Aprendizado profundo no reconhecimento de sinais estáticos de Libras

Eros G. A. Caiafa[1], Fabiana F. Fonseca[2], Amaro A. Lima[1], Gabriel M. Araujo[1], Eduardo A. B. da Silva[2],

***Resumo*— Um problema bem comum, talvez o mais comum entre as pessoas com deficiência auditiva, é terem dificuldade em interagir com outros indivíduos, pois mesmo tendo uma língua para a sua integração, poucos a conhecem. Muitos trabalhos tem usado soluções tecnológicas pra esse problema, mas a maioria possui algum custo financeiro (por usar dispositivos para vestir) ou são atualmente limitados (visão computacional). Neste trabalho, tentamos avançar em soluções usando visão computacional para o problema de reconhecimento de gestos em Libras (Lingua Brasileira de Sinais). Para tanto, compilamos alguns trabalhos recentes na área, bem como mostramos algumas simulações usando aprendizado profundo em bases de dados contendo gestos ou sinais estáticos em Libras. As simulações contém uma comparação das arquiteturas LeNet, InceptionV3, VGG e ResNET no reconhecimento dos sinais em duas bases de dados, atingindo acurácias maiores que 99%.**

***Palavras-Chave*— Libras, Reconhecimento de Gestos, Visão Computacional, Acessibilidade.**

***Abstract*— A very common problem, perhaps the most common among the hearing impaired, is that they have difficulty to communicate with other people because even though they have a sign language for their integration, few individuals know it. Many works proposed technological solutions to these problems, but most of them are expensive (e.g., wearable gadgets) or limited (e.g., computer vision). In this paper, we try to push forward a computer vision solution for the problem of gesture recognition in Libras (Brazilian Sign Language). In order to do so, we compile some recent works in the area as well as show some simulations for deep learning in static signs or gestures datasets. Simulations contain a comparison of the LeNet, VGG, InceptionV3 and ResNet architectures on two datasets, reaching accuracies higher than 99%.**

***Keywords*— Libras, Gesture Recognition, Computer Vision, Accessibility.**

I. INTRODUÇÃO

De acordo com o último censo realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), o Brasil possui mais de 9,7 milhões de pessoas com alguma deficiência auditiva [1]. Pessoas surdas são consideradas uma minoria linguística. No Brasil essas pessoas fazem uso da Língua Brasileira de Sinais (Libras) para se comunicar. Como toda língua de sinais [2], Libras é considerada uma língua natural pois atende todos os critérios linguísticos, a saber: léxico, sintaxe e capacidade de gerar infinitas sentenças [3]. A Libras foi instituída como segunda língua oficial do Brasil em 2002 [4]. Entretanto, a comunicação é um problema presente na sociedade brasileira, já que poucas pessoas conhecem a Libras, principalmente entre os não-surdos.

Diversos estudos e soluções tecnológicas foram concebidos com o intuito de minimizar esse problema. Alguns fazem uso de dispositivos que podem ser vestidos pelo usuário, outros utilizam visão computacional. As soluções que fazem uso de dispositivos que podem ser vestidos tendem a ser bastante assertivas, mas possuem dois problemas principais: são intrusivos e possuem alto custo financeiro. Por outro lado, soluções que utilizam visão computacional podem ser muito baratas e não intrusivas, mas ainda são limitadas em termos de aplicabilidade ou confiabilidade. Considerações mais aprofundadas acerca dessas tecnologias aplicadas no reconhecimento de Libras podem ser vistas na Seção III.

Um dos principais problemas do uso de visão computacional em Libras, sobretudo utilizando técnicas baseadas em aprendizado profundo, está justamente nas bases de dados disponíveis [5]. Existem dois dicionários *online* com muitos verbetes em vídeo, mas apresentam apenas uma amostra por verbete em baixa resolução e ambiente controlado (os dicionários foram projetados para o ensino). A maioria dos trabalhos na área acaba criando as próprias bases de dados, em geral contendo somente as configurações de mão empregadas durante a execução dos gestos, ou então poucos sinais, no máximo centenas de sequências com até algumas dezenas de palavras diferentes. Isso dificulta a criação de aplicações mais abrangentes e/ou a comparação entre os trabalhos.

Ainda assim, existem algumas bases de dados de sinais estáticos ou contendo configurações de mão. Algumas dessas bases de dados estão consolidadas com trabalhos que descrevem métodos com bons resultados nessas bases, facilitando a comparação de desempenho. Métodos que reconhecem sinais estáticos não são suficientes para efetuar a transcrição Libras-Português, mas podem consistir numa etapa importante para o desenvolvimento dessa aplicação. Nesse trabalho, são avaliados os desempenhos de arquiteturas neurais profundas bastante conhecidas nas bases consolidadas de sinais estáticos de Libras. Em muitos casos, os resultados obtidos chegam próximos a 100% de acurácia na classificação dos sinais. Além disso, esse trabalho apresenta uma extensa revisão de literatura de trabalhos recentes bem como faz uma comparação das bases de dados disponíveis.

Um resumo de conceitos fundamentais para a Libras são apresentados na Seção II. A Seção III contém uma revisão da literatura recente sobre o assunto. As principais bases de dados de sinais estáticos de Libras estão descritas nas Seção IV. Na Seção V discutimos sobre a aplicação de aprendizado profundo em sinais estáticos e apresentamos alguns resultados encorajadores. As conclusões estão na Seção VI.

II. ALGUMAS CARACTERÍSTICAS DA LIBRAS

Libras é uma língua da modalidade espaço-visual, pois é produzida pelas mãos e recebida pelos olhos. Entretanto, possui os mesmos princípios de qualquer outra língua, como um léxico (símbolos convencionais) e uma gramática. Os símbolos são decompostos nos seguintes parâmetros: configuração da mão, movimento, locação, orientação da mão e expressões

[1]Centro Federal de Educação Tecnológica, Nova Iguaçu, RJ, 26041-271, Brasil.[2]PEE/COPPE/DEL/POLI, Universidade Federal do Rio de Janeiro, Cx. P. 68504, Rio de Janeiro, RJ, 21945-970, Brasil. E-mails: erosg17@gmail.com, fabiana.ferreira@poli.ufrj.br, {amaro.lima, gabriel.araujo}@cefet-rj.br, eduardo@smt.ufrj.br

não manuais. Nas línguas de sinais, esses parâmetros não possuem significado isoladamente e são representados de maneira simultânea [6]. A Tabela I contém um resumo destes parâmetros e como eles são manifestados.

Parâmetros	**Como são manifestados**
Configuração da mão	Gesto feito com a mão
Movimento	Ação da mão
Locação	Local onde as mãos realizam ação
Orientação da mão	Direção da palma na mão
Expressões não manuais	Movimentos da face, cabeça ou tronco

TABELA I: Paramâmetros de Libras e suas manifestações [6].

Cada um dos parâmetros resumidos na Tabela I possui diversas configurações, que variam de acordo com a literatura sobre o assunto. Por exemplo, no trabalho em [6], foram abordadas 74 configurações de mão (alguns autores reportam até 81 configurações); diversas locações (por exemplo, a ponta do nariz, testa, frente do tronco etc) [3]; 4 categorias de movimento e 11 subcategorias de acordo com [7] (por exemplo, contato, dobramento de pulso, repetição etc); 6 orientações da mão [3] (para a esquerda, para a direita, para frente, para o corpo, para baixo ou para cima); e 24 expressões não manuais em 4 posições distintas [8] (por exemplo, inflar bochechas, franzir o nariz, inclinar a cabeça para trás, balancear alternadamente os ombros etc). Detalhes mais aprofundados sobre a fonologia em Libras podem ser encontradas em [3], [6], [7], [8], [9].

III. Trabalhos relacionados

Um estudo sobre a conversão automática das 79 configurações padrão de mão utilizadas em Libras para escrita de língua de sinais (ELS) é realizado em [10]. O intuito da utilização da ELS é devido a maior naturalidade para as crianças surdas associarem a escrita ao movimento das mãos, similarmente ao que ocorre em crianças ouvintes e a associação grafema-fonema durante o aprendizado. Para a realização do projeto, foram gravadas as 79 configurações de mão de 5 voluntários diferentes utilizando o Microsoft Kinect com o intuito de capturar as imagens de mão juntamente com informação de profundidade. As imagens foram representadas em escala de cinza usando 54×54 *pixels*. Uma rede neural convolucional (CNN) foi treinada utilizando 316 imagens e testada nas 79 imagens restantes, propiciando uma acurácia de 87,5%. Cada imagem identificada é associada a um conjunto de símbolos conforme o protocolo *Formal Sign Writing* (FSW).

Já os trabalhos [11] e [12] consistem no reconhecimento automático de sinais de Libras utilizando descritores de forma e Rede Neurais Artificiais (RNA) do tipo *Multi Layer Perceptron* (MLP) como classificadores. O trabalho foi desenvolvido com a criação de uma base de dados que consiste em 9.600 imagens de 40 sinais em Libras, sendo algumas letras, números e palavras, evitando sinais que pudessem provocar confusão na desambiguação, e com 120 repetições de cada gesto realizados por até 5 especialistas em Libras. O sistema de reconhecimento projetado se baseia na utilização de Histograma de Gradientes Orientados (HOG) e Momentos Invariantes de Zernik (MIZ), além de um sistema de pré-processamento de detecção de pele também utilizando MLP. O treinamento foi realizado com 3.600 imagens, já o teste e validação utilizaram 800 e 400 imagens respectivamente. Todas as etapas foram feitas utilizando validação cruzada de 6 *folds* gerando uma acurácia média final de 96,77%. A base de dados gerada por esse trabalho está melhor descrita na Seção IV e é a mesma utilizada em nossos experimentos com aprendizado profundo.

O trabalho descrito em [13] utilizou a mesma base de dados gerada por [12]. No entanto, propôs o uso de uma combinação de HOG com SVM para fazer o reconhecimento dos gestos. O artigo reportou uma acurácia de 95,16%.

Um sistema para reconhecimento totalmente automatizado de 61 configurações de mão em Libras utilizando um classificador de novidade proposto previamente pelos autores onde as técnicas de extração de atributos podem ser a *two-dimensional* LDA ou a *two-dimensional* PCA é apresentado em [14]. Os autores utilizam o Microsoft Kinect para capturar as imagens produzindo uma nova base de dados de 12.200 imagens com 200 repetições de cada uma das 61 configurações de mãos realizadas por 10 pessoas diferentes. O conjunto de treinamento foi de 5.246 imagens e 6.954 imagens para teste, resultando, no melhor caso, em uma acurácia média de 96,31% usando KNN de $k = 1$ e distância Euclidiana.

A tese de mestrado em [15] aborda a utilização de redes neurais convolucionais (CNN) na classificação de 61 configurações de mãos de LIBRAS. O trabalho utiliza a mesma base de dados aplicada em [14] e testa 3 variações de arquitetura: 1) Alexnet [16]; 2) Variação da arquitetura apresentada em [17]; e 3) Variação da arquitetura apresentada em [18], combinadas com as técnicas de regularização. Desta forma são gerados 12 modelos, cada uma das 3 arquiteturas sem regularização, com regularização *dropout*, com regularização L2 e com *dropout*+L2. Os testes incluem uma vasta busca no intuito de otimizar os hiperparâmetros das redes e o treinamento utilizou um número fixo de 500 épocas. O melhor desempenho foi 97,98% de acurácia obtido com arquitetura da Alexnet usando *dropout*+L2.

Um sistema de reconhecimento de sinais em Libras baseado em modelos 3D e projeções 2D da mão é desenvolvido no artigo [19]. O trabalho é baseado em 61 configurações de mão realizadas por 5 pessoas com duas repetições. Os vídeos gravados foram manualmente segmentados para se extrair um quadro com vista frontal e um com vista lateral da mão que serviram de base para a construção dos modelos 3D das mãos usando os métodos *shape from sillhouette* e rotação, translação e invariância a escala de harmônicos esféricos. Já os atributos baseados nas projeções 2D vertical e horizontal foram obtidos através dos quadros manualmente segmentados. O classificador utilizado foi o SVM (*Support Vector Machine*). Quando baseado nos modelos 3D de mãos, obteve um desempenho médio de 96,83%. Em contrapartida, o SVM baseado nas projeções 2D alcançou 98,36% de acurácia. Os 2 classificadores utilizaram os dados dos 610 vídeos gravados e realizaram o treinamento e teste considerando a proporção de 70% e 30%, respectivamente, da base de dados.

O artigo em [20] aborda o reconhecimento dos dígitos de 0 a 9 em Libras e para tal, foi utilizada uma CNN com arquitetura LeNet [21] e base de dados de sinais estáticos sendo 2.640 imagens para treino e 1.360 para teste atingindo 98,57% de acurácia. Uma nova base de dados foi criada por uma única pessoa, que não participou nem do treinamento e nem do teste anteriormente realizado, gerando 1.000 imagens que obtiveram 82,5% de acertos.

IV. Bases de dados disponíveis

Essa seção contém a descrição das bases de dados de sinais estáticos disponíveis atualmente.

A. *Sinais estáticos de Libras (Bastos et al., 2015)*

Para tarefas de reconhecimento de gestos estáticos, pode se utilizar a base de dados do Laboratório de Pesquisa em Sistemas Inteligentes e Cognitivos (LASIC) da Universidade Estadual de Feira de Santana (UEFS) [12]. Essa base de dados contém 40 categorias, distribuídas igualmente em 9.600 imagens. As categorias são as letras do alfabeto: A, B, C, D, E, F, G, I, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y; números: 1, 2, 4, 5, 7 e 9; e as palavras: adulto, américa, avião, casa, gasolina, identidade, junto, lei, palavra, pedra, pequeno, e verbo. Cada categoria possui 240 imagens das mãos com resolução 50×50, sendo que 120 são em níveis de cinza e 120 são máscaras binárias das mãos. As principais desvantagens dessa base de dados são a falta de variabilidade da iluminação e *background*, bem como a ausência do restante do corpo (principalmente tronco, braços e face, que são muito importantes em Libras). Exemplos dessa base de dados podem ser vistos na Figura 1.

Fig. 1: Exemplos das imagens em nível de cinza da base (Bastos et al., 2015).

B. *Sinais estáticos de Libras (Costa et al., 2017)*

Uma base de dados de 61 configurações de mão pode ser encontrada em [14]. A base foi executada por 10 pessoas diferentes com 200 repetições cada, totalizando 12.200 imagens. As imagens foram adquiridas por meio do Microsoft Kinect. A base de dados disponibilizada já está separada nos conjuntos de treinamento e teste, com 6.100 imagens cada. As imagens possuem resolução de 139×135 *pixels* e já estão pré-processadas. Cada uma delas contém o mapa de profundidade de uma configuração de mão, já com uma máscara aplicada. Exemplos estão ilustrados na Figura 2.

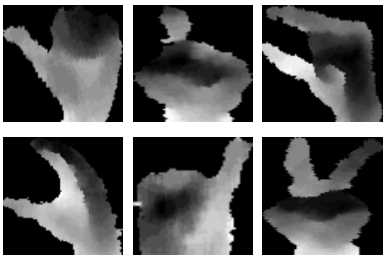

Fig. 2: Exemplos dos mapas de profundidade da base de dados em (Costa et al., 2017).

V. Aprendizado profundo em sinais estáticos da Libras

É sabido que ainda não há uma base de dados de Libras de vídeo que seja referência para o uso em aplicações em visão computacional. De um modo geral, os trabalhos na área geralmente envolvem a criação de uma pequena base de dados para validação dos métodos, o que dificulta uma comparação objetiva entre eles. Por outro lado, esse tipo de comparação é mais fácil nas bases contendo sinais estáticos. Até onde os autores desse artigo conseguiram apurar, é possível encontrar alguns trabalhos que utilizam a base de dados gerada por [12] (descrita na Seção IV).

Nesse trabalho, nós avaliamos o desempenho de quatro redes amplamente encontradas na literatura nesta base de dados (Bastos et. al, 2015) [12]. Foram selecionadas redes bem conhecidas e muito utilizadas em tarefas de visão computacional: LeNet [22], VGG16 [23], VGG19 [23], InceptionV3 [24] e ResNet50 [25]. Em todas as redes, a função de ativação ReLu foi utilizada, com exceção da camada de saída, que utilizou a função Softmax. A LeNet foi construída com as camadas convolucionais sendo $32\times5\times5$, $64\times3\times3$, $64\times3\times3$ e $64\times3\times3$ e como classificador duas camadas densas de 128 neurônios. Já a VGG16, VGG19 e ResNet50, utilizaram um classificador MLP com três camadas densas de 256 neurônios. A Inception não utiliza MLP para fazer a classificação, e sim a camada de Global Average Pooling. Além disso, a função custo utilizada em todas as redes foi a entropia cruzada categórica esparsa.

A. *Resultados*

Na base de dados em (Bastos et. al, 2015) [12], todos os algoritmos foram testados usando validação cruzada por *k-fold* com 10 *folds*. O gráfico da Figura 3 contém a evolução média da função custo das redes utilizadas ao longo das épocas. É possível notar que a convergência das redes VGG16, VGG19 é mais rápida que a das redes LeNet, InceptionV3 e ResNet50.

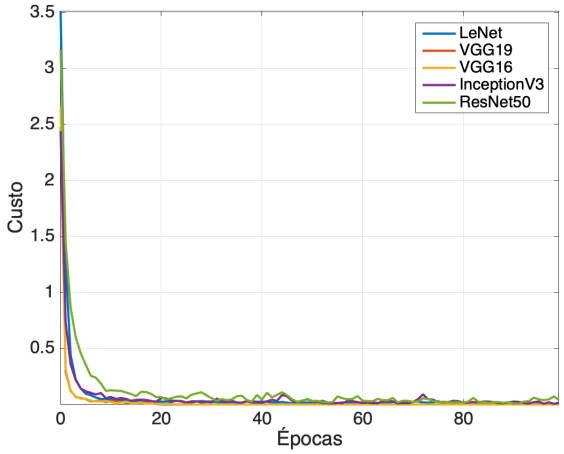

Fig. 3: Evolução da função custo ao longo das épocas no conjunto de treinamento da base de dados (Bastos et. al, 2015) [12]. As curvas contém os valores médios da função custo nos 10 *folds*.

As acurácias de teste (base de dados (Bastos et. al, 2015) [12]) em todos os *folds* estão na Tabela II e a Figura 4 sintetiza os resultados da tabela, que podem ser considerados todos satisfatórios. As redes LeNet, VGG16 e VGG19 apresentaram resultados muito parecidos, com alta taxa média de acerto e baixo desvio padrão. Essas redes também possuem

uma implementação mais simples e, portanto, menor quantidade de parâmetros. Isso implica a necessidade de menos sinais e convergência mais rápida, o que concorda com as curvas na Figura 3. As Redes InceptionV3 e ResNet50, apesar de possuírem uma mediana alta, apresentaram baixa acurácia média e desvio padrão bem mais elevado. Por outro lado, essas duas últimas possuem maior poder de abstração e são capazes de aprender padrões mais complexos.

	LeNet	**VGG16**	**VGG19**	**InceptionV3**	**ResNet50**
folds	0,9958	0,9854	0,9937	0,9917	0,8896
	0,9958	0,9958	0,9979	1,0000	0,9979
	0,9750	0,9896	0,9896	0,9979	0,9833
	0,9958	0,9917	0,9937	0,4083	0,9396
	0,9937	0,9937	1,0000	0,9312	0,9979
	1,0000	1,0000	1,0000	0,5938	1,0000
	0,9979	0,9937	0,9958	1,0000	1,0000
	0,9937	0,9917	0,9958	0,9979	0,6479
	0,9979	1,0000	0,9958	0,9541	0,9979
	0,9979	0,9958	0,9917	0,6417	0,3625
Média	0,9944	0,9937	0,9954	0,8517	0,8817
STD	0,0067	0,0043	0,0032	0,2075	0,2016
Mediana	0,9958	0,9937	0,9956	0,9427	0,9615

TABELA II: Resultados das redes nos conjuntos de teste da base de dados (Bastos et. al, 2015) [12] de todos os *folds*.

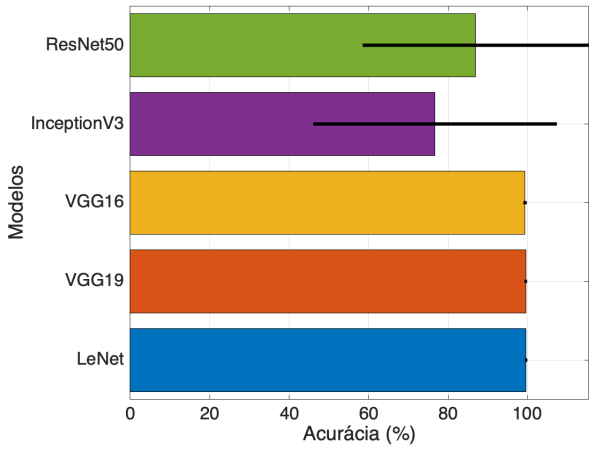

Fig. 4: Desempenho das redes utilizadas no conjunto de testes da base de dados (Bastos et. al, 2015) [12]. As barras coloridas representam as acurácias médias (nos 10 *folds*) e as linhas horizontais são os desvios padrão.

Com os resultados mostrados, é possível afirmar que a rede VGG19 teve um resultado bastante satisfatório na base de dados em questão. A Tabela III contém uma comparação do resultado na VGG19 (pior dos *folds*) com os encontrados na literatura. Apesar de todos os resultados serem satisfatórios, temos evidências de que redes profundas tendem a apresentar os melhores resultados nesse caso.

Método	**Acurácia**
(Bastos et. al, 2015) [12]	0,9677
(Pessoa et. al, 2016) [13]	0,9516
VGG19	0,9917

TABELA III: Comparação entre resultados na base de dados (Bastos et. al, 2015) [12].

As redes escolhidas também foram avaliadas na base de dados em (Costa et. al, 2017) [14]. Nesse caso, a validação cruzada *k-fold* não foi utilizada para facilitar a comparação, pois a base de dados já é separada em conjuntos de treino e teste. O gráfico da Figura 5 contém a evolução média da função custo das redes utilizadas ao longo das épocas. Assim como no caso anterior, a convergência das redes VGG16, VGG19 é mais rápida que a das demais. Os resultados no conjunto de teste foram: LeNet - 93,49% (27 épocas); VGG16 - 91,59% (27 épocas); VGG19 - 91,57% (30 épocas); InceptionV3 - 99,13% (49 épocas); ResNet50 - 93,44% (40 épocas). Uma versão gráfica desses resultados está na Figura 6.

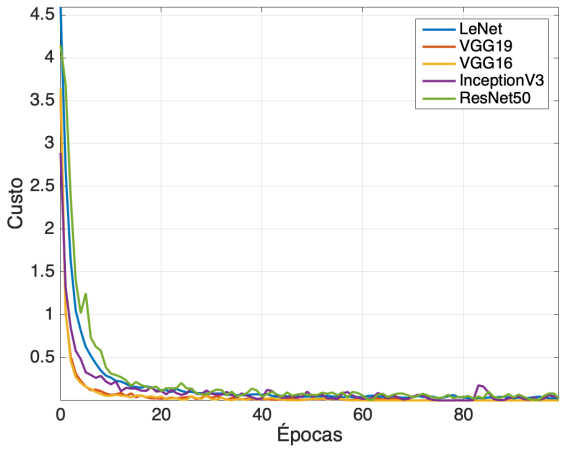

Fig. 5: Evolução da função custo ao longo das épocas no conjunto de treinamento da base de dados (Costa et. al, 2017) [14].

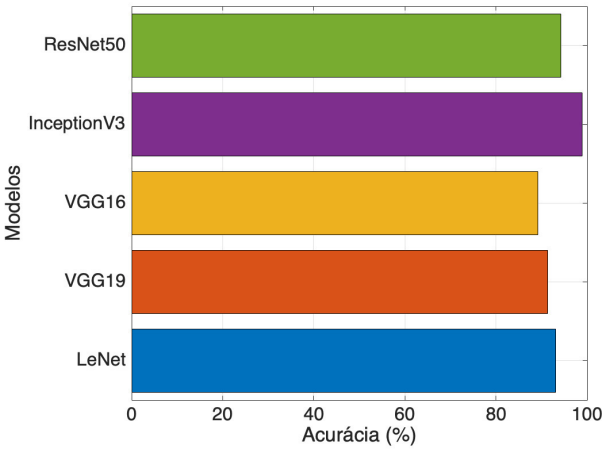

Fig. 6: Desempenho das redes utilizadas no conjunto de testes da base de dados (Costa et. al, 2017) [14]. As barras coloridas representam as acurácias no conjunto de teste.

Na base de dados (Costa et. al, 2017) [14], a rede InceptionV3 apresentou os melhores resultados. Uma comparação entre os resultados obtidos e os encontrados na literatura podem ser vistos na Tabela IV. Novamente, todos os resultados podem considerados satisfatórios e o método em (Oliveira, 2019) [15] já utiliza uma abordagem neural. Ainda assim, a rede InceptionV3 obteve a maior acurácia.

A partir dos resultados expostos nessa seção, é possível afirmar que as arquiteturas de aprendizado profundo mais co-

Método	Acurácia
(Costa et. al, 2017) [14]	0,9631
(Oliveira, 2019) [15]	0,9798
InceptionV3	0,9913

TABELA IV: Comparação entre resultados na base de dados (Costa et. al, 2017) [14].

nhecidas atualmente possuem um bom desempenho na classificação de sinais estáticos em Libras. Vale ressaltar que as bases de dados estáticas, até onde foi possível verificar, somente contêm as mãos executando os sinais e/ou configurações de mão. Esses resultados são apenas promissores e podem ter serventia quando combinados com outras estratégias (reconhecimento de expressões faciais, pose dos membros, classificação de movimento, processamento de linguagem natural, etc) em um sistema de transcrição Libras-Português.

VI. Conclusões

Nesse trabalho, foi discutido o problema de reconhecimento de sinais estáticos em Libras. Uma ampla revisão de literatura contemplando trabalhos recentes foi feita. As arquiteturas de aprendizado profundo LeNet, VGG, InceptionV3 e ResNet foram avaliadas em duas bases de dados consolidadas. Nos nossos experimentos, a rede VGG19 obteve uma acurácia de 99,17% na base de dados (Bastos et. al, 2015) [12] e a rede InceptionV3 obteve 99,13% na base de dados (Costa et. al, 2017) [14].

Conforme detalhado neste texto, a Libras possui vários detalhes importantes, como expressão facial, posicionamento e orientação das mãos, movimento e configuração das mãos. Um sistema genérico de transcrição Libras-Português usando visão computacional deve ser capaz de modelar adequadamente todas essas nuances. Entretanto, o correto reconhecimento da configuração das mãos é fundamental e é neste cenário que este trabalho se encaixa.

A partir dos resultados expostos nessa seção, é possível afirmar que as arquiteturas de aprendizado profundo mais conhecidas atualmente possuem um bom desempenho na classificação de sinais estáticos em Libras, com acurácia maior do que arquiteturas projetadas exclusivamente para esta tarefa.

Os resultados satisfatórios dos experimentos conduzidos nesse trabalho sugerem que um importante trabalho futuro seria atacar os outros problemas relacionados com a Libras (pose, movimento, etc). Como foi explicitado, existem alguns trabalhos que utilizam visão computacional em sinais dinâmicos (vídeos), mas geralmente cada um deles propõe uma pequena base de dados, o que torna a comparação dos resultados mais difícil, sobretudo utilizando aprendizado profundo. Propor uma base de dados de vídeos que seja escalável e genérica o suficiente para uso em diversos trabalhos envolvendo visão computacional em Libras também está entre os trabalhos futuros.

Referências

[1] Instituto Brasileiro de Geografia e Estatística (IBGE), "Censo2010," https://censo2010.ibge.gov.br/, 2010.

[2] W. Stokoe, "Sign language structure," *Annual Review of Anthropology*, vol. 9, no. 1, pp. 365–390, Oct. 1980.

[3] R. M. de Quadros and L. B. Karnopp, *Língua de Sinais Brasileira: Estudos Lingüísticos*. Artmed Editora, 2009.

[4] Brasil, "Lei no 10.436. Brasília, Presidência da República, Casa Civil, Subchefia para Assuntos Jurídicos, 24 de abril de 2002. Dispõe sobre a Língua Brasileira de Sinais - Libras e dá outras providências." 2002.

[5] G. Z. de Castro, R. R. Guerra, M. M. de Assis, T. M. Rezende, G. T. B. de Almeida, S. Almeida, C. L. de Castro, and F. G. G. aes, "Desenvolvimento de uma Base de Dados de Sinais de Libras para Aprendizado de Máquina: Estudo de Caso com CNN 3D," in *Anais do XIV Simpósio Brasileiro Automação Inteligente (SBAI)*, Ouro Preto, Out. 2019.

[6] F. F. Fonseca, "Visão computacional aplicada ao reconhecimento de imagens relacionadas à Língua Brasileira de Sinais," Monografia de conclusão de curso, Universidade Federal do Rio de Janeiro - Escola politécnica, Rio de Janeiro, RJ, Brasil, 2020.

[7] L. F. Brito, *Por uma gramática de língua de sinais*. Editora Tempo Brasileiro, 1995.

[8] R. Passos, "Parâmetros físicos do movimento em libras: um estudo sobre intensificadores," Tese de doutorado, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil, 2014.

[9] W. Stokoe, "Uma abordagem fonológica dos sinais da LSCB," *Revista Espaço - Informativo Técnico-Científico do INES*, vol. 1, no. 1, pp. 20–43, Jan. 1990.

[10] J. D. B. Junior, "Tradução automática de línguas de sinais: do sinal para a escrita," Monografia de conclusão de curso, Universidade Federal do Pampa - Ciência da Computação, Alegrete, RS, Brasil, 2016.

[11] I. L. O. Bastos, "Reconhecimento de sinais da libras utilizando descritores de forma e redes neurais artificiais," Tese de mestrado, Universidade Estadual de Feira de Santana - Ciência da Computação, Salvador, BA, Brasil, 2015.

[12] I. L. O. Bastos, M. Angelo, and A. Loula, "Recognition of static gestures applied to brazilian sign language (Libras)," in *Proc. of the 28th SIBGRAPI Conf. on Graphics, Patterns and Images*, Salvador, Aug. 2015.

[13] A. C. P. Pessoa, G. Braz, L. B. Maia, R. M. P. Pereira, and T. A. Silva, "Reconhecimento de gestos manuais para identificação de letras do alfabeto da lıngua brasileira de sinais (Libras)," Universidade Federal do Maranhão, Relatório Técnico, 2016.

[14] C. F. F. C. Filho, R. S. de Souza, J. R. dos Santos, B. L. dos Santos, and M. G. F. Costa, "A fully automatic method for recognizing hand configurations of Brazilian sign language," *Research on Biomedical Engineering*, vol. 33, no. 1, pp. 78–89, March 2017.

[15] A. S. Oliveira, "Uso de técnicas de aprendizagem profunda na classificação de configurações de mão de língua de sinais," Tese de mestrado, Universidade Federal do Amazonas - Programa de Pós-graduação em Engenharia Elétrica, Manaus, AM, Brasil, 2019.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, Dec. 2017.

[17] X. Zhu, W. Liu, X. Jia, and K. K. Wong, "A two-stage detector for hand detection in ego-centric videos," in *Proc. of the IEEE Winter Conf. on Applications of Comp. Vision (WACV)*, Lake Placid, March 2016.

[18] L. Pigou, S. Dieleman, P. J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Comp. Vision - ECCV 2014 Workshops - 13th European Conf.*, Sept. 2015.

[19] A. J. Porfirio, K. L. Wiggers, L. E. S. Oliveira, and D. Weingaertner, "Libras sign language hand configuration recognition based on 3d meshes," in *Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics*, Manchester, Oct. 2013.

[20] A. V. Santos, I. F. Bacurau, J. M. Silva, T. B. Viana, and R. G. F. Feitosa, "Rede neural artificial convolucional aplicada ao reconhecimento de configuração de mão nos símbolos de 0 a 9 da língua brasileira de sinais (LIBRAS)," in *Anais Estendidos do XV Simpósio Brasileiro de Sistemas de Informação*, Aracaju, Maio 2019.

[21] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Müller, E. Säckinger, P. Simard, and V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," in *Proc. of the Int. Conf. on Artificial Neural Networks*, Paris, Oct. 1995.

[22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of the Int. Conf. on Learning Representations*, San Diego, May 2015.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. of the IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR)*, Las Vegas, June 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR)*, Las Vegas, June 2016.