, adicionando os pesos aprendidos na base ImageNet, para verificar o impacto decorrente da utilização desse conjunto. A Figura 5.6 apresenta o resultado obtido pelo modelo com as entradas RGB e RGB-Flow. Como pode ser visto, o modelo com a entrada RGB alcançou melhor acurácia.

Figura 5.6: Resultados alcançados pelo modelo com as entradas RGB e RGB-Flow — LIBRAS_APOEMA-84.



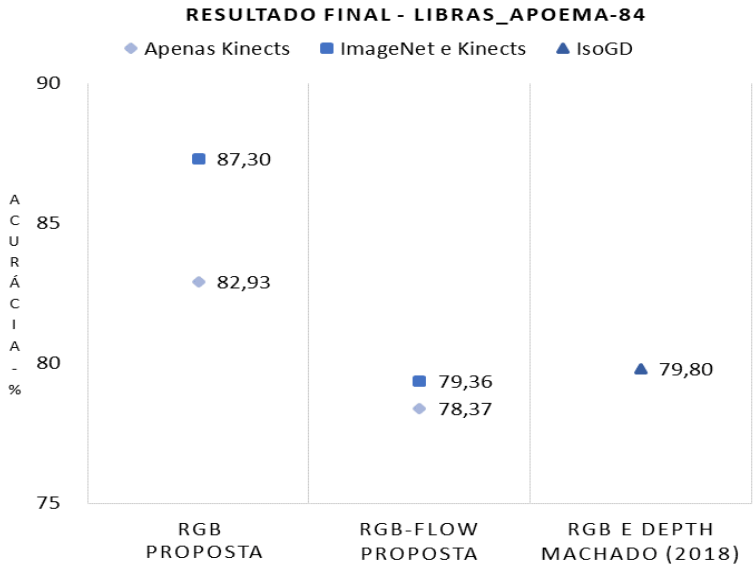
Fonte: Próprio Autor.

Na Figura 5.7 podem ser analisados os resultados finais dos experimentos, incluindo o resultado apresentado por Machado (2018). Como pode ser observado, o modelo apresentou melhor desempenho quando treinado a partir de imagens RGB, utilizando pesos aprendidos nas bases ImageNet e Kinetics. Esse é um comportamento esperado, pois apesar de a ImageNet representar um contexto diferente, essa base é muito maior em quantidade de instâncias e classes, fazendo com que conceitos simples sejam bem aprendidos.

A matriz de confusão é apresentada na Figura 5.11 em um mapa de calor, em que é possível perceber que as confusões ocorreram majoritariamente entre classes em que há elevada semelhança. Além disso, pode haver mudança na execução do sinal, dependendo do intérprete, acentuando a similaridade interclasse. As Figuras 5.8, 5.9 e 5.10 apresentam alguns exemplos para ilustrar esse ponto:

- o sinal de CARRO confundido com o sinal ANDAR_(DE-BICICLETA);

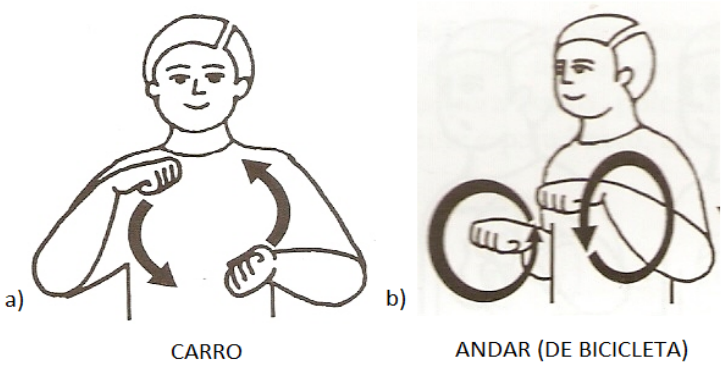
Figura 5.7: Resultado de acurácia obtida pelo modelo com as entradas RGB e RGB-Flow, variando as bases de transferência — LIBRAS_APOEMA-84.



Fonte: Próprio Autor.

- o sinal referente à letra D confundido com o sinal DEUS;
- e o sinal referente à letra R confundido com o sinal da letra K.

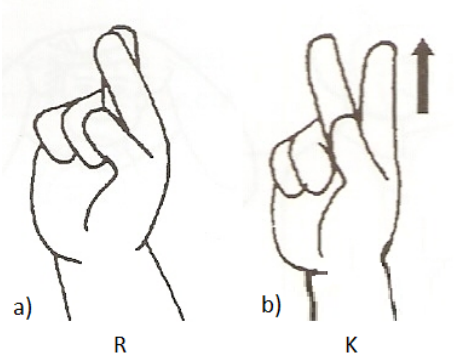
Figura 5.8: Execução dos sinais (a) CARRO e (b) ANDAR_(DE-BICICLETA).



Fonte: Capovilla e Raphael (2001).

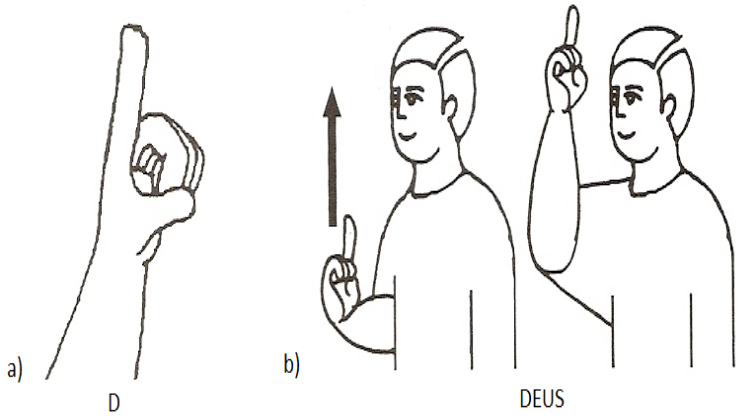
Na Tabela 5.2 pode ser visto um comparativo entre o resultado do modelo treinado com a base LIBRAS_APOEMA-84 e o resultado obtido por Machado (2018), em

Figura 5.9: Execução dos sinais (a) letra R e (b) letra K.



Fonte: Capovilla e Raphael (2001).

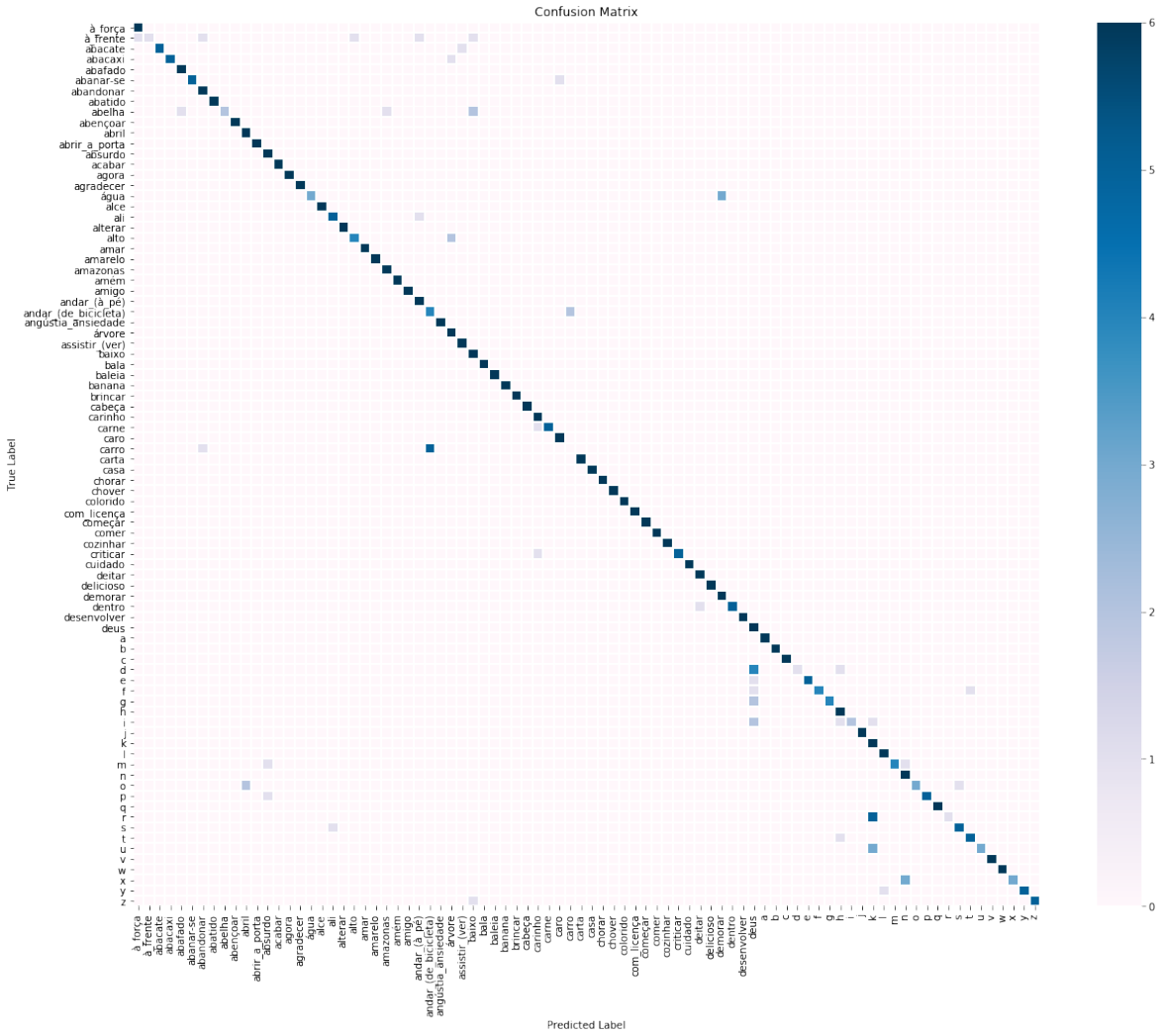
Figura 5.10: Execução dos sinais (a) letra D e (b) DEUS.



Fonte: Capovilla e Raphael (2001).

que são apresentados o tipo de entrada, a base em que o modelo foi pré-treinado, o percentual de acurácia obtido, a forma de aquisição dos dados, e o modelo utilizado. Como pode ser visualizado, o modelo de 3D-CNN utilizado nesta pesquisa obteve resultados superiores, ainda que utilizando somente os dados RGB, sem aplicação de aumento ou fusão de dados. Além disso, este trabalho tem como vantagem a aquisição de dados, que é realizada de maneira simples, em contraste ao sensor específico necessário para se adquirir os dados de profundidade utilizados por Machado (2018).

Figura 5.11: Matriz de confusão da amostra composta por 84 classes: dados RGB.



Fonte: Próprio Autor.

5.2.3 Terceira Série de Experimentos: 560 classes

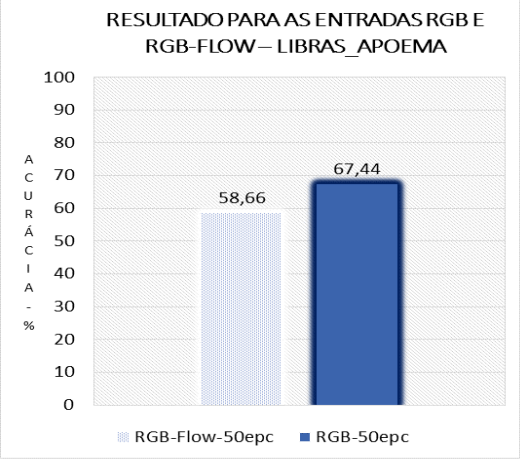
Por fim, foram realizados experimentos com a base de dados completa, mantendo os mesmos hiperparâmetros utilizados nas séries anteriores. Nesta terceira série de experimentos, o modelo também apresentou melhor resultado quando recebidas diretamente as sequências de dados RGB, conforme pode ser visualizado na Figura 5.12. Por outro lado, ao compararmos esses novos resultados com os resultados obtidos em partições menores da base, obtidos nas séries anteriores, percebe-se que o desempenho do modelo reduziu, em consequência do aumento do número de classes.

Para tentar estudar melhor a baixa taxa de acurácia, são exibidas na Figura 5.13

Tabela 5.2: Comparação para a Amostra de 84 classes.

Autor	Entrada	Transferência de Aprendizado	Resultados	Aquisição	Modelo
Machado (2018)	RGB+ DEPTH	IsoGD	**79.8%**	Kinect	3D-CNN+ LSTM
Proposta	RGB-FLOW	Kinetics	78.36%	Câmera simples (RGB)	3D-CNN (I3D)
	RGB-FLOW	ImageNet + Kinetics	79.37%		
	RGB	Kinetics	82.93%		
	RGB	ImageNet + Kinetics	**87.3%**		

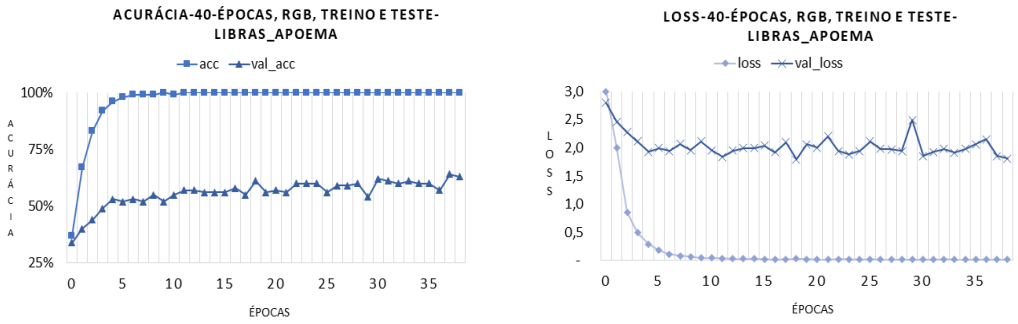
Figura 5.12: Resultados alcançados para a base LIBRAS_APOEMA.



Fonte: Próprio Autor.

as curvas de acurácia e das funções de perda nas partições de treino e validação. Nessa figura pode ser percebido que a diferença entre a curva de acurácia obtida na base de treino e a obtida na base de validação é muito alta, o que significa que o modelo apresentou superajuste aos dados de treino. De fato, a partir da época 5, os resultados de acurácia já estão muito próximos de 100%, e os da função perda chegando a 0, na partição de treino. Esse comportamento, entretanto, já estava previsto, visto que houve um aumento significativo do número de classes, mas a quantidade de instâncias por classe permaneceu a mesma. Para solucionar esse problema, foram aplicadas as técnicas previamente estabelecidas: aumento de dados e fusão de canais.

Figura 5.13: Gráficos de acurácia e função de perda no treino e na validação, utilizando dados RGB: treinamento de 40 épocas, LIBRAS_APOEMA.



Fonte: Próprio Autor.

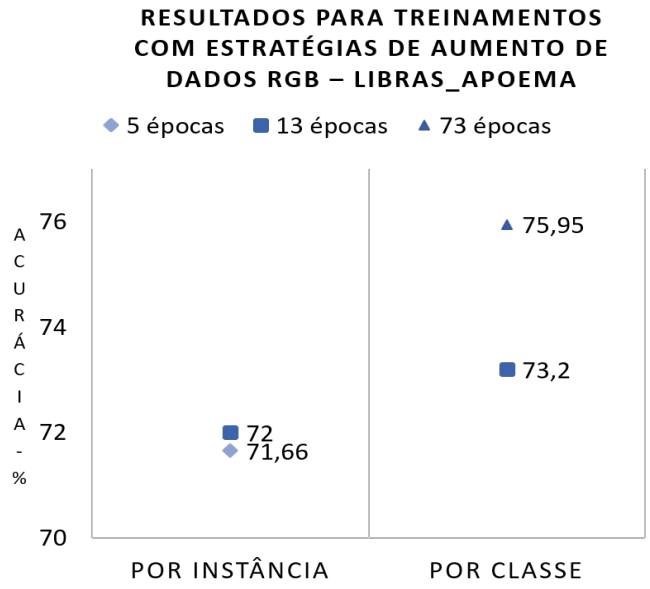
5.2.3.1 Aumento de dados

Duas estratégias de implementação foram consideradas para ampliação da base com aumento de dados: gerar transformações aleatórias por instância; gerar transformações aleatórias por classe. Para o primeiro cenário, foram geradas 20 transformações para cada instância da base de dados. Alcançou-se 71% de acurácia. Devido o longo tempo de treinamento (com a duração média de 10 horas por época), reconheceu-se a inviabilidade de se treinar com essa quantidade de transformações.

Por esses motivos, fixou-se a quantidade de 5 transformações, das 20 estabelecidas, para a implementação do segundo cenário: cada instância da base recebeu cinco transformações aleatórias fixadas por classe, ou seja, instâncias pertencentes à mesma classe receberam o mesmo tipo de transformação. Foram executadas 12 épocas para o ajuste. Os resultados chegam a 73,2% de acurácia. Para 100 épocas, o treinamento foi interrompido na época 73, após 20 épocas sem apresentar melhoria na acurácia, alcançando 75,95% de acurácia no teste.

A Figura 5.14 apresenta os resultados obtidos, em que pode ser verificado que, embora a diferença seja baixa, o melhor desempenho geral do modelo ocorreu quando treinado com a versão da base aumentada em 5 transformações aleatórias fixadas por classe. Entretanto, dada a pouca diferença entre os resultados de acurácia, não se pode afirmar que essa estratégia é melhor que a primeira, sendo necessário um estudo mais aprofundado sobre o assunto. Ainda assim, gerou-se, para essa versão da base, o conjunto com estimativas de fluxo ótico. Com o intuito de melhorar os resultados aproveitando melhor a informação disponível nos dados, o tamanho da sequência temporal foi alterado, de 20, para 30 quadros por sinal. Na Figura 5.15 pode ser visto o

Figura 5.14: Resultados obtidos após aplicação das estratégias de aumento de dados RGB — LIBRAS_APOEMA.



Fonte: Próprio Autor.

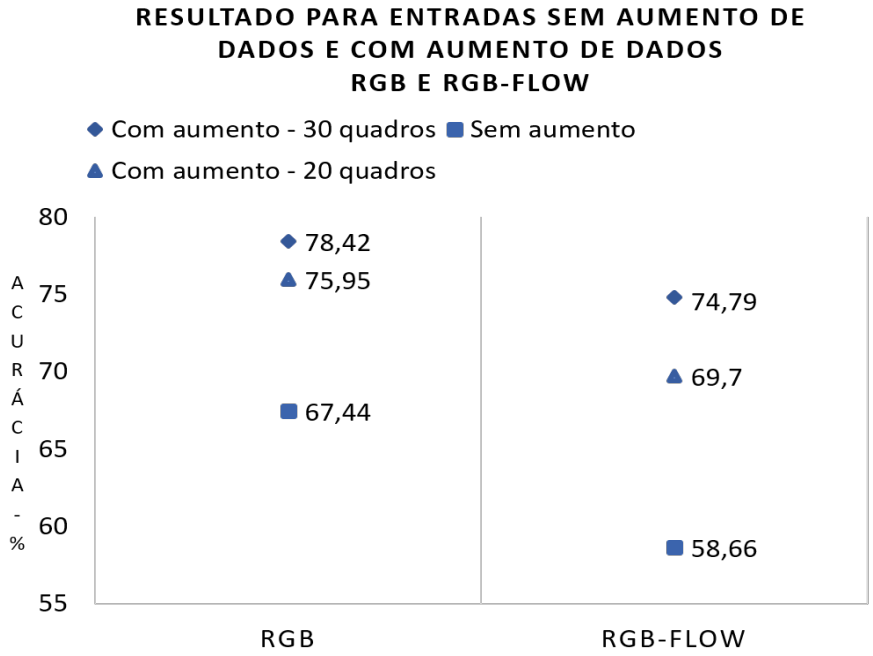
resultado para RGB e RGB-Flow, dos modelos individuais: sem aumento de dados e 20 quadros; com aumento de dados e 20 quadros; e com aumento de dados e 30 quadros.

Como pode ser observado, os resultados de acurácia de $75,95\%$ alcançados pelo modelo trouxeram um ganho estatístico de 8% em relação ao resultado de $67,44\%$ obtido pelo melhor modelo de classificação individual com dados RGB. Esse índice ainda melhorou em, aproximadamente, 3% com a alteração na quantidade de quadros por vídeo. Ainda maior foi a diferença entre os resultados de classificação do modelo treinado com dados RGB-Flow, que passaram de $58,66\%$ para $69,7\%$ com o aumento de dados, e depois para $74,79\%$ com a alteração na normalização do quadros, totalizando 16% de ganho estatístico. Portanto, pode ser afirmado que o aumento de dados foi um recurso fundamental de contribuição no combate ao superajuste do modelo aos dados de treino.

5.2.3.2 Fusão de Canais

A fase de integrar os resultados obtidos pelos modelos iniciou com a fusão em nível de características. Entretanto, não encontrou-se um conjunto de hiperparâmetros capaz de fazer com que o modelo conseguisse generalizar. Portanto, o foco principal durante essa etapa esteve voltado para a fusão em nível de decisão. Como mencionado, para

Figura 5.15: Comparação de resultado: sem aumento de dados e com aumento de dados — LIBRAS_APOEMA.



Fonte: Próprio Autor.

Tabela 5.3: Comparação Entre Resultados Individuais e Resultados Com Fusão

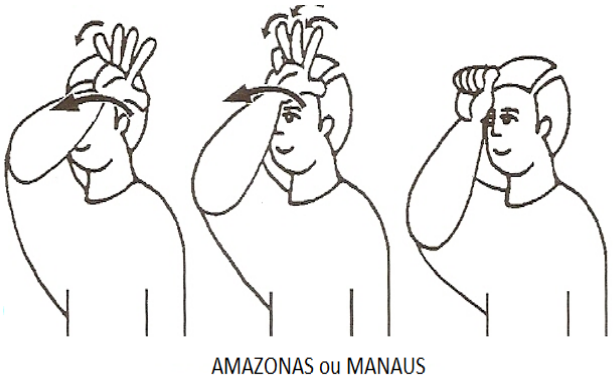
Classificação Final	**Entrada**	**Acurácia Top 1**
Individual	RGB	78.42%
	RGB-FLOW	74.79%
Com fusão	Máximo	81.69%
	Média	82.20%
	Produto	**83.75%**

a tomada de decisão, os vetores das probabilidades *a posteriori* foram integrados a partir do cálculo do valor máximo, do valor médio e do produto entre eles, resultando em um único vetor para cada operação. Um comparativo dos resultados obtidos pode ser visualizado na Tabela 5.3. Como pode ser observado, com os resultados de acurácia de $83,75\%$ para fusão de probabilidades utilizando a regra do produto, houve uma ganho estatístico de 5% em relação ao resultado de $78,42\%$ obtido pelo melhor modelo de classificação individual.

Também foi analisada a matriz de confusão do modelo resultante da estratégia

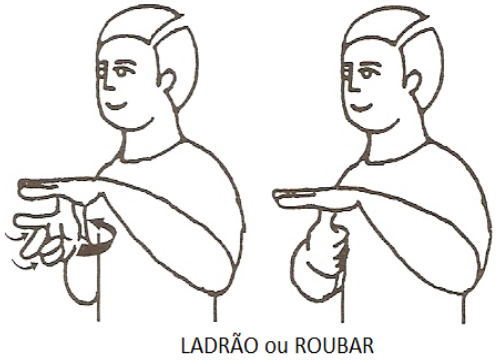
que obteve melhor acurácia — que, devido à grande quantidade de classes, não pôde ser adicionada a este texto. Nela, é possível observar que o modelo apresentou alta taxa de precisão para a maioria das classes. Entretanto, da mesma forma que ocorreu com as amostras de 50 e 84 sinais, alguns sinais foram confundidos especificamente com outros. A partir desse resultado, foram identificadas polissemias na base de dados, isto é, sinais com execução idêntica, mas que possuem significados distintos. Nesses casos, o significado real que o emissor pretende transmitir é dependente do contexto da fala. As Figuras 5.16, 5.17 e 5.18 ilustram três exemplos de polissemia: o sinal referente às palavras AMAZONAS e MANAUS; o sinal referente às palavras LADRÃO e ROUBAR; e o sinal referente à expressão DE_NOVO e à palavra OUTRO.

Figura 5.16: Ilustração do sinal que representa as palavras AMAZONAS ou MANAUS.



Fonte: Capovilla e Raphael (2001).

Figura 5.17: Ilustração do sinal que representa as palavras LADRÃO ou ROUBAR.



Fonte: Capovilla e Raphael (2001).

Diante disso, foi levantada a hipótese de que, em uma situação como essa, as probabilidades podem ser muito próximas, visto que o modelo pode estar muito con-

Figura 5.18: Ilustração do sinal que representa a expressão DE_NOVO ou a palavra OUTRO.

Fonte: Capovilla e Raphael (2001).

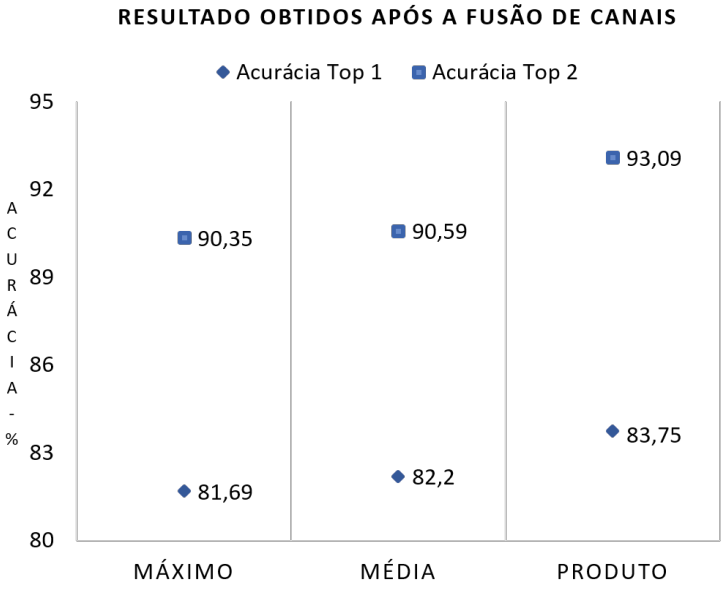
Tabela 5.4: Exemplos de Cenários de Rejeição.

	y_1	y_2	d	$t = 1$	$t = 5$	$t = 7$	$t = 10$
Instância 1	38.42%	37.75%	0.67%	Rejeita	Rejeita	Rejeita	Rejeita
Instância 2	52.27%	47.42%	4.85%	-	Rejeita	Rejeita	Rejeita
Instância 3	45.96%	38.98%	6.98%	-	-	Rejeita	Rejeita
Instância 4	54.30%	45.50%	8.80%	-	-	-	Rejeita
Instância 5	60.01%	39.81%	20,20%	-	-	-	-
Instância 6	33.57%	32.64%	4,93%	-	Rejeita	Rejeita	Rejeita
Instância 7	55.30%	45,33%	9,97%	-	-	-	Rejeita
		Total de Rejeições		**1**	**3**	**4**	**6**

fuso ao atribuir uma instância à determinada classe. Assim, para fins de análise, duas alternativas foram implementadas: a análise da acurácia Top 2, para saber se a classe alvo está entre y_1 e y_2, as duas maiores probabilidades previstas; e — para o classificador com a melhor acurácia Top 2 — a análise acurácia-rejeição, para verificação do comportamento do classificador diante da possibilidade de não classificar uma instância caso a diferença d não ultrapasse o limiar t, onde d corresponde à diferença entre y_1 e y_2. Neste trabalho, o limiar t pode assumir valores entre 0% e 10%. Assim, caso a diferença seja $0 < d <= t$, a classificação da instância é rejeitada. A Tabela 5.4 apresenta alguns cenários para exemplificar como essa análise é feita. Nesse exemplo, pode ser percebido que o valor de t representa o quão flexível o classificador é ao comparar as duas primeiras probabilidades. As Figuras 5.19 e 5.20 apresentam os resultados finais dessas análises.

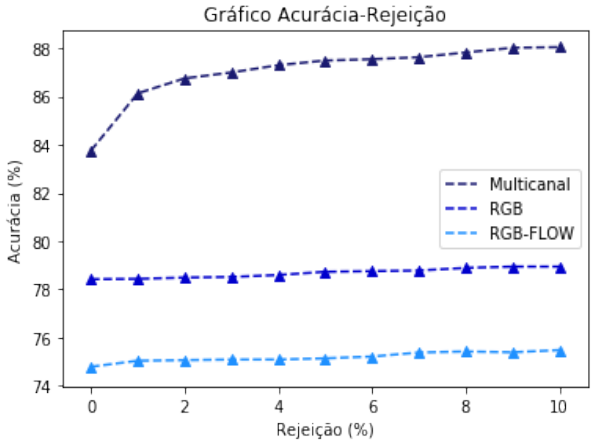
Os resultados apresentados na Figura 5.20 mostram que, de fato, há uma porcentagem significativa de instâncias para as quais as probabilidades mais altas são bem

Figura 5.19: Resultados de acurácia para as três estratégias de fusão de decisão — LIBRAS_APOEMA.



Fonte: Próprio Autor.

Figura 5.20: Análise da relação Acurácia-Rejeição — LIBRAS_APOEMA.



Fonte: Próprio Autor.

próximas, com a diferença de até 10% entre elas. Ainda nesse gráfico, pode ser observado que o classificador multicanal foi muito mais beneficiado com a possibilidade de rejeição. Isso deve-se ao fato de que o classificador em questão é resultante do produto das probabilidades *a posteriori* dos modelos individuais, regra que acentua a diferença entre as probabilidades.

Entretanto, ao se comparar os resultados de acurácia, da relação acurácia-rejeição, de 88,04%, aos da acurácia Top 2, de 93,09%, não é possível comprovar a hipótese de que o classificador ficou severamente dividido entre duas classes. Porém, foi confirmado que, em pelo menos 80% dos casos de polissemia, o modelo apresenta acima de 75% de certeza ao atribuir uma instância à uma classe. Sendo assim, em um sistema de RLS, é mais apropriado que, para casos de polissemia, o modelo faça sugestões das prováveis classes, visto que não há como saber o real significado do sinal sem ter o conhecimento do contexto da fala. Vale ressaltar que, a alternativa de sugerir um significado não é novidade, tendo em vista que essa é a maneira com a qual os dicionários tratam as palavras polissêmicas das línguas faladas.

5.3 Considerações Finais

Considerando o conteúdo desta seção, pode ser concluído que as três técnicas principais nas quais esta pesquisa focou melhoraram significativamente as taxas de acurácia em reconhecimento de sinais da Libras.

Entretanto, sabendo dos resultados obtidos, pode ser levantado um questionamento sobre as taxas alcançadas, se estão dentro do esperado na tarefa de reconhecimento de línguas de sinais de outros países, ou de reconhecimento de ações, por exemplo. Em resposta, os resultados podem ser comparados: aos obtidos pelos trabalhos apresentados na seção de RLS de outros países (Seção 3.2), onde as taxas de acurácia para os trabalhos que utilizam somente dados RGB variam de 74% a 99%, sendo esse último resultado obtido ao classificar somente 40 sinais; e — embora devam ser consideradas as peculiaridades de cada base de dados — aos obtidos no trabalho de Carreira e Zisserman (2017), em que modelos multimodais alcançam 97.9% e 80.2% de acurácia em bases de ações, compostas por no máximo 101 classes de ações.

Capítulo 6

Conclusão

Este trabalho apresenta uma estratégia de classificação de sinais da Libras. Após estudo inicial dos trabalhos relacionados, constatou-se que há um grande incentivo e interesse no desenvolvimento e aprimoramento de modelos para reconhecimento de língua de sinais, por se tratar de uma questão de grande impacto social, além de se levar em consideração a legislação vigente no Brasil. Entretanto, foi detectado que, para treinamento de modelos de redes neurais profundas, não há uma base de dados de sinais em Libras disponível publicamente, que apresente dados simples, como vídeo em RGB, ou mais específicos, como de profundidade ou esqueleto. Além disso, há poucos trabalhos que investigam a classificação de sinais dinâmicos, e quando o fazem, poucos são os sinais considerados.

A partir desses trabalhos foi possível identificar padrões candidatos a serem utilizados nesta pesquisa, relacionados ao tipo de dado, de modelagem temporal, e de transferência de aprendizado. Diante desse cenário, uma investigação foi realizada neste trabalho para definir-se qual desses padrões se adapta melhor ao reconhecimento de sinais da Libras, considerando a quantidade de classes apresentada na base de dados disponível. O resultado dessa investigação foi: em relação ao tipo de dado, foram utilizados dados RGB e de fluxo ótico estimado do RGB; em relação à modelagem temporal, foi utilizado o modelo de 3D-CNN (Carreira e Zisserman, 2017), que compõe o estado da arte em reconhecimento de ações; para a transferência de aprendizado, considerou-se a transferência de tarefas de reconhecimento de ações. Portanto, nesse capítulo serão feitas as considerações finais sobre os resultados alcançados e as perspectivas para o futuro desta pesquisa.

6.1 Considerações Finais

Neste trabalho é abordada a tarefa de reconhecimento de línguas de sinais, considerando sinais isolados. Como mencionado, a tarefa de reconhecimento de gestos em si não é simples, visto que muitos fatores podem influenciar a classificação do gesto. Especialmente no âmbito das línguas de sinais, os parâmetros constituintes dos sinais podem ser considerados complexos, principalmente por incluírem movimento ou por induzirem uma semelhança entre sinais diferentes. Por esses motivos, conforme apresentado no Capítulo 1, ainda há algumas lacunas em aberto nesta área, relacionadas principalmente a três fatores: os métodos atuais são limitados à classificação de poucos sinais; ou demandam alto custo financeiro de implantação; ou são intrusivos.

Portanto, esta pesquisa identificou uma estratégia, validada experimentalmente, para reconhecimento de sinais estáticos e dinâmicos da Libras combinando informações sobre dimensões espaciais e temporais, aumento de dados, fusão de dados de múltiplos canais e transferência de aprendizado, por meio de um modelo de 3D-CNN.

Em relação à quantidade se sinais classificados, a base de dados utilizada neste trabalho é superior a todos os trabalhos estudados. Quanto ao custo financeiro, este trabalho é de baixo custo, visto que foram utilizados como fonte de dados apenas sequências de imagens em RGB. Quanto à categorização, isto é, se é intrusivo ou não, trata-se de uma abordagem não intrusiva, pois é baseada em visão computacional sem a utilização de sensores portáteis.

Ainda destacamos que este trabalho diferencia-se dos demais ao aplicar um modelo pertencente ao atual estado da arte em reconhecimento de ações ao reconhecimento da Libras, identificando a melhor estratégia para treinamento, considerando os tipos de dados de entrada, aumento de dados, transferência de aprendizado, e as formas de fusão de dados, abordagem não encontrada na literatura.

6.1.1 Limitações

A carência de bases de dados de sinais da Libras é um dos limitadores principais do avanço de pesquisas nessa área. No caso deste trabalho, foi utilizada a base de dados LIBRAS_APOEMA, que foi filmada em ambiente padronizado, com fundo estático. Isso gera uma limitação no modelo, pois faz com que ele aprenda padrões específicos dessa base de dados, prejudicando o desempenho do modelo se utilizado em um sistema de RLS atuante em ambiente não controlado.

Outra limitação deste trabalho é a falta de investigação sobre a fusão de dados em nível de características. Devido aos longos períodos de treinamento, foram inter-

rompidas as tentativas de encontrar um conjunto de hiperparâmetros que fizesse com que a fusão de dados em nível de características obtivesse bons resultados.

Também não foram realizadas investigações sobre transferências de aprendizado do modelo treinado em bases de línguas de sinais de outros países.

6.1.2 Trabalhos Futuros

Esta pesquisa deu ênfase aos benefícios obtidos por meio da fusão de diferentes canais. Uma das principais vantagens da abordagem multicanal empregada é a fácil viabilidade de implementação. Primeiramente, por envolver visão computacional, o que significa que não é uma abordagem intrusiva. Em segundo lugar, por não serem necessários sensores específicos para a aquisição de dados, visto que uma versão RGB do vídeo já é suficiente, a qual pode ser obtida por meio de uma câmera simples.

Por outro lado, ainda no campo de visão computacional, uma abordagem multimodal, que envolva a utilização de dados adquiridos por diferentes fontes de dados, obtidos por diferentes sensores, como dados de profundidade ou de pontos de articulações do esqueleto (apresentados mais detalhadamente na Seção 2.2), pode apresentar resultados ainda melhores. Portanto, pretende-se ampliar a base de dados para que componha informações dos pontos de articulação do emissor, por serem informações mais precisas.

Além disso, como mencionado, a estratégia de fusão em nível de características tem se mostrado eficaz, apesar de que, durante o desenvolvimento deste trabalho, a mesma não foi amplamente estudada. Entretanto, para trabalhos futuros, pretende-se focar nessa categoria de fusão.

Outra possibilidade é a utilização de diferentes tipos de modelagem temporal: no âmbito das redes convolutivas, a arquitetura ResNet3D também pode ser aplicada ao reconhecimento e classificação de vídeos; no contexto das redes com memória, as BiLSTM e as GRUs. Nesse caso, o objetivo seria comparar os resultados dos modelos, ou associá-los, formando um comitê de classificadores mais robusto.

Referências Bibliográficas

Asadi-Aghbolaghi, M., Clapés, A., Bellantonio, M., Escalante, H. J., Ponce-López, V., Baró, X., Guyon, I., Kasaei, S., e Escalera, S. (2017). Deep learning for action and gesture recognition in image sequences: A survey. Em *Gesture Recognition*, pgs. 539–578. Springer.

Atrey, P. K., Hossain, M. A., El Saddik, A., e Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379.

Barros Junior, J. D. (2016). Tradução automática de línguas de sinais: do sinal para a escrita. *Dissertação - Universidade Federal do Pampa*.

Bracewell, R. N. (1986). *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York.

Brasil (1999). Decreto nº 3.298, de 20 de dezembro de 1999. regulamenta a lei nº 7.853, de 24 de outubro de 1989, dispõe sobre a política nacional para a integração da pessoa portadora de deficiência, consolida as normas de proteção, e dá outras providências. *Diário Oficial da União*.

Brasil (2002). Lei nº 10.436, de 24 de abril de 2002. dispõe sobre a língua brasileira de sinais-libras e dá outras providências. *Diário Oficial da União*.

Brasil (2003). Portaria nº 3.284, de 2003.

Brasil (2005). G. f. do. decreto no 5.626, de 22 de dezembro de 2005.

Camgoz, N. C., Hadfield, S., Koller, O., e Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. Em *2017 IEEE International Conference on Computer Vision (ICCV)*, pgs. 3075–3084.

Capovilla, F. C. e Raphael, W. D. (2001). *Dicionário enciclopédico ilustrado trilíngüe da língua de sinais brasileira: sinais de M a Z*, volume 2. EdUSP.

Capovilla, F. C. e Raphael, W. D. (2004). *Enciclopédia da língua de sinais brasileiras: o mundo do surdo em libras*, volume 8. Edusp.

Carreira, J. e Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. Em *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pgs. 6299–6308.

Censo, I. (2010). Disponível em : <http://www.censo2010.ibge.gov.br/>. Acesso em 05 de setembro de 2018.

Cui, R., Liu, H., e Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. Em *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pgs. 1610–1618.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763.

Felipe, T. A. (2007). Libras em contexto: curso básico. *Livro do estudante. Brasília*, 8.

Felipe, T. A. e Monteiro, M. S. (2007). Libras em contexto: curso básico. *Livro do professor. Brasília*, 7.

Gonçalves, L. C., Saad, E. F., Andrade, R. B., Romero, B. A., e de Campos, R. D. (2016). Redes neurais artificiais e processamento de imagem no reconhecimento de libras, usando o kinect. *Jornal de Engenharia, Tecnologia e Meio Abiente - JETMA*, 1(1):32–37.

Gonzalez, R. C., Woods, R. E., e Eddins, S. L. (2004). *Digital image processing using MATLAB*. Pearson Education India.

Goodfellow, I., Bengio, Y., Courville, A., e Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Graves, A., Fernández, S., Gomez, F., e Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Em *Proceedings of the 23rd international conference on Machine learning*, pgs. 369–376. ACM.

Gunes, H. e Piccardi, M. (2005). Affect recognition from face and body: early fusion vs. late fusion. Em *2005 IEEE international conference on systems, man and cybernetics*, volume 4, pgs. 3437–3443. IEEE.

Hara, K., Kataoka, H., e Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? Em *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pgs. 6546–6555.

Haykin, S. (2007). *Redes neurais: princípios e prática*. Bookman Editora.

He, K., Zhang, X., Ren, S., e Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916.

Horn, B. K. e Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185–203.

JI, S., Xu, W., Yang, M., e Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

Jung, A. B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F.-M., Weng, C.-H., Ayala-Acevedo, A., Meudec, R., Laporte, M., et al. (2020). Imgaug. Disponível em : <https://github.com/aleju/imgaug>. Acesso em 05 de outubro de 2019.

Kaur, P., Ganguly, P., Verma, S., e Bansal, N. (2018). Bridging the communication gap: With real time sign language translation. Em *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pgs. 485–490.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kishore, P. V. V., Prasad, M. V. D., Kumar, D. A., e Sastry, A. S. C. S. (2016). Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. Em *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pgs. 346–351.

Kitchenham, B. e Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *Technical Report, Evidence-Based Software Engineering (EBSE)*, 2.3(4):43.

Kittler, J., Hatef, M., Duin, R. P., e Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239.

Klima, E. S. e Bellugi, U. (1979). *The signs of language*. Harvard University Press.

Koller, O., Ney, H., e Bowden, R. (2016). Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. Em *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pgs. 3793–3802.

Kopuklu, O., Kose, N., e Rigoll, G. (2018). Motion fused frames: Data level fusion strategy for hand gesture recognition. Em *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pgs. 2103–2111.

Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Em *Advances in neural information processing systems*, pgs. 1097–1105.

Kumar, E. K., Kishore, P. V. V., Sastry, A. S. C. S., Kumar, M. T. K., e Kumar, D. A. (2018). Training cnns for 3-d sign language recognition with color texture coded joint angular displacement maps. *IEEE Signal Processing Letters*, 25(5):645–649.

Leal, M. M. (2018). Singapp: um modelo de identificação de língua de sinais através de captura de movimento em tempo real. *Dissertação - Universidade do Vale do Rio dos Sinos*.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

Lin, M., Chen, Q., e Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Lira, G. e Souza, T. (2018). Dicionario da língua brasileira de sinais. 1. Disponível em: http://www. dicionarioLibras.com.br. Acesso em 05 de setembro de 2018.

Liu, T., Zhou, W., e Li, H. (2016). Sign language recognition with long short-term memory. Em *2016 IEEE International Conference on Image Processing (ICIP)*, pgs. 2871–2875.

Machado, M. C. (2018). Classificação automática de sinais visuais da língua brasileira de sinais representados por caracterização espaço-temporal. *Dissertação - Universidade Federal do Amazonas*.

Magalhaes, G. I. (2018). Reconhecimento de gestos da língua brasileira de sinais através de redes neurais. *Dissertação - Instituto Tecnológico de Aeronáutica*.

Mattelart, A. e Mattelart, M. (2011). *História das teorias da comunicação*. Edições Loyola.

Maturana, D. e Scherer, S. (2015). Voxnet: A 3d convolutional neural network for real-time object recognition. Em *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pgs. 922–928. IEEE.

Mikołajczyk, A. e Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. Em *2018 international interdisciplinary PhD workshop (IIPhDW)*, pgs. 117–122. IEEE.

Mohandes, M., Deriche, M. A., e Aliyu, S. O. (2017). Arabic sign language recognition using multi-sensor data fusion. US Patent 9,672,418.

Munib, Q., Habeeb, M., Takruri, B., e Al-Malik, H. A. (2007). American sign language (asl) recognition based on hough transform and neural networks. *Expert systems with Applications*, 32(1):24–37.

Narayana, P., Beveridge, J. R., e Draper, B. A. (2019). Analyzing multi-channel networks for gesture recognition. Em *2019 International Joint Conference on Neural Networks (IJCNN)*, pgs. 1–8. IEEE.

Oprea, S., Garcia-Garcia, A., Garcia-Rodriguez, J., Orts-Escolano, S., e Cazorla, M. (2017). A recurrent neural network based schaeffer gesture recognition system. Em *2017 International Joint Conference on Neural Networks (IJCNN)*, pgs. 425–431.

Pan, S. J. e Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Quadros, R. M. (2004). *O tradutor e intérprete de língua brasileira de sinais e língua portuguesa*. SEESP.

Rao, G. A., Syamala, K., Kishore, P. V. V., e Sastry, A. S. C. S. (2018). Deep convolutional neural networks for sign language recognition. Em *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, pgs. 194–197.

Schunck, B. G. (1989). Image flow segmentation and estimation by constraint line clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(10):1010–1027.

Shin, S., Baek, Y., Lee, J., Eun, Y., e Son, S. H. (2017). Korean sign language recognition using emg and imu sensors based on group-dependent nn models. Em *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pgs. 1–7. IEEE.

Silva, B. C. R. (2018). Desenvolvimento de tecnologia baseada em redes neurais artificiais para reconhecimento de gestos da língua de sinais. *Dissertação - Universidade Federal de Goiás*.

Simonyan, K. e Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Em *Advances in neural information processing systems*, pgs. 568–576.

Snoek, C. G., Worring, M., e Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. Em *Proceedings of the 13th annual ACM international conference on Multimedia*, pgs. 399–402.

Stokoe, W. C. (1960). Sign language structure: An outline of the visual communication systems of the american deaf. *Studies in Linguistics. Occasional paper*, 8.

Stokoe Jr, W. C. (2005). Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., e Rabinovich, A. (2015). Going deeper with convolutions. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, pgs. 1–9.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., e Liu, C. (2018). A survey on deep transfer learning. Em *International conference on artificial neural networks*, pgs. 270–279. Springer.

Taylor, L. e Nitschke, G. (2017). Improving deep learning using generic data augmentation. *arXiv preprint arXiv:1708.06020*.

Voigt, J. F. (2018). Aprendizagem profunda para reconhecimento de gestos da mão usando imagens e esqueletos com aplicações em libras. *Dissertação - Universidade Federal de Alagoas*.

Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., e Li, S. Z. (2016). Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. Em

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pgs. 56–64.

Xavier, A. e Barbosa, P. (2014). Os efeitos semânticos da duplicação do número de mãos na produção de sinais da língua brasileira de sinais (libras).

Xavier, A. N. (2006). *Descrição fonético-fonológica dos sinais da língua de sinais brasileira (LIBRAS)*. PhD thesis, Universidade de São Paulo.

Yang, S. e Zhu, Q. (2017). Video-based chinese sign language recognition using convolutional neural network. Em *2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)*, pgs. 929–934.

Yasir, F., Prasad, P. W. C., Alsadoon, A., Elchouemi, A., e Sreedharan, S. (2017). Bangla sign language recognition using convolutional neural network. Em *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pgs. 49–53.

Zach, C., Pock, T., e Bischof, H. (2007). A duality based approach for realtime tv-l 1 optical flow. Em *Joint pattern recognition symposium*, pgs. 214–223. Springer.

Zheng, L., Liang, B., e Jiang, A. (2017). Recent advances of deep learning for sign language recognition. Em *Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on*, pgs. 1–7. IEEE.