Modelo de Rede Neural Convolucional para Classificação da Linguagem de Sinais

E Lima, E Veloso, G Júnior, M Pinheiro e R Rego

Resumo **- Este projeto utiliza técnicas de aprendizado de máquina, em particular, redes neurais convolucionais (CNNs), para processar imagens de gestos em linguagens de sinais e identificar os símbolos correspondentes. O modelo apresentou 99,82% de acurácia. Os resultados indicam um grande potencial para a aplicação prática deste projeto.**

Palavras chaves **– Libras, ASL, CNN, aprendizagem de máquina.**

I. INTRODUÇÃO

A Língua Brasileira de Sinais (Libras) é uma língua visual-espacial utilizada pela comunidade surda no Brasil. Ela é reconhecida oficialmente como meio de comunicação e expressão pela Lei nº 10.436/2002 e pelo Decreto nº 5.626/2005 [1][2]. Diferentemente do português, Libras possui gramática própria, baseada em gestos, expressões faciais e movimentos corporais, sendo uma língua naturalmente adaptada à comunicação visual. É uma ferramenta essencial para a inclusão e participação das pessoas surdas na sociedade [3].

Libras não é uma linguagem universal, ou seja, cada país possui sua própria língua de sinais. No caso do Brasil, a Libras é a principal forma de comunicação para a comunidade surda, sendo fundamental para a educação, trabalho e interação social.

É importante compreender que cada país possui a sua própria língua de sinais, adaptada à sua cultura e comunidade surda. Por exemplo, nos Estados Unidos, a Língua de Sinais Americana (ASL) é predominante, enquanto na França, utiliza-se a Língua de Sinais Francesa

(LSF) [3][4]. Essas línguas possuem gramáticas e expressões próprias, o que enfatiza a diversidade linguística que envolve a comunicação com a comunidade surda.

A importância da Libras vai além da esfera individual, influenciando diretamente políticas públicas, direitos civis e a promoção da inclusão. É crucial para educadores, profissionais de saúde e todos os que desejam estabelecer uma comunicação eficaz com a comunidade surda [4].

No entanto, apesar de ser essa importante ferramenta de comunicação e inclusão de pessoas surdas, no Brasil, Libras ainda é uma língua pouco difundida, onde têm-se uma a ausência de políticas públicas eficazes para promover o ensino e uso da língua.

Pensando em formas de suprir essa deficiência no conhecimento de libras, o presente trabalho tenta apresentar uma ferramenta de identificação de símbolos em linguagem de sinais utilizando aprendizagem de máquina.

II. LEVANTAMENTO BIBLIOGRÁFICO

Porfirio et al. (2013) utilizaram vídeos gerados pelo kinect e aplicaram técnicas de malhas 3D para reconhecer as configurações das mãos feitas na LIBRAS. Uma configuração de mão é uma forma apresentada pelas mãos durante a execução de um sinal, e segundo os autores, a LIBRAS possui 61 possíveis configurações de mãos. Os quadros dos vídeos existentes tiveram que ser identificados manualmente para que as posições fossem relacionadas e as imagens exportadas no formato

jpeg, dado que, segundo os autores, os vídeos adquiridos pelo kinect associados à metodologia não foram capazes de produzir os detalhes necessários para a geração da malha, e por isso passaram a utilizar imagens das posições frontal e lateral da mão tratadas manualmente. Os autores indicam resultados com taxa de acerto de 96%.

No trabalho descrito em Pizzolato, Anjo e Pedroso (2010), a estratégia adotada envolveu a extração automática de características através de uma rede neural artificial, a fim de agrupar imagens de gestos da LIBRAS que apresentassem posturas semelhantes. Posteriormente, a saída desta rede foi empregada como entrada para outra rede neural artificial, com o propósito de determinar a correspondência do sinal com um símbolo específico do alfabeto. Os resultados mais promissores foram alcançados ao empregar uma rede neural do tipo Multi-Layer Perceptron (MLP) com uma camada escondida composta por 300 neurônios, juntamente com a função de ativação tangente hiperbólica. Além disso, foi aplicada uma combinação dessa saída com Modelos Ocultos de Markov (HMM) para detectar as transições entre letras em cada palavra, bem como para reconhecer situações em que múltiplas palavras fossem representadas por transições de gestos com significados distintos. A taxa de acerto descrita pelos autores foi de 98%.

Com base nessa pesquisa, Anjo, Pizzolato e Feuerstack (2012) exploraram diferentes combinações de técnicas ao desenvolver um sistema conhecido como "GestureUI - Interface de Usuário por Gestos", elaborado para identificar gestos estáticos em vídeos em tempo real, utilizando informações de profundidade adquiridas por meio do dispositivo Kinect.O classificador resultante dessa rede neural é composto por uma camada de entrada composta por 625 neurônios, acompanhada por uma camada escondida com 100 neurônios, e a saída apresenta 5 opções de classificação possíveis. Foi obtida uma taxa de precisão de 100% apenas para os gestos estáticos utilizados no treinamento e reconhecimento, os quais incluíram as vogais (A, E, I, O e U) e as consoantes (B, C, L, F e V). Nenhum gesto dinâmico foi incluído no conjunto de dados.

No estudo realizado por Bastos, Angelo e Loula (2015), foi criado um conjunto de dados que abrange 40 sinais da LIBRAS, englobando a maioria das letras do alfabeto, excluindo aquelas cujos sinais envolvem algum tipo de movimento, alguns números e 12 palavras. Empregou-se uma estratégia de classificação em dois estágios, o primeiro destinado a identificar a categoria à qual a imagem do sinal pertence e a encaminhar o processo de reconhecimento para a rede neural subsequente. Esse estágio direciona qual rede neural deve ser ativada com base na categoria identificada. A taxa de reconhecimento média foi de 96,77%.

Donahue et al. (2014) apresentaram um sistema de categorização de vídeos que introduziram como LRCN (Long-term Recorrent Convolutional Network), uma rede convolucional recorrente de longo prazo. Neste sistema, camadas convolucionais processam individualmente cada quadro de entrada do vídeo e compartilham seus resultados com uma camada LSTM, uma rede de memória de curto e longo prazo.

No estudo conduzido por Passos et al. (2021), foi empregada a técnica conhecida como Gait Energy Image (GEI) para a identificação de gestos. Essa técnica foi utilizada para codificar informações relacionadas aos movimentos das mãos, braços e cabeça. Abordaram a identificação de sinais isolados por meio de algoritmos tradicionais de aprendizado de máquina, obtendo resultados comparáveis às arquiteturas de redes neurais mais avançadas.

Mittal et al. (2019) introduzem uma versão adaptada da Long Short Term Memory (LSTM) com o propósito de reconhecer sequências ininterruptas de sinais sem relação aparente ou sequências de sinais contínuos interligados. As amostras de sinais coletados foram segmentadas em subunidades para serem compatíveis com

modelos de redes neurais. Em outras palavras, cada sequência de sinais foi subdividida em várias unidades, permitindo que cada sinal fosse processado pelo modelo como uma entidade isolada.

O trabalho Recognizing hand gestures with Microsoft Kinect, tem como objetivo estudar a viabilidade do reconhecimento de gestos em uma escala pequena, como por exemplo agarrar e apontar. É identificado gestos simples, que uma pessoa pode realizar com sua mão, ao invés de tentar reconhecer ações de corpo inteiro, como acenar e pular. Segundo Tang (2011), são utilizados os recursos da câmera RGB e o sensor de profundidade do Kinect. Como também, foi utilizado um rastreador de corpo inteiro, que identifica a localização da mão de uma pessoa, a qual deve estar posicionada na extremidade de um braço na estrutura esquelética. A segunda etapa consiste em reconhecer os pixels que constituem a área referente à mão. Na última etapa são identificados os movimentos em uma sequência de imagens e posições pré-definidas em um conjunto de treinamento para obter-se o gesto atribuído. Como resultado foi obtida uma taxa de sucesso de 90% para reconhecimento de gestos de agarrar e soltar.

O aplicativo ProDeaf é um dicionário e tradutor de libras desenvolvido pela ProDeaf Tecnologias Assistivas em parceria com a Bradesco Seguros (Brandt, 2015). Ele possibilita traduzir palavras para gestos utilizando computadores e smartphones. É permitido digitar a palavra, procurar no dicionário dentro do aplicativo ou utilizar o reconhecimento de voz. Para utilizar este recurso, é necessário ter conexão com a internet. Após escolher a palavra a ser traduzida, o aplicativo retorna para a tela inicial e o Avatar realiza os gestos, indicando a palavra logo acima. Caso a palavra não esteja disponível no dicionário,a mesma é soletrada utilizando o alfabeto de libras.

André Henrique Brandt em 2015, desenvolveu a ferramenta que intitula-se LiRANN - *Libras Recognition based on Artificial Neural Networks with Kinect* (Sistema de reconhecimento de libras baseado em redes neurais artificiais com kinect) e tem como objetivo, desenvolver uma aplicação que interprete e traduza o alfabeto da língua brasileira de sinais em tempo real, utilizando o Kinect e Redes Neurais Artificiais. De modo geral, a aplicação desenvolvida pode ser dividida em dois módulos. O primeiro, utiliza a câmera RGB do Kinect, em que são feitas capturas dos padrões a serem reconhecidos. O segundo módulo, corresponde a uma rede neural artificial para fazer o reconhecimento de posições baseado nas informações capturadas.

Giulia Zanon de Castro em 2020, desenvolveu o trabalho que se intitula como Reconhecimento de Línguas de Sinais Utilizando Redes Neurais Convolucionais e Transferência de Aprendizado e tem como objetivo, a criação de um modelo de reconhecimento automático de sinais em Libras, utilizando as redes neurais convolucionais tridimensionais (CNN 3D). Se trata de uma geração de um modelo de aprendizado de máquina para a tarefa de reconhecimento de sinais em Libras, empregando uma metodologia independente de sensores de profundidade e de luvas instrumentadas.

Jhon et al. propuseram uma solução baseada em redes neurais profundas para o reconhecimento e tradução de sinais de libras para a língua portuguesa escrita. Trata-se de um estudo exploratório em que três sinais em Libras são submetidos ao treinamento, reconhecimento e classificação em duas arquiteturas de redes neurais (LSTM e BiLSTM). Dado isto, os autores construíram modelos de aprendizagem de máquina para categorização de imagens em modo contínuo e em tempo real e teve como primeiro passo, a construção de um sistema de tradução de Libras para a língua portuguesa escrita. Foram aplicadas técnicas de visão computacional e aprendizado profundo com redes neurais artificiais, visando realizar o reconhecimento de um conjunto de sinais básicos em Libras.

Johann Felipe Voigt em 2018, realizou um trabalho de aprendizagem profunda para reconhecimento de gestos da mão usando imagens e esqueletos com aplicações em libras, com a finalidade de experimentar a utilização de redes neurais juntamente com o uso dos esqueletos, para diferenciar diferentes gestos e movimentos na interação entre homem e computadores. Neste trabalho, os autores realizaram a avaliação dos 26 sinais da Linguagem Brasileira de Sinais (Libras), em que os resultados poderão ser ampliados para qualquer gesto ou movimento realizado apenas com uma mão. Através do dispositivo Leap Motion, foram capturadas tanto imagens quanto dados do esqueleto da mão e foram avaliadas diversas arquiteturas de Redes Neurais para reconhecer gestos, com ênfase em sinais de Libras.

Marcelo Chamy Machado em 2018 desenvolveu um trabalho para classificação automática de sinais visuais da língua brasileira de sinais e teve como objetivo, explorar a aplicação de características espaço-temporais para realizar a classificação de vídeos de sinais da LIBRAS. Nesse cenário, o autor tende-se a demonstrar que o uso dessas abordagens pode ser de muita utilidade para o reconhecimento de linguagens de sinais, com a utilização de redes neurais profundas para capturar características espaço-temporais dos vídeos.

III. ARQUITETURA DA REDE

Para o desenvolvimento da pesquisa, foi utilizado um modelo de Rede Neural Convolucional (CNN), cujo principal propósito é o reconhecimento de imagens. No projeto, foram utilizadas as seguintes bibliotecas: NumPy, pandas, TensorFlow, Matplotlib e Scikit-Learn. A base de dados utilizada é a Linguagem de Sinais Americana, com exceção das letras "j" e "z", que envolvem movimentos.

Após a importação dos dados, o primeiro passo é a extração das labels nos dados de treinamento, que servirão como rótulos para a validação dos dados. Em seguida, é realizado o procedimento de "reshape" para padronizar os dados, tanto de teste quanto de treinamento. A função "LabelBinarizer" é utilizada para transformar as labels, de teste e treino, em uma matriz binária.

A biblioteca "ImageDataGenerator" é empregada para realizar a manipulação e aumento de dados de imagem, sendo útil para melhorar o desempenho de modelos de aprendizado profundo, tornando-os mais robustos e capazes de generalizar melhor. A etapa de normalização também é essencial, uma vez que os valores dos pixels variam de 0 a 255, representando a intensidade de cor de cada pixel. O processo de normalização consiste em dividir os valores por 255, para que todos os valores fiquem no intervalo de 0 a 1.

Na rede neural, foi aplicada a técnica de parada antecipada conhecida como EarlyStopping, que tem a finalidade de prevenir o sobreajuste (overfitting) do modelo. Se a perda no conjunto de validação não estiver apresentando melhorias, isso indica que o modelo pode ter atingido seu desempenho máximo no conjunto de validação e está começando a se ajustar em excesso aos dados de treinamento. Portanto, interromper o treinamento nesse ponto pode auxiliar na prevenção do sobreajuste do modelo, resultando em economia de tempo e recursos computacionais.

A arquitetura de rede do modelo é composta por 4 camadas, sendo a última conhecida como camada totalmente conectada. As camadas de processamento, que são as 3 primeiras, são compostas pelas camadas convolucionais e de pooling. A camada convolucional é responsável pelas operações matriciais que são realizadas nos dados, que são os filtros aplicados a cada conjunto de dados em cada iteração. Já a camada de pooling é responsável por combinar unidades próximas, reduzindo o tamanho dos dados de entrada na próxima camada. Pode ser aplicado o pooling máximo ou médio. Quando ocorre o pooling máximo, é utilizado o valor máximo do

conjunto de dados para representá-lo, enquanto no pooling médio são usados os valores médios do conjunto de dados [23]. As 4 camadas possuem respectivamente 128, 128, 64 e 256 neurônios com a função de ativação "ReLU", como mostrado na figura 1. Na camada de saída, é aplicada a função de ativação softmax para realizar a classificação dos dados.

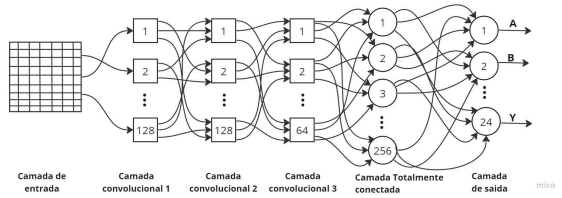

Fig. 1. Arquitetura do modelo. Fonte: Autoria própria.

Foi utilizada uma técnica de regularização na rede, conhecida como Dropout, que tem como função a redução do sobreajuste, que acontece quando o modelo se torna muito complexo, ajustando-se demais aos dados de treinamento, mas tendo dificuldade em generalizar para outros dados. Durante os períodos de treinamento, o Dropout irá desligar aleatoriamente um determinado número de neurônios, que, no caso do nosso modelo, foi de 15% nas 3 camadas convolucionais e 10% na camada totalmente conectada. A principal intenção do Dropout é evitar que a rede neural se torne muito dependente de neurônios específicos durante o treinamento, tornando a rede mais robusta e prevenindo o sobreajuste, já que em cada iteração o conjunto de neurônios utilizados é diferente.

O otimizador escolhido para o modelo foi o "Adam". Para o treinamento, foi definido um total de 100 épocas, utilizando a função EarlyStopping com uma paciência de 3, monitorando o valor de "val_loss" para garantir que o treinamento termine no momento em que o modelo deixar de apresentar bons resultados de aprendizagem.

Após o treinamento, o gráfico de evolução do treinamento é exibido, seguido pela matriz de confusão para analisar o desempenho do modelo no treinamento e validação dos dados.

IV. RESULTADOS DA REDE NEURAL

A entropia cruzada é uma métrica clássica amplamente empregada na resolução de problemas de classificação. É comumente utilizada em modelos para avaliar o quão bem eles estão realizando a tarefa de categorizar instâncias em diferentes classes ou grupos, considerando as distribuições probabilísticas associadas a cada classe. Quanto mais distintas e bem separadas essas classes forem, menor será o valor da entropia cruzada. A Fórmula 1 é empregada para o cálculo dessa métrica, na qual "y_j" representa os valores reais das classes e "p_j" representa as probabilidades previstas pelo modelo para essas classes [21].

$$L = -\sum_{j=1}^{m} y_j * log(p_j) \qquad (1)$$

Com o propósito de visualizar a evolução da função de perda durante o treinamento e a validação, foi criado na Figura 2, no qual é possível observar como a entropia cruzada diminui em relação ao número de épocas decorridas.

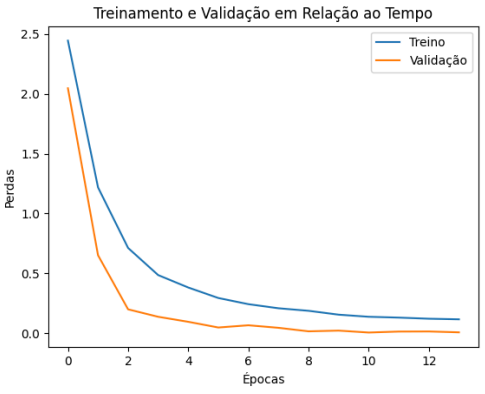

Fig. 2. Curva da função perda aplicada aos dados de treinamento e validação em relação a quantidade de épocas. Fonte: Autoria própria.

Para avaliar a eficácia do modelo de rede neural construído, foram empregadas distintas

métricas, incluindo a matriz de confusão e a acurácia.

A matriz de confusão é uma ferramenta representada em forma de tabela que auxilia na compreensão da frequência com que um classificador atribui instâncias a cada classe. No qual, cada linha representa instâncias reais de uma classe, enquanto cada coluna representa instâncias previstas para essa classe. A matriz de confusão ideal possui classificadores com contagens elevadas na diagonal principal, representando altos índices de verdadeiros positivos. Na Figura "3", pode-se observar a matriz de confusão gerada pela rede neural em questão.

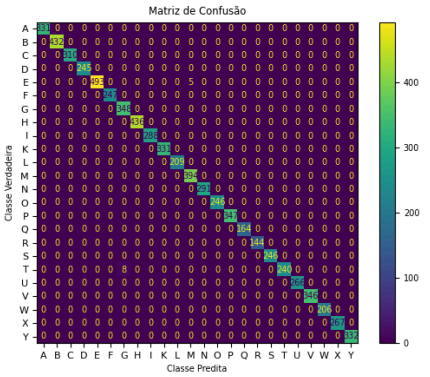

Fig. 3. Matriz de confusão do modelo. Fonte: Autoria própria.

A acurácia é uma ferramenta utilizada para calcular o percentual de predições corretas dessa matriz, ou seja, é uma medida direta de quão bem o modelo está fazendo previsões corretas em relação ao total de previsões, e para avaliar a validade de um modelo é fundamental analisar métricas adicionais e considerar as implicações de prever falsos positivos (f_p) e falsos negativos (f_n). A acurácia é calculada dividindo o número de previsões corretas pelo número total de previsões, como exposto na Fórmula 2, onde t_p e t_n representam respectivamente verdadeiros positivos e verdadeiros negativos [22].

$$acurácia = \frac{(t_p + t_n)}{(t_p + t_n + f_p + f_n)} \qquad (2)$$

O modelo de rede neural construído obteve uma acurácia de aproximadamente 99,82%, sendo possível afirmar que o modelo apresentou um ótimo ajuste com os dados repassados.

V. DIFICULDADES ENCONTRADAS

A construção da rede neural obteve como início a implementação do banco de dados constituído por imagens da língua brasileira de sinais (LIBRAS), porém pelo curto período de tempo, a equipe obteve dificuldades na tratativa dos dados, impossibilitando de realizar um treinamento ideal com a rede neural construída. Em um âmbito geral, com os dados iniciais, o modelo não conseguiu um bom aprendizado, resultando em um valor alto para o cálculo de erro, tendendo crescer, e consequentemente uma baixa acurácia de aproximadamente 9.00%.

A língua brasileira de sinais possui movimento em alguns de seus símbolos, mais especificamente quatro deles, sendo a letra “H”, ”J”, ”K” e ”Z”. Outros trabalhos na área, tendo por exemplo o de Lucas Lacerda [6], possuem uma base de dados de treino restritas às letras estáticas. Na base de dados LIBRAS, construída neste trabalho, se viu necessário realizar a conversão dos vídeos obtidos por uma sequência de imagens, para ser repassado corretamente para a rede neural construída.

Na construção da rede neural atual, se viu necessário a modificação de alguns parâmetros, dentre eles o número de camadas existentes e a quantidade de neurônios das mesmas, inicialmente, a rede apresentava três camadas, sendo elas constituídas por 16, 32 e 64 neurônios, sendo a última a camada totalmente conectada, como mostra a Figura 4. O número reduzido de camadas e neurônios interferia negativamente na taxa de aprendizado do modelo, sendo necessário o aumento de uma camada e dos neurônios presentes nela para 128,

128, 256 e 128 respectivamente, como apresentado na Figura 5.

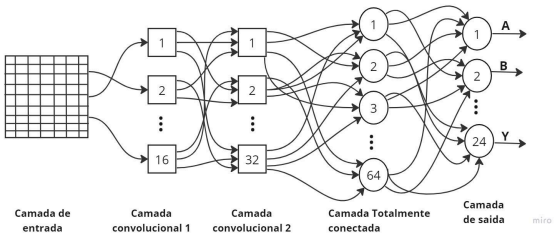

Fig. 4. Arquitetura do modelo anterior. Fonte: Autoria própria.

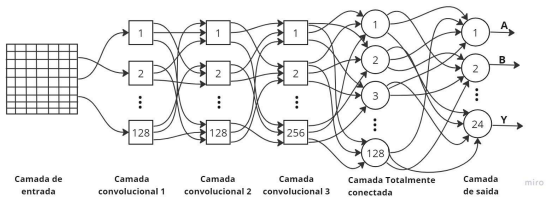

Fig. 5. Arquitetura do modelo anterior. Fonte: Autoria própria.

Por fim, por consequência da ineficiência da rede anterior, realizou-se a modificação da quantidade de neurônios, para respectivamente 128, 128, 64 e 256 formando as 4 camadas presentes no modelo.

VI. CONCLUSÃO

Neste trabalho, foi desenvolvido uma rede neural de classificação utilizando o modelo de rede neural convolucional. Inicialmente, foi realizada a busca por imagens da linguagem brasileira de sinais, e a construção de um banco de dados com fotografias coletadas pela própria equipe. Apesar do banco construído não ter sido utilizado no presente trabalho, existe a possibilidade de realizar a tratativa dos dados e ser aplicado em trabalhos futuros. O modelo foi construído e analisado através de métricas. O gráfico da função perda permite visualizar que a curva de treinamento e a de validação assemelham-se, finalizando o treinamento com o valor da entropia cruzada menor que 0.50, permitindo afirmar a ausência de erros como overfitting e underfitting. A matriz de confusão possibilita observar valores elevados na diagonal dos verdadeiros, aproximando-se do ideal. Para avaliar a validade de um modelo é fundamental analisar o valor da acurácia, no qual o modelo presente apresenta o valor de 99,8187%, ganhando destaque pelo alto valor obtido.

VII. REFERÊNCIAS

[1] Brasil. Lei nº 10.436, de 24 de abril de 2002. Dispõe sobre a Língua Brasileira de Sinais - Libras e dá outras providências. Acesso em: https://www.planalto.gov.br/ccivil_03/leis/2002/L10436.htm

[2] Brasil. Decreto nº 5.626, de 22 de dezembro de 2005. Regulamenta a Lei nº 10.436, de 24 de abril de 2002, que dispõe sobre a Língua Brasileira de Sinais - Libras, e o art. 18 da Lei nº 10.098, de 19 de dezembro de 2000. Acesso em: https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5626.htm

[3] Lopes, M. C. P., & Macedo, E. C. (2002). Surdos: processos identitários e educacionais. Editora Autores Associados.

[4] Quadros, R. M. (2010). O tradutor e intérprete de língua brasileira de sinais e língua portuguesa: reflexões sobre práticas discursivas. Editora Arara Azul.

[5] Souza, C. A., & Filho, J. M. (2014). A formação do intérprete de língua de sinais brasileira no Brasil: avanços e desafios. Revista Educação Especial.

[6] lucaaslb. "cnn-libras." GitHub. Acesso em: https://github.com/lucaaslb/cnn-libras.

[7] "Sign Language MNIST." Kaggle. Acesso em: https://www.kaggle.com/datasets/datamunge/sign-language-mnist/data.

[8] ANJO, M. d. S.; PIZZOLATO, E. B.; FEUERSTACK, S. A real-time system to recognize static gestures of brazilian sign language (libras) alphabet using kinect. In: BRAZILIAN COMPUTER SOCIETY.

Proceedings of the 11th Brazilian Symposium on Human Factors in Computing Systems. [S.l.], 2012.

[9] BASTOS, I. L.; ANGELO, M. F.; LOULA, A. C. Recognition of static gestures applied to brazilian sign language (libras). In: IEEE. Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on. [S.l.], 2015.

[10] BRANDT, A. H., LiRANN – Sistema de Reconhecimento de LIBRAS baseado em Redes Neurais Artificiais com Kinect, 2015.

[11] CASTRO, G. Z.., Reconhecimento de Línguas de Sinais Utilizando Redes Neurais Convolucionais e Transferência de Aprendizado, 2020.

[12] DONAHUE, J. et al. Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389, 2014. Disponível em: <http://arxiv.org/abs/1411.4389>.

[13] Jhon at al., Reconhecimento e Tradução de Sinais de Libras para Língua Portuguesa Escrita usando Redes Neurais Profundas, Brasil.

[14] MACHADO, M. C., Classificação Automática de Sinais Visuais da Língua Brasileira de Sinais Representados por Caracterização Espaço-Temporal, 2018.

[15] Mittal, A., Kumar, P., Roy, P.P., Balasubramanian, R., and Chaudhuri, B.B. (2019). A modified lstm model for continuous sign language recognition using leap motion. IEEE Sensors Journal, 19(16), 7056–7063. doi:10.1109/ JSEN.2019.2909837.

[16] Passos, W.L., Araujo, G.M., Gois, J.N., and de Lima, A.A. (2021). A gait energy image-based system for brazilian sign language recognition. IEEE Transactions on Circuits and Systems I: Regular Papers, 68(11), 4761–4771. doi:10.1109/TCSI.2021.3091001.

[17] PIZZOLATO, E. B.; ANJO, M. dos S.; PEDROSO, G. C. Automatic recognition of finger spelling for libras based on a two-layer architecture. In: ACM. Proceedings of the 2010 ACM Symposium on Applied Computing. [S.l.], 2010.

[18] PORFIRIO, A. J. et al. Libras sign language hand configuration recognition based on 3d meshes. In: IEEE. Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. [S.l.], 2013.

[19] TANG, M. Recognizing Hand Gestures with Microsoft's Kinect - Department of Electrical Engineering Stanford University.

[20] VOIGT, J. F., Aprendizagem profunda para reconhecimento de gestos da mão usando imagens e esqueletos com aplicações em libras, 2018.

[21] CECCON, Denny. Conceitos sobre IA - Fundamentos de ML: funções de custo para problemas de classificação, 2019. Disponível em: < https://iaexpert.academy/>. Acesso em: 20, agosto. 2023.

[22] HARRISON, Matt. Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python. (2019). Brasil: Novatec Editora. Brasil: Novatec Editora, 2019.

[23] GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep learning. MIT press, 2016.