Reconhecimento de sinais dinâmicos em LIBRAS

Ana Cristina Oliveira, Icaro Vasconcelos, Letícia Rezende, e Mateus Oliveira

Abstract—A área de reconhecimento automático de linguagem de sinais tem o intuito de atenuar possíveis obstáculos existentes para pessoas surdas e/ou com deficiência auditiva, e assim melhorar e aumentar a comunicação e integração com a sociedade de maioria ouvinte. Sendo assim necessária a aplicação de técnicas de reconhecimento de imagem para desenvolvimento de tecnologias que possam estar voltadas para esta finalidade, facilitando a comunicação entre pessoas que tenham ou não o conhecimento de LIBRAS. O objetivo deste trabalho além de se basear em e contribuir para pesquisas nessa área, foi elaborar e analisar a eficiência do protótipo de uma ferramenta que realiza o reconhecimento de sinais em LIBRAS (Língua Brasileira de Sinais). Após explorar algumas técnicas de reconhecimento de imagens com o auxílio de uma rede neural, percebeu-se que o protótipo criado é um passo inicial para reconhecimento de sinais, com ainda pouca precisão, mas promissor. Possíveis melhorias que podem aumentar sua precisão e eficiência foram identificadas.

Abstract—The automatic recognition of the sign language study field is intended to mitigate possible existing obstacles for people who are deaf or have a hearing impairment, and thus improve and increase communication and integration with the society of the listener majority. Therefore, it is necessary to apply image recognition techniques for the development of technologies that may be aimed at this purpose, facilitating communication between people who have or do not have the knowledge of LIBRAS. The objective of this work, in addition to being based on and contributing to research in this area, was to elaborate and analyze the efficiency of the prototype of a tool that performs the recognition of signs in LIBRAS (Brazilian Sign Language). After exploring some image recognition techniques with the assistance of a neural network, it was noticed that the prototype created is an initial step toward signal recognition, with small precision, but promising. Possible improvements that can increase its accuracy and efficiency have been identified.

Index Terms—Reconhecimento, LIBRAS, visão computacional.

1 Introdução

A Língua Brasileira de Sinais, também conhecida como LIBRAS, é uma língua gesto-visual de extrema importância para a inclusão social. Muito utilizada na comunicação de pessoas surdas, em 24 de abril de 2002 através da lei nº 10.436, foi reconhecida como um meio legal de comunicação. Os sinais da linguagem são realizados através de gestos formulados por movimentos das mãos e articulações e também expressões corporais e faciais. Os gestos desta linguagem, também conhecidos como sinais, são realizados pela junção de movimentos das mãos e articulações e também de expressões faciais e corporais. Em uma pesquisa divulgada em 2020, de acordo com o IBGE - Instituto Brasileiro de Geografia Estatística, cerca de 10 milhões de pessoas no Brasil são surdas. A estimativa, segundo a OMS - Organização Mundial da Saúde, é de que até 2050 aproximadamente 900 milhões de pessoas no mundo possam desenvolver surdez.

Atualmente, existem diversas abordagens computacionais para o reconhecimento de linguagem de sinais que utilizam de variadas metodologias como redes neurais, reconhecimento da mão através da cor da pele, classificação dos sinais utilizando k-NN (k-Nearest Neighbors), dentre outras. Apesar de diversas, nem todas são muito assertivas já que as linguagens são dinâmicas e vários fatores como, por exemplo, iluminação, qualidade da imagem capturada e ruídos na imagem, dificultam o reconhecimento dos sinais. No contexto do Brasil, ainda não há um reconhecimento computacional muito preciso quando se trata de sinais da LIBRAS.

A falta do reconhecimento dos sinais de libras além de acentuar barreiras de comunicação, dificulta a inclusão de pessoas surdas no âmbito tecnológico e o desenvolvimento de tecnologias que possam vir a acrescentar na vida das pessoas que compõem a comunidade surda brasileira.

A utilização e o reconhecimento da LIBRAS também são formas de garantir a preservação da identidade das pessoas surdas e contribuem para a valorização e reconhecimento de sua cultura. Infelizmente no Brasil não há softwares eficientes no reconhecimento e traduções dos sinais da língua.

Havendo um software que possibilite seu reconhecimento e tradução, é possível tornar a língua acessível para ouvintes que não possuíram contato com a língua anteriormente, facilitando e expandindo os horizontes de comunicação de pessoas surdas fluentes em LIBRAS. Esse software também proporcionaria um novo meio de interação, isentando surdos de serem obrigados a utilizarem do português como principal língua para interação computacional.

A pergunta pesquisa deste trabalho é: o reconhecimento computacional da LIBRAS pode ser alcançado e se tornar um meio que auxiliará pessoas surdas na inclusão digital? O objetivo do presente trabalho é desenvolver e avaliar uma ferramenta computacional que contribua com o reconhecimento e tradução dos sinais da LIBRAS, tendo como foco o reconhecimento de seu alfabeto. Almeja-se que esta ferramenta seja mais assertiva que as que encontramos e utilizamos para identificar as dificuldades do reconhecimento dos sinais, e possa ser evoluída com o tempo para

identificar mais sinais da linguagem. Técnicas de processamento digital de imagens, em conjunto com uma rede neural, serão utilizadas para compor a ferramenta que terá como objetivo específico identificar e traduzir símbolos e gestos que compõem o alfabeto da LIBRAS. A ferramenta será testada com diversas imagens e vídeos e os resultados obtidos serão descritos neste trabalho.

2 Trabalhos Correlatos

No trabalho [1] podemos verificar que apesar de existirem diversos trabalhos na área de processamento digital de imagens focados em realizar o reconhecimento de partes do corpo humano, atualmente não existem abordagens relevantes interessadas em detectar pontos-chave em mãos para imagens RGB. Uma abordagem como essa é interessante para resolver problemas do mundo real relacionados ao uso das mãos em papéis importantíssimos em nossas atividades diárias: ao usar ferramentas, tocar instrumentos musicais, entre outras. Dada essa questão, os autores decidiram propor um método que permitisse o rastreamento 2D de mãos em tempo real em vídeo de visualização única e também a captura de movimento de mãos em 3D. A metodologia escolhida por eles foi a Multiview Bootstrapped Training, um sistema que irá fornecer visualizações da mão onde a detecção de pontos chave seja fácil, para que logo após seja feita uma triangulação das posições 3D dos pontos-chave. O resultado disso é reutilizado para casos onde há visões difíceis e com falhas na detecção, e logo após passar por esse processo, o detector é treinado novamente para que seja aprimorado e consiga agora realizar o reconhecimento em visualizações mais complexas.

Em [2] uma pesquisa realizada por membros de universidades situadas em Campinas, Santo André e São Paulo, descreve uma abordagem de solução para uma ASLR (tecnologia de reconhecimento automático de linguagem de sinais). Tal tecnologia é conhecida por traduzir gestos de linguagens de sinais para uma língua de pessoas ouvintes. A solução utiliza redes neurais convolucionais e codificação para reconhecimento de expressões faciais, tais são, respectivamente: CNN padrão, uma combinação de CNN e LSTM, e Facial Action Coding System (FACS). O trabalho ressalta que expressões faciais presentes na língua de sinais nem sempre estão relacionadas às emoções humanas, já que as expressões faciais de emoções estão apenas em uma parcela das expressões faciais comuns na linguagem de sinais, tal parcela é conhecida como expressões faciais afetivas (AFE). Com isso, as taxas de precisão de uma ASLR se tornariam melhores caso o reconhecimento de expressões faciais fosse implementada.

A metodologia do trabalho trata-se, com o auxílio de especialistas em Libras, em utilizar as unidades de ação (AU) rotuladas pelo FACS para codificar as expressões de rosto ligadas tanto à emoções humanas e faz adaptações para casos em que as expressões faciais não denotem emoções humanas, aumentando ainda mais a quantidade de rótulos, utilizando-se 80 categorias. Além disso, regiões da face são segmentadas em duas partes: porção inferior e porção superior para a análise ao invés de se usar a face inteira, como também um banco de dados para vídeos de intérpretes de Libras em que se há movimentações de cabeça (HM-Libras). No trabalho proposto tiveram problemas com vídeos em que a iluminação estava baixa e/ou com variação no fundo.

A metodologia utilizada em [3] aborda um conjunto de testes através do treinamento de múltiplas imagens de cada letra que compõem o alfabeto da American Sign Language. Cada letra possui pontos de referência, sendo esses pontos responsáveis por produzirem dois descritores (um de 72 pontos e outro de 180 pontos) e que são armazenados em um banco de dados. O descritor da imagem que está sendo testada é comparado com os descritores presentes no banco de dados. A distância euclidiana é responsável por classificar imagens de teste na letra reconhecida.

A metodologia deste trabalho mostrou-se divergente dos demais analisados, apresentando uma proposta mais inovadora. Notou-se que a ideia adotada proporcionou um processamento significativamente mais eficiente em relação ao tempo. Obteve-se uma taxa de acerto notável, com um total de 75% com o descritor de 72 pontos e 79.9% com um descritor de 180 pontos. Também pôde-se observar e contemplar um conteúdo explicativo e resultados bem representados neste trabalho, o que facilitou seu entendimento.

Apesar da notável taxa de acerto, a solução apresentada por este trabalho mostrou resultados não satisfatórios para algumas letras específicas, o que pode ser atribuído ao fato de algumas letras possuírem representações similares na ASL. Apesar de justificável, o trabalho também apresentou uma abordagem restrita no que diz respeito à quantidade de termos reconhecidos.

O trabalho [4] é motivado por fornecer classificação de sinais de séries temporais e uma validação robusta e abordagens de teste. O estudo propõe classificar Língua de Sinais Americana com base em dados fornecidos pelo LeapMotion. O objetivo da pesquisa também é propor uma validação cruzada robusta de k-fold independente do usuário e testes para a tal, em contraste com estudos anteriores que se concentraram principalmente em testes intra-usuário de seus modelos. O trabalho utiliza do dispositivo LeapMotion para detectar e mapear as mãos dos usuários.

Foram utilizados os modelos k-nearest neighbors (k-NN), random forest (RF) e support vector para o treinamento; e para a implementação foi usado o modelo DeepConvLSTM que integra camadas convolucionais e recorrentes com células de Long-Short Term Memory e o modelo ConvNet. Para evitar o overfitting do modelo e ajudar a generalizar a classificação, o aumento de dados foi considerado. Os resultados demonstraram que o maior valor de precisão foi do modelo DeepConvLSTM com aumento de dados em comparação com todos os outros modelos (p ¡ 0,05 para cada comparação), exceto com ConvNet com aumento de dados (p ¿ 0,05).

O trabalho [5] baseia-se em processamento de imagens no MATLAB para o método de reconhecimento de sinais, utilizando uma câmera webcam para recepção dos vídeos. Das imagens são extraídos os recursos que contém os sinais manuais. Estes são comparados com os recursos já em banco de dados com técnicas de processamento de imagem presentes na função ”bagOfFeatures”, assim de acordo com classificadores treinados do banco de dados, realiza a classificação conforme as semelhanças presentes.

Então é construído o vocabulário visual com imagens de sinais diferentes da American Sign Language(ASL).

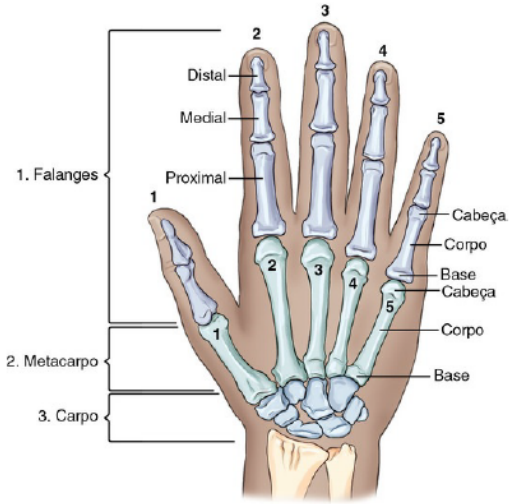
Tendo cada imagem 277x277 pixels de dimensão. Usando da técnica Speeded UP Robust Features (SURF) e das K-means clustering, que particiona os recursos. Depois de particionados os recursos com características semelhantes, são usados para a geração do histograma de ocorrências de palavras visuais, que será a base de treinamento da categoria da imagem classificadora. E é treinado utilizando SVM (Support Vector Machine). Assim é obtido uma precisão de 85%, sendo satisfatória que foi conduzido para reconhecimento de três letras A, B e C.

O objetivo do estudo [6] é propor um novo modelo para aumentar a precisão dos métodos existentes para reconhecimento ASL utilizando quatro bancos de dados publicamente disponíveis. O algoritmo utiliza um modelo CNN, que contribui para o desempenho ideal por uma seleção adequada de camadas de convolução e o número de neurônios. É também proposto uma nova arquitetura para esse modelo CNN que foi chamado pelos autores de SLRNet-8, que consiste de seis camadas de convolução, três camadas de pooling e uma camada totalmente conectada além das camadas de entrada-saída. O aumento de dados também foi empregado, alterando atributos das imagens aleatoriamente. Os resultados apresentam que o modelo reconhece sinais de dígito e alfabeto de cada conjunto de dados com cerca de 100% de precisão. A menor precisão (99,90%) foi relatada para os dígitos do conjunto de dados de dígitos da língua de sinais.

3 METODOLOGIA

Nesta seção aborda-se a construção de uma aplicação que auxilia na detecção de símbolos do alfabeto da LIBRAS. O intuito é realizar a adaptação dos métodos vistos na revisão bibliográfica, algumas técnicas apresentadas por Simon em [1] e adicionar a capacidade de detectar sinais da LIBRAS para a aplicação. A aplicação tem como base 3 módulos: i) a verificação do ponto, identificando se o mesmo se encontra "acima" ou "abaixo"; ii) verificação da configuração dos dedos, se estes se encontram esticados ou recolhidos; iii) a análise da proximidade entre os pontos chaves encontrados. O resultado final será composto por um conjunto de dados vetoriais equivalente as configurações dos símbolos do alfabeto da LIBRAS.

No início, realiza-se a captura de cada frame do vídeo analisado para obter-se os pontos chaves presentes na configuração da mão. Após a coleta desses pontos, o próximo passo é o processamento dos mesmos. Estes pontos levantados pela rede neural (CMU-Perceptual-Computing-Lab, 2021) foram utilizados como base para determinar as configurações de mão, tendo assim, 22 pontos fornecidos pela rede que são catalogados em relação as partes que compõem a mão, como mostra a figura:



A princípio, obtém-se os pontos mais significativos para a identificação do sinal, os quais são capazes de determinar a qual letra do alfabeto da LIBRAS a configuração de mão se refere. A determinação da forma como a mão está articulada, para "cima" ou para "baixo" é feita pelo módulo de comparação das alturas dos pontos na vertical e horizontal. Tendo como início o ponto compreendido no carpo, pode-se excluir as possibilidades de configurações de mão que são representadas torcendo a articulação do punho, como a letra B, que possui em sua configuração o punho abaixo dos metacarpo e os falanges. Com os resultados obtidos pode-se analisar a posição dos falanges comparados a palma da mão, onde cada sinal apresenta uma configuração de mão específica, podendo-se explorar dessa característica para constatar se os falanges estão "dobrados" ou "esticados". Em sequência, verifica-se a distância entre os dedos para definir se estão na mesma altura, como pode-se observar na configuração de mão da letra U, que possui os dedos médio e indicador próximos e esticados, já os demais dobrados e também próximos. O vetor que é obtido como resultado desta arquitetura possui a descrição da configuração do sinal, simplificando o reconhecimento desse símbolo pelo alfabeto. Python foi a linguagem utilizada para o desenvolvimento da aplicação deste trabalho. Aproveitou-se também da biblioteca OpenCV e o deep learning framework Caffe, que apresenta ferramentas e métodos para manipulações em imagens. Através da rede neural utilizada mapeou-se os pontos chaves da mão em ações e ângulos distintos. A partir da imagem do frame fornecida como entrada para a rede neural, o resultado obtido com os pontos mapeados percorre a arquitetura proposta e resulta na identificação do sinal.

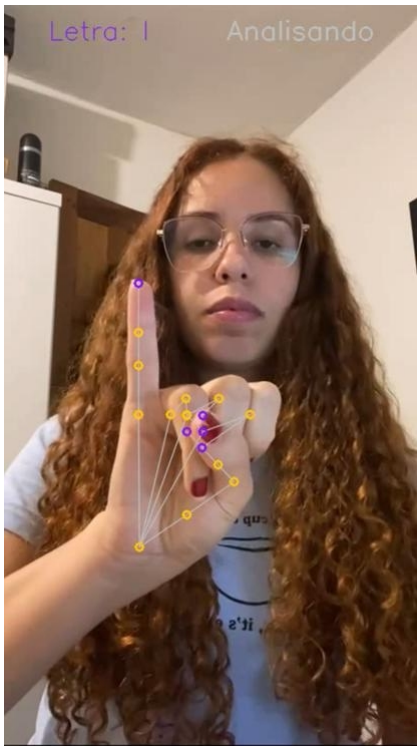
4 RESULTADOS

Para obtenção dos resultados, testamos em dois cenários de testes distintos. Em um primeiro momento verificamos a interferência da qualidade do vídeo em diferentes situações, ambientes, porcentagem de ruído, iluminação e qualidades de gravações. Notou-se que havia interferência também devido a proximidade da mão ao rosto e a câmera, e posições mais complexas dos dedos, como quando estes se sobrepunham. Observou-se também que o espelhamento dos sinais interferia, caso eles fossem feitos pela mão direita ou esquerda. Devido a isso, elaborou-se um cenário de testes com

os espelhamentos da disposição das configurações de mão e mesmas interferências do ambiente, e outro analisando a proximidade da mão à lente que está captando a imagem fundo sem interferências. Para a análise da eficiência foi utilizada a porcentagem de classificações corretas, o número de acertos nas classificações dividido pelo total de sinais.

No primeiro cenário, foram realizados testes com com as letras P, D e I, para os quais foram submetidos 3 vídeos à rede, com baixa, média e alta luminosidade. Neste contexto, os vídeos submetidos estavam com a imagem invertida. Com isso, obtivemos uma taxa de acerto de 11,11% de acerto, 1 acerto em 9 sinais testados. As imagens foram invertidas após o primeiro teste para uma nova avaliação dos acertos, e obteve-se uma taxa de acerto similar mesmo após a inversão.

Após a inversão, em um contexto em que a mão estava mais próxima a câmera que captava a imagem, a letra D foi reconhecida após invertida. No mesmo processo, a letra I que havia sido reconhecida anteriormente, não foi reconhecida após o espelhamento.



Independente da inversão, pôde-se observar baixa eficácia na identificação de letras com sobreposição de dedos e dedos com ângulos mais complexos, havendo dificuldade em verificar a ordem correta dos dedos.

Logo após, testou-se as letras com configurações de mão com a disposição de dedos menos complexa. Foram testadas as letras B, O, N e I, nas condições novamente de baixa, média e alta luminosidade com rostos aparecendo no vídeo. Neste cenário, obteve-se 50% de acerto, 6 acertos em 12 sinais testados.

Outro cenário de testes foi realizado com a imagem focando apenas na mão, o fundo do vídeo em branco e com o ambiente bem iluminado. Neste contexto, testou-se as letras A, B, C, D, E e F. Houve 33,33% de taxa de acerto, 6 acertos em 18 sinais testados.

5 Conclusão

Através do presente trabalho desenvolvido podemos verificar a importância do uso de tecnologias para auxiliar na acessibilidade da comunicação de pessoas surdas e os ganhos significativos que a tecnologia traz para este fim. A ferramenta desenvolvida e testada apresentou eficácia baixa de acerto, considerando-se erros derivados de posicionamento, iluminação, mão em que o sinal foi realizado e complexidade do sinal. Sistemas que se baseiam em visão computacional são dependentes de uma diversa quantidade de detalhes e métricas para serem benéficos e efetivos na realização de suas tarefas. Como melhorias para a ferramenta, tem-se como pensamento a implementação de funções de análise e comparação de frames sequenciais para solucionar as questões que envolvem classificação dos sinais que possuem movimento em suas configurações. Outra possível melhoria seria a identificação e correção de defeitos presentes nas imagens com baixa iluminação e mal posicionamento da configuração de mão. Uma interface gráfica também auxiliaria na manipulação e utilização da aplicação.

6 Referências

[1] SIMON, Tomas et al. Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017. p. 1145-1153.

[2] SILVA, Emely Pujólli da et al. Recognition of affective and grammatical facial expressions: a study for Brazilian sign language. In: European Conference on Computer Vision. Springer, Cham, 2020. p. 218-236.

[3] GAUTAM, Amit Kumar; KAUSHIK, Ajay. American sign language recognition system using image processing method. International Journal on Computer Science and Engineering (IJCSE), v. 9, n. 07, 2017.

[4] HERNANDEZ, Vincent; SUZUKI, Tomoya; VENTURE, Gentiane. Convolutional and recurrent neural network for human activity recognition: Application on American sign language. PloS one, v. 15, n. 2, p. e0228869, 2020.

[5] SRIDEVI, Parama et al. Sign Language recognition for Speech and Hearing Impaired by Image processing in matlab. In: 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). IEEE, 2018. p. 1-4.

[6] RAHMAN, Md Moklesur et al. A new benchmark on american sign language recognition using convolutional neural network. In: 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE, 2019. p. 1-6.