Aplicação de Visão Computacional no Reconhecimento de Sinais da Língua Americana de Sinais (ASL)

1st Jamil Soares da Silva Júnior
Instituto de Informatica
Universidade Federal de Goias
Goiania, Brasil
jamiljunior@discente.ufg.br

2nd Letícia de Cerqueira Xavier
Instituto de Informatica
Universidade Federal de Goias
Goiania, Brasil
cerqueiraleticia@discente.ufg.br

3rd Victor Guerreiro Pimenta
Instituto de Informatica
Universidade Federal de Goias
Goiania, Brasil
victorguerreiro@discente.ufg.br

4th Francieli Moreira de Carvalho
Instituto de Informatica
Universidade Federal de Goias
Goiania, Brasil
francielimoreira@discente.ufg.br

***Abstract*—Este projeto de visão computacional tem como objetivo aprimorar a comunicação entre pessoas surdas e ouvintes, focalizando na identificação e reconhecimento da Língua Americana de Sinais (ASL). Inicialmente, concentra-se na identificação dos gestos e sinais correspondentes às letras do alfabeto, abrangendo de A a Z. Isso possibilitará aos usuários uma comunicação mais eficiente em atividades básicas de escrita e leitura, promovendo inclusão e acessibilidade. Embora a aplicação inicial não contemple a identificação imediata de outros gestos e palavras da ASL, o principal objetivo é facilitar a compreensão e uso do alfabeto na língua.**

***Index Terms*—ASL, reconhecimento, sistema, alfabeto, acessibilidade.**

I. Introdução

A tarefa de classificação desempenha um papel crucial em nossa vida diária, uma vez que frequentemente nos deparamos com a necessidade de categorizar objetos com base em suas características (SILVA, 2019). O desenvolvimento de algoritmos capazes de distinguir e agrupar objetos automatiza uma variedade de atividades. Um exemplo notável é a identificação de placas de veículos, onde um algoritmo analisa as letras e as associa ao alfabeto, de maneira semelhante à identificação de símbolos. A habilidade de criar algoritmos para identificar símbolos de uma linguagem também abre possibilidades para a tradução. No entanto, é essencial reconhecer a complexidade da classificação da língua de sinais, frequentemente desafiadora mesmo para indivíduos fluentes na Língua de Sinais Americana (ASL, do inglês American Sign Language). No cenário atual, já existem aplicativos que auxiliam a tradução da língua inglesa para a ASL por meio de animações 3D de intérpretes individuais. Contudo, ainda existe uma lacuna em relação a soluções que possam inverter esse processo, ou seja, reconhecer o significado do gesto por meio de uma câmera e traduzi-lo. Esse desafio é acentuado ainda mais devido à rica diversidade nos movimentos das mãos, exigindo uma distinção precisa das configurações das mãos e dedos. Esse aspecto adiciona complexidade ao reconhecimento do significado do sinal. Tendo em vista que a comunicação é fundamental para a socialização, é crucial explorar abordagens que facilitem a interação entre os usuários da ASL e aqueles menos familiarizados com essa língua, a fim de promover a inclusão social. A utilização de Redes Neurais Artificiais (RNA) para criar sistemas inteligentes capazes de realizar essas traduções surge como um meio de aproximar a comunidade surda, que se comunica por meio da ASL, daqueles que não dominam essa língua. A evolução dos modelos de Aprendizado de Máquina (AM), impulsionados por bibliotecas como TensorFlow, Keras e Numpy, tem sido consideravelmente simplificada. A disponibilidade dessas ferramentas tornou a tecnologia mais acessível e eficaz. Para a construção de modelos de classificação de sinais em imagens capturadas por câmera, optou-se por adotar a arquitetura de Redes Neurais Convolucionais (CNN, do inglês Convolutional Neural Network). Essa escolha se justifica pelo fato de que as CNNs são uma das abordagens mais destacadas no campo de Visão Computacional (VC) nos dias de hoje. As CNNs oferecem diversas vantagens notáveis, incluindo a capacidade de lidar com transformações geométricas (reconhecendo objetos independentemente das alterações de aparência), bem como a capacidade de compartilhar parâmetros e estabelecer conectividade esparsa (LOCA; RAUBER, 2019). Essas redes são especialmente projetadas para extrair padrões de extensos conjuntos de dados de imagens, e demonstram uma notável capacidade de generalização com base nos dados de entrada. A abordagem da CNN representa uma escolha estratégica para enfrentar os desafios de reconhecimento de padrões em imagens geradas por câmeras, permitindo assim uma identificação

precisa de gestos e sinais em tempo real.

II. Fundamentos Teóricos

Este capítulo tem como objetivo realizar uma contextualização de todos os temas necessários para o desenvolvimento deste trabalho. Para tal, esse capítulo se estrutura em seis tópicos, a saber: Língua de Sinais Americana; Visão Computacional; Aprendizado de Máquina; Redes Neurais Artificiais; Aprendizado Profundo e Redes Neurais Convolucionais.

A. Língua de Sinais Americana (ASL)

A Língua de Sinais Americana (ASL) é uma língua visual-gestual vibrante e complexa que desempenha um papel crucial na comunicação e identidade da comunidade surda nos Estados Unidos e em partes do Canadá. A ASL é uma língua completa, com sua própria gramática, sintaxe e léxico, e é utilizada como meio de comunicação primário entre pessoas surdas (Academia de Libras, 2019). A ASL tem uma história rica e evolutiva, que remonta ao século XIX. Desenvolveu-se organicamente dentro das comunidades surdas, emergindo como uma resposta à necessidade de comunicação eficaz e uma forma de expressão cultural única. A língua foi moldada por diversos fatores, incluindo influências linguísticas de línguas de sinais francesas e de origem local, bem como a contribuição da comunidade surda em sua constante evolução. A estrutura da ASL difere fundamentalmente das línguas orais. Em vez de palavras faladas, a ASL utiliza movimentos das mãos, expressões faciais, corporais e espaço para transmitir informações. Os sinais nas ASL podem variar em significado com base na configuração das mãos, movimento, localização e expressões faciais. A gramática da ASL também difere significativamente da gramática da língua falada inglesa, com regras de ordem de palavras distintas e marcadores gramaticais visuais. A ASL não é apenas um meio de comunicação, mas também desempenha um papel central na identidade cultural da comunidade surda. Por meio da ASL, os indivíduos surdos podem compartilhar suas experiências, expressar emoções e pensamentos complexos e se conectar com sua cultura e história compartilhada. Embora a ASL tenha ganhado reconhecimento como uma língua legítima, ainda enfrenta desafios, como a falta de compreensão e acessibilidade por parte da sociedade ouvinte. No entanto, avanços tecnológicos, como o desenvolvimento de aplicativos e recursos de tradução, estão ampliando as oportunidades de interação entre a comunidade surda e a sociedade ouvinte.

B. Visão Computacional

O sistema visual humano é responsável por capturar e transformar os padrões de luz do ambiente externo em informações coerentes sob a forma de imagens. Essa capacidade permite que informações sejam extraídas simultaneamente, bem como processadas e interpretadas (BALLARD, 1982). Inspirada nessa habilidade humana e impulsionada pelos avanços da Inteligência Artificial, surgiu a disciplina da Visão Computacional (VC). A Visão Computacional compreende a capacidade de uma máquina ”enxergar”, ou seja, interpretar o ambiente ao seu redor por meio de imagens capturadas por câmeras de vídeo, sensores, scanners e outros dispositivos. Essas informações possibilitam o reconhecimento, manipulação e análise dos elementos presentes em uma imagem (BALLARD, 1982). A maioria das aplicações de Visão Computacional é derivada de outras áreas de pesquisa e visa resolver problemas específicos, envolvendo desde o reconhecimento de objetos em imagens até a transformação desses objetos em dados processáveis por sistemas especializados. Autores como Davies (2012) e Conci, Azevedo e Leta (2008) descrevem etapas comuns na maioria dos sistemas de Visão Computacional, tais como: Aquisição de Imagem: É a etapa inicial de um sistema de VC, envolvendo a captura de imagens ou conjuntos de imagens por meio de câmeras ou outros sensores. Os pixels dessas imagens contêm informações de luz e propriedades físicas, podendo representar imagens bidimensionais, tridimensionais ou sequências. Pré-processamento: Essa fase ocorre antes da extração efetiva de informações da imagem. Consiste em aplicar métodos específicos para melhorar a identificação de objetos, como realce de contornos, detecção de bordas ou destaque de figuras geométricas. Detecção e segmentação: Nessa etapa, as regiões de interesse na imagem são delimitadas e segmentadas, facilitando o processamento subsequente. Extração de características: Envolve a obtenção de descritores matemáticos que caracterizam aspectos da imagem, como textura, bordas, formas e movimento. Reconhecimento de padrões: Essa etapa inclui a tomada de decisão sobre os dados obtidos. Por exemplo, identificar a categoria à qual um objeto pertence, baseado em critérios predefinidos. A Visão Computacional não apenas simula a capacidade visual humana, mas também oferece oportunidades significativas em diversas áreas, como automação industrial, medicina, veículos autônomos e segurança, melhorando a maneira como interagimos com o mundo digital e físico.

C. Redes Neurais Artificiais e Aprendizado Profundo

Os modelos tradicionalmente utilizados para lidar com problemas de Regressão e Classificação possuem características analíticas e computacionais notáveis, sendo aplicáveis em uma ampla gama de cenários práticos. No entanto, esses modelos enfrentam limitações quando confrontados com dados que possuem múltiplas dimensões (FARIA, 2017). Diante dessa situação, quando se deparam com problemas que envolvem funções não lineares em alta dimensionalidade, frequentemente recorre-se a algoritmos conhecidos como Redes Neurais Artificiais.

A exploração das Redes Neurais Artificiais, também conhecidas como Artificial Neural Networks em inglês, foi largamente influenciada pela observação de que os sistemas de aprendizado biológico consistem em intrincadas redes de neurônios interconectados. Assim, os algoritmos de Redes Neurais Artificiais buscam, em certa medida, emular as capacidades de processamento de um cérebro humano por meio de unidades de processamento simples, que modelam os neurônios biológicos (MITCHELL; MICHALSKI; CARBONELL, 2013).

Vários pesquisadores buscaram simular esse sistema de neurônios biológicos em ambientes computacionais, tendo como base a estrutura e o funcionamento dos neurônios biológicos. Em 1943, McCulloch e Pitts (1943) apresentaram o conceito de Perceptron, um dos primeiros modelos reconhecidos que, de forma simplificada, incorporava os componentes e o funcionamento de um neurônio biológico. No cerne desse conceito, as Redes Neurais Artificiais realizam cálculos como a soma ponderada das várias entradas, seguida da aplicação de uma função e a propagação do resultado adiante. Consequentemente, a função logística do Perceptron é representada por $f(b_k + n\sum_{i=1}^{n} x_i w_{ki})$, permitindo uma análise mais detalhada do seu funcionamento e das variáveis envolvidas.

Sinais de Entrada ($x_1, x_2, ..., x_n$): Estes são os dados externos que alimentam o início do modelo preditivo após passarem por uma normalização.

Pesos Sinápticos ($w_1, w_2, ..., w_n$): São valores atribuídos para ponderar os sinais de entrada da rede. Esses valores são ajustados durante o processo de treinamento da rede neural.

Combinador Linear ($\sum$): Soma os sinais de entrada após serem multiplicados pelos seus respectivos pesos sinápticos, gerando um potencial de ativação.

Limiar de Ativação (θ): Define o ponto no qual o resultado do combinador linear leva à ativação do neurônio. É um valor determinante para disparar ou não a saída do neurônio.

Potencial de Ativação (u_k): É o resultado da diferença entre o valor produzido pelo combinador linear e o limiar de ativação. Esse potencial é decisivo para que o neurônio seja ativado ou não, resultando em uma saída binária.

Função de Ativação (ϕ): Limita a saída do neurônio a um intervalo específico de valores. É uma etapa importante para introduzir não-linearidades no processo de ativação.

Sinal de Saída (y_k): Representa o valor resultante da saída do neurônio. Esse sinal pode ser utilizado como entrada para outros neurônios subsequentes na rede.

Cada neurônio na rede possui a capacidade de aprender padrões que são linearmente separáveis. Por isso, o Perceptron, que é a primeira forma de Rede Neural Artificial é composto por apenas um neurônio, é adequado somente para resolver problemas de classificação binária que possuem uma separação linear. Ao incorporar mais neurônios, a estrutura da rede neural se divide em três camadas: a camada de entrada, responsável por introduzir os dados na rede; a camada oculta, que executa o processamento dos dados; e, por último, a camada de saída, que exibe o resultado final da operação.

A combinação de vários neurônios em uma camada oculta, que é independente das camadas de entrada e saída, viabiliza a detecção de padrões que não são lineares (MARTINS, 2018). Essa abordagem dá origem à estrutura conhecida como Perceptron Multicamada (MLP, de Multi-Layer Perceptron), que consiste em redes neurais com múltiplas camadas ocultas e capacidade de solucionar uma ampla gama de problemas, tanto em classificação quanto em regressão, particularmente em domínios não lineares. O uso mais comum do MLP ocorre na resolução de desafios de classificação com características não lineares.

No contexto do treinamento de uma MLP, existem dois algoritmos fundamentais: o algoritmo de propagação e o algoritmo de retropropagação (ou backpropagation) (CARVALHO, 2014). O algoritmo de retropropagação é mais amplamente empregado, pois inclui a etapa de propagação. Esse método se baseia na regra de aprendizado por correção de erro, onde o erro nas saídas é calculado e utilizado como entrada para a próxima etapa do processo (CONTE et al., 2008). De acordo com Teixeira (2014), a etapa de propagação envolve disseminar os dados oriundos da camada de entrada por toda a rede, resultando em um conjunto de saídas. Essa saída é comparada com os valores desejados. A partir desse ponto, o processo de retropropagação é iniciado, ajustando os pesos da rede com base na diferença obtida, visando fazer com que a resposta da rede se aproxime do resultado desejado.

A arquitetura MLP vai além de uma única camada oculta, incorporando múltiplas camadas ocultas. Estruturas semelhantes, mas com várias camadas ocultas adicionais, se enquadram no domínio das Redes Neurais Profundas (DL, de Deep Learning). Ter um número maior de neurônios e camadas ocultas oferece maior capacidade de generalização, mas também requer um conjunto de dados abrangente, hardware especializado e um período prolongado de treinamento para obter resultados significativos (LENT; JR, 2018).

As Redes Neurais Profundas operam com a combinação de neurônios artificiais e conjuntos de dados extensos, resultando em estruturas como ilustrado na Figura 6. As arquiteturas MLP conseguem identificar padrões mais sutis no conjunto de treinamento, o que habilita a rede neural a tomar decisões levando em conta informações globais. Isso se traduz em resultados mais robustos, mesmo quando lidando com dados de múltiplas dimensões (PONTI; COSTA, 2018).

D. Rede Neural Convolucional

Dentro das várias arquiteturas de Aprendizado Profundo, as Redes Neurais Convolucionais (CNNs) emergiram como o estado da arte para resolver desafios relacionados a áudio, vídeo e imagens (LECUN; BENGIO; HINTON, 2015). A estrutura da CNN é uma evolução das redes de Perceptron de Múltiplas Camadas, inspirada pelo processo biológico de processamento visual. Similar aos métodos tradicionais de Visão Computacional, as CNNs têm a capacidade de extrair informações de dados visuais, preservando as relações de proximidade entre os pixels da imagem durante o processamento na rede (VARGAS; PAES; VASCONCELOS, 2016).

REFERENCES

SILVA, G. F. d. Tecnologias assistiva e a deep learning: aplicativo com reconhecimento de imagem no auxílio a deficientes visuais. 2019.

LOCA, A.; RAUBER, T. Uso de uma rede neural convolucional unidimensional para detecção de falhas em processos industriais. In: SBC. Anais da XIX Escola Regional de Computação Bahia, Alagoas e Sergipe. [S.l.], 2019.

BALLARD, D. H. Ballard d. and brown cm 1982 computer vision. Image, 1982.

DAVIES, E. R. Computer and machine vision: theory, algorithms, practicalities. [S.l.]: Academic Press, 2012.

CONCI, A.; AZEVEDO, E.; LETA, F. R. Computação gráfica. [S.l.]: Elsevier, 2008.

FARIA, A. C. B. Classificação de estilos visuais utilizando aprendizado profundo. 2017.

MITCHELL, R.; MICHALSKI, J.; CARBONELL, T. An artificial intelligence approach. [S.l.]: Springer, 2013.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, Springer, v. 5.

MARTINS, V. E. Aplicação de deep learning para detecção de veias em carne suína. 2018.

CARVALHO, H. M. Aprendizado de máquina voltado para mineração de dados: árvores de decisão. 2014.

TEIXEIRA, J. de F. Inteligência artificial. [S.l.]: Pia Sociedade de São Paulo-Editora Paulus, 2014.

LENT, D. M. B.; JR, M. L. P. Detecção de anomalias utilizando redes neurais convolucionais. 2018.

PONTI, M. A.; COSTA, G. B. P. D. Como funciona o deep learning. arXiv preprint arXiv:1806.07908, 2018.

LECUN, Y. et al. Comparison of learning algorithms for handwritten digit recognition. In: PERTH, AUSTRALIA. International conference on artificial neural networks. [S.l.], 1995. v. 60.

ARGAS, A. C. G.; PAES, A.; VASCONCELOS, C. N. Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. In: SN. Proceedings of the xxix conference on graphics, patterns and images. [S.l.], 2016.