UNIVERSIDADE FEDERAL DE OURO PRETO
DEPARTAMENTO DE CIÊNCIA DE COMPUTAÇÃO
PROCESSAMENTO DIGITAL DE IMAGENS

Projeto de Pesquisa

RECOGNITION OF DYNAMIC SIGNS IN A SIGN LANGUAGE

Alunos:

Gustavo Lucas Moreira,

Paula Franca Toledo,

Valmir Rodrigues Bueno Júnior.

Resumo

O reconhecimento automático de linguagens de sinais promete facilitar a comunicação entre a comunidade surda com o não praticantes de LIBRAS. Apesar de existirem estudos atuais para técnicas de reconhecimento de língua de sinais, este ainda é um assunto complexo. A inclusão dos surdos na sociedade tem enfrentado a falta de conhecimento de grande parte dos ouvintes sobre línguas de sinais, tornando consideravelmente complicada a comunicação entre eles. Assim, pode-se pensar na aplicação de técnicas de reconhecimento de imagens no desenvolvimento de tecnologias que auxiliem a comunicação entre surdos e ouvintes. A finalidade deste trabalho é elaborar um protótipo de uma ferramenta que obtenha o reconhecimento de sinais do alfabeto da Língua Brasileira de Sinais (LIBRAS), objetivando a facilitação da comunicação entre os ouvintes que não tenham conhecimentos sobre LIBRAS e os sujeitos surdos, além de procurar contribuir para pesquisas nesse cenário.

1 Introdução

É sabido que existem comunidades formadas por surdos, compostas de milhões de pessoas ao redor do mundo. Aproximadamente 400 milhões de pessoas no mundo possuem perda auditiva "incapacitante"e uma parcela dessas pessoas são totalmente surdas (WHO, 2021). Devido a necessidade das pessoas se comunicarem, faz-se necessário a criação de uma língua, tão complexa e importante quanto qualquer outra: as línguas envolvendo as necessidades especiais.

A língua gestual, também conhecida por espaço-visual, é transmitida através de sinais e recebida através da visão. Especificamente denominadas como Língua de Sinais (LS), existe uma diversidade grande de LS pela qual a comunidade surda utiliza como meio de comunicação, onde cada qual possui características próprias. No Brasil, a Língua Brasileira de Sinais (LIBRAS) é a LS oficial (CIVIL, 2002).

Visto a constante evolução no campo das tecnologias da informação, sobre a perspectiva de inclusão que cingem a acessibilidade, a criação de meios de comunicação e expressão, no caso desse projeto, da comunidade surda no ambiente digital com a interação com a comunidade ouvinte não praticamente de LIBRAS. Considerando que a comunidade de usuários de LIBRAS é formada principalmente por: *a)* os deficientes auditivos que tem como língua nativa; *b)* surdos que aprenderam a LIBRAS ainda durante o período de alfabetização; *c)* indivíduos que se relacionam com a comunidade surda; *d)* e por aqueles que desejam aprender por interesses pessoais. Evidência a dificuldade na interação de comunicação entre a comunidade surda e ouvinte, mostrando que ficam reclusas e as margens da sociedade.

Uma contribuição importante na inclusão digital dessas pessoas seria o reconhecimento de LIBRAS pelos computadores (Koroishi & Silva, 2015). Porém, o reconhecimento das línguas de sinais pelos computadores ainda é uma área aberta à pesquisa. Duas são as abordagens usadas para o reconhecimento das línguas de sinais, sendo elas:

- Uso de luvas especiais para facilitar a detecção e o rastreamento dos movimentos das mãos.
- A outra se baseia nas técnicas de Visão Computacional.

A abordagem que tem-se apresentado promissora é a da aplicação de técnicas compreendidas no campo da visão computacional, como processamento de imagens, redes neurais convolucionais

entre outras, dado a facilidade de acesso dos usuários à tecnologia por não exigir gastos adicionais como a abordagem das luvas especiais. (do Nascimento Siqueira, n.d.).

Nesse sentido, um sistema tecnológico digital possibilite a comunicação entre a comunidade surda com a ouvinte e vice-versa, reduzirá o impacto no desenvolvimento sociocultural e digital do individuo surdo e sua valorização perante a sociedade, além de evidenciar os anseios e reduzir as aflições que atingem a comunidade.

Tendo em vista o exposto, o presente estudo tem como objetivo apresentar uma proposta de reconhecimento automático de símbolos em LIBRAS.

1.1 Objetivos

O objetivo geral deste trabalho é desenvolver uma aplicação capaz de realizar a detecção e a classificação de imagens que contenham os símbolos do alfabeto em LIBRAS.

1.2 Organização do Trabalho

O restante deste trabalho encontra-se organizado como se segue. No Capitulo 2, é apresentado o referencial bibliográfico, relacionado ao tema proposto e necessário para o entendimento deste trabalho. No Capitulo 3, a metodologia proposta neste trabalho, envolvendo suas características é descrita. No Capitulo 4, os experimentos realizados são apresentados junto aos resultados obtidos. E por fim, no Capitulo 5, são apresentados as conclusões e perspectivas de trabalho futuro.

2 Revisão Bibliográfica

Esta seção descreve, com o apoio das principais referências bibliográficas, aspectos e questões de forma a sustentar o presente trabalho, auxiliar na definição dos objetivos e na delimitação ao tema proposto. A subseção está organizada como se segue. A Subseção 2.1 discorre sobre os trabalhos relacionados.

2.1 Trabalhos Relacionados

Na literatura , há diversos trabalhos relacionados ao reconhecimento e a tradução de línguas de sinais por computadores. É possível encontrar diferentes abordagens computacionais para o reconhecimento de linguagens de sinais, cada uma delas utilizando metodologias diferentes. Serão apresentados a seguir metodologias abordadas em diferentes artigos revisados.

Primeiramente, em Saad et al. (2016) é realizado a captura de imagens gestuais via *Kinect* e aplica treino da Rede Neural Artificial (RNA), entrelaçados em uma gama de conceitos e métodos, possibilitando distinguir os gestos obtidos, que define as letras do alfabeto da Libras. Para reconhecimento gestuais, utilizaram técnicas de rastreamento de movimento, processamento de imagens e inteligência artificial. O cerne é rastrear o movimento da mão do usuário e assim as imagens terão seu fundo removido e convertidos para escala de preto. E posterior convertidas para escala de cinza e aplicando o filtro laplaciano, destacando as bordas contidas na imagem. Uma rede neural artificial, com a matriz resultante da imagem, é treinada para reconhecer o gesto, obtendo 88% de acertos.

Semelhantemente, Barros Junior (2016) desenovelaram uma aplicação de reconhecimento das 16 configurações de mãos da Libras, executadas por cinco pessoas. As imagens, capturas pelo sensor de profundidade do *Kinect* foram convertidas para escala de cinza. Na etapa do pré-processamento foram extraídas apenas as regiões onde se localizam as mãos. Uma CNN foi utilizada para extração de características e classificação, alcançando acurácia de 87.5%.

Ribeiro & Rodrigues (2017) apresentam o mapeamento das letras do alfabetos em libras, para criação de um *dataset* de imagens sólidas, utilizando a abordagem presente na biblioteca *Opencv (Open Source Computer Vision Library)* e seus classificadores, mapea-se os sinais que expressam as letras do alfabeto assim capazes de recepcionar uma imagem através de uma câmera estacionária calibrada, identificar o sinal capturado e traduzir para a língua portuguesa.

Outro artigo, Sridevi et al. (2018) apresenta um método de reconhecimento de sinais baseado em processamento de imagem no MATLAB, utilizando uma *webcam* para recepção do video e posteriori extração de recursos da imagem contendo os sinais manuais, estes recursos são comparados com os recursos das imagens do banco de dados e logo técnicas de processamento de imagem presentes na função MATLAB *"bagOfFeatures"* a classifica de acordo com as semelhan-

ças presentes nos classificadores treinados do banco de dados, pelo qual é construído o vocabulário visual com imagens de sinais diferentes da *American Sign Language(ASL)* (cada imagem com dimensão de 277x277 pixeis) partir da técnica SURF *(Speeded UP Robust Features)* e fazendo uso da *K-means clustering* particiona-se os recursos em k, representando o conjunto de recursos com características semelhantes, apos gera-se o histograma das ocorrências dessas palavras visuais, este será a base para treinar a categoria da imagem classificadora e treinado utilizando *Support Vector Machine(SVM)* quadrática. Obtendo uma precisão satisfatória de 85%, conduzidos para o reconhecimento das três letras diferentes (A, B e C) da *ASL*.

Analisando Nandhini et al. (2021) que aborda adoção da metologia de *Convolutional Neural Network (CNN)* para treinamento e classificação imagens contendo sinais, e utilização da filtragem das imagens, devido as tonalidades de pele e as condições de iluminação ambiente poderem prejudicar a classificação. O sistema proposto consiste em determinar a região do símbolo mostrado nas mãos, redimensiona-la e posteriori tem-se o processo de conversão de cor a escala de cinza sobre os pixels RGB, subtração do segundo plano e mascaramento da imagem para assim ocultar partes e dar ênfase as porções das bordas e características das imagens, a imagem em tons de cinza é convertida em uma imagem binária, o processo CNN, é responsável pela derivação de múltiplas características dos dados, treinando sobre um total de 1750 imagens estáticas e sendo capaz de classificar 125 palavras, apresentando uma acurácia de 90%.

Em Voigt et al. (2018) propõe uma *Deep Learning* para reconhecimento de gestos estáticos e dinâmicos da mão, com aplicações em sinais de LIBRAS. Através de dados capturados pelo dispositivo *Leap Motion*, incluindo tanto imagens quanto esqueletos da palma da mão, buscando um arquitetura de rede neural para reconhecer os gestos. Nisso a metodologia abordada consiste em três etapas. Primeira, reconhecer os gestos estáticos (poses) usando redes *perceptron* multicamadas (MLP) para os dados do esqueleto, redes convolucionais (CNN) para as imagens, e redes de múltiplas entradas, utilizando ambos os tipos de informação. Segunda, classificação individual dos gestos que incluam movimento (letras: H, J, K, W, X, Y e Z), e para tanto inclui-se camadas recorrentes *Long Short-Term Memory* (LSTM) e ainda assim aplica-se Transferência de Aprendizado nos blocos convolucionais, trazendo os parâmetros já treinados com as poses estáticas para dentro da rede projetada para os gestos dinâmicos, comparando os resultados com e sem a transferência. Por fim, apresenta-se um algoritmo capaz de reconhecer online os mesmos gestos dinâmicos da

etapa anterior, mas executados de forma sequencial, sem pausas ou segmentações. O caso estático obteve sucesso, atingindo uma maior acurácia do que em ambos casos individuais. Atingindo uma acurácia em apenas imagens como entrada de 79% na base de validação, ao esqueleto como entrada obteve acurácia de apenas 55% na base de validação e em imagens e esqueleto como entrada obteve 60% na base de validação.

Já em Amaral et al. (2017), é proposto um método para reconhecer gestos estáticos da mão representados por imagens de profundidade. Primeiramente, segmenta-se a mão do plano de fundo e após é realizado o cálculo da Transformada de Distância para treinamento da rede neural convolucional (CNN) que é usada para classificar poses da mão. Para testes em contexto prático, fazem uso de uma base de dados contendo 1400 imagens representando 14 classes de configurações de mão distintas representando sinais da Língua Brasileira de Sinais (LIBRAS). Alcançando uma taxa de reconhecimento de 96.42% em média.

de Almeida et al. (2018) se propuseram a investigar e obter a mais apta entre as técnicas de comparação de imagens por meio de histogramas estudadas. Para efeito de comparação, na solução ao problema relacionado a analise de vídeos com sinais dinâmicos em Libras, onde faz-se necessário o processamento das imagens que o compõe. Entretanto, os vídeos possuem grande quantidade de imagens similares, o que dificulta o processo de seleção manual, assim tem-se a necessidade de comparar computacionalmente a similaridade ou dissimilaridade entre as imagens. Isto é realizado através da comparação de histogramas, no entanto, existem diversas técnicas, como por exemplo: as distancias de Interseção, Chebyshev, Manhattan e Euclidiana. Após os testes de comparação, conclui-se que a distância de Chebyshev apresentou melhor desempenho e então foi considerada a mais apta entre as demais.

Em Escobedo Cárdenas (2020), é proposto uma metodologia para o reconhecimento de sinais contínuos da Língua Brasileira de Sinais (LIBRAS) utilizando como dados de entrada de um sinal as informações fornecidas pelo dispositivos *Kinect*. Assim, o método proposto utiliza janelas deslizantes para buscar segmentos candidatos de serem sinais dentro de um fluxo continuo de video. Equitativamente, propõe-se o uso de imagens dinâmicas para codificar as informações espaço-temporais fornecidas pelo *Kinect*, reduzindo a complexidade da arquitetura CNN proposta. Baseado no conceito de pares mínimos e um novo banco de dados de LIBRAS denominado LIBRAS-UFOP (possui tanto sinais isolados como sinais contínuos) e o comparando com as

literaturas existentes. Assim, o método poderá reconhecer de forma continua os gestos da língua, se baseando na geração de imagens dinâmicas fazendo uso das técnicas de *rank-pooling* e *Skeleton Optical Spectra*. Dado a complexidade do tema abordado, a pesquisa limitou-se ao estudo dos parâmetros primários de um sinal (movimento, localização e configuração de mãos). Os métodos propostos apresentaram uma diminuição no tempo de treinamento e teste ao ser comparados com métodos da literatura baseados em redes mais complexas (3DCNN, res3D, LSTM).

Simon et al. (2017) desenvolveram em seu artigo uma abordagem que se utiliza de um sistema multi-câmera para treinar detectores de baixa granularidade para pontos-chave que são propensos à oclusão, como as articulações de uma mão. Denominaram o este processo descrito como *bootstrapping multiview*: primeiro, um ponto-chave inicial detector é usado para produzir rótulos barulhentos em múltiplas visualizações a mão. As detecções ruidosas são então trianguladas em 3D usando geometria multi-vista ou marcados como *outliers*. Finalmente, as triangulações reprojetadas são usadas como novos dados de treinamento rotulados para melhorar o detector. Repete-se este processo, para garantir mais dados rotulados em cada iteração. Assim, deriva resultado que relaciona analiticamente o número mínimo de visualizações para atingir as taxas alvo de verdadeiros e falsos positivos para um determinado detector. Da mesma forma, o método é usado para treinar um detector de ponto chave manual para imagens únicas. O detector de ponto chave manual para imagens únicas. O detector de ponto-chave resultante é executado em tempo real em imagens RGB e tem uma precisão comparável aos métodos que usam sensores de profundidade. O detector de visão única, triangulado em várias visualizações, permite a captura de movimentos 3D sem marcadores com interações de objetos complexos.

Após a analise de diversas metodologias, foi possível compreender melhor o problema e como podemos resolve-los de diferentes maneiras. Sendo assim, a seguir é detalhada a metodologia escolhida na implementação deste trabalho.

3 Metodologia

Neste capitulo, é abordado o desenvolvimento de uma aplicação que solucione o problema de detecção de símbolos do alfabeto em LIBRAS. A proposta é realizar uma adaptação dos métodos já catalogados na literatura, algumas técnicas apresentadas por (Simon et al., 2017) e agregar a

capacidade de detecção de sinais em LIBRAS para a aplicação. Tal aplicação é composta por 3 módulos: (a) realização da verificação se o ponto está "acima"ou "abaixo"; (b) verificação se os dedos se encontram recolhidos ou esticados; (c) comparação da proximidade entre os pontos chaves detectados. O resultado final será um conjunto de dados vetoriais que corresponde as configurações do simbolo do alfabeto em LIBRAS.

A arquitetura da aplicação proposta é apresentada na Figura 3 . Inicialmente, é realizado a captura de cada *frame* do vídeo em análise, para que seja realizada a determinação dos pontos chaves da mão. Em seguida o processamento destes pontos.

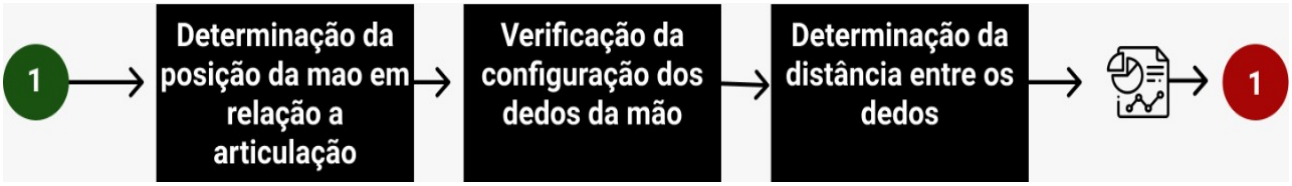

Figura 1: Arquitetura da aplicação. Fonte: Elaborado pelo autor

Estes pontos catalogados pela rede neural (CMU-Perceptual-Computing-Lab, 2021) serviram de base para a determinação das configurações de mão, sendo assim, os 22 pontos fornecidos pela rede, são catalogados em relação aos constituintes da mão, como mostra a Figura 3

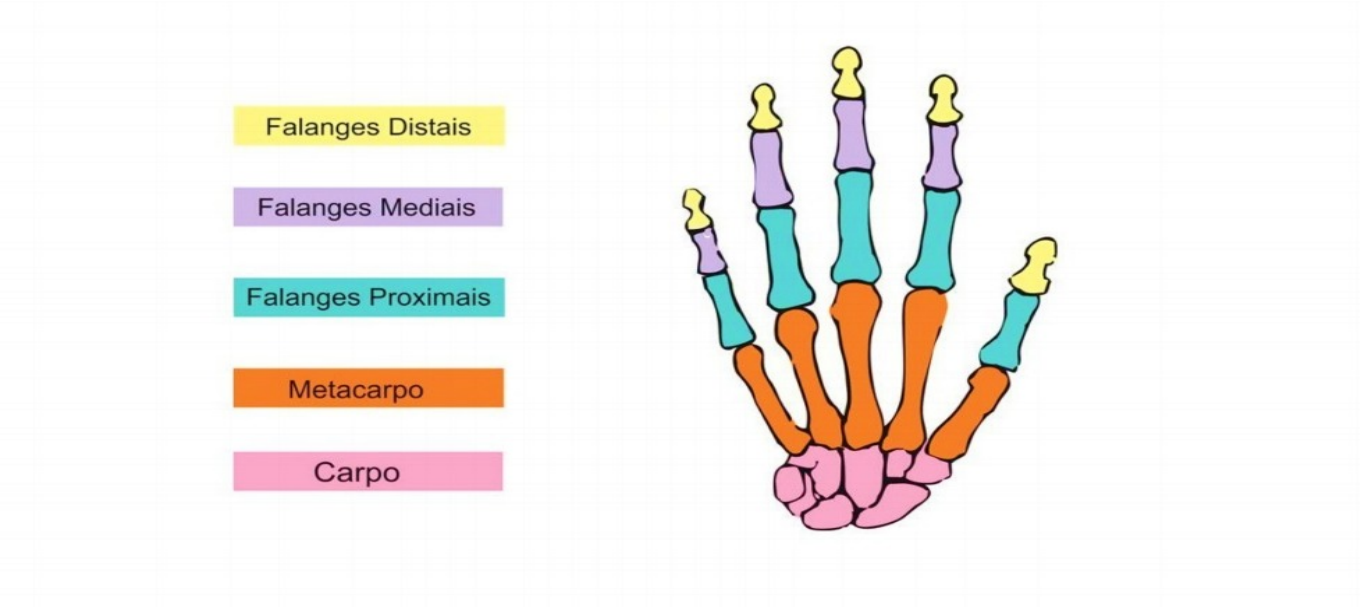

Figura 2: Constituição dos ossos da mão. Fonte: Diana (2017)

O próximo passo é obter os pontos de significância para as configurações do sinal, onde po-

demos determinar a qual letra do alfabeto se refere a configuração de mão. Assim, ao modulo de comparação das alturas dos pontos na vertical e horizontal determinaram se a mão está articulada para "baixo"ou para "cima", a partir do ponto compreendido no carpo podemos descartar as possibilidades de configurações de mãos que são expressas torcendo a articulação do punho, como a letra B, que apresenta o punho abaixo dos metacarpo e os falanges. A partir dos resultados podemos verificar a posição dos falanges em relação palma da mão, onde cada sinal apresenta uma configuração de mão única, podemos explorar essa característica, para assim definir se os falanges estão "dobrados"ou "esticados", em seguida verificamos a proximidade entre os dedos, para definir se estão na mesma altura, como a configuração de mão da letra N, que apresenta os dedos indicador e médio próximos e esticados, e os demais dobrados e também próximos. Com o vetor resultado dessa arquitetura, temos a descrição da configuração do sinal, assim torna simples a interpretação desse simbolo pelo alfabeto.

A aplicação proposta neste trabalho foi desenvolvida na linguagem *Python*, com auxilio da biblioteca *OpenCV* e o *deep learning framework Caffe* (*CAFFE*, 2012), que possui métodos e ferramentas para a realização de manipulações em imagens. A rede neural utilizada é a ganhadora do prêmio *COCO 2018 Keypoint Detection Task* (2020), qual apresentou o mapeamento dos pontos chaves da mão em diferentes ângulos e ações. A imagem do *frame* e submetida a rede neural, o resultado dos pontos mapeados passam pela arquitetura proposta e por fim obtém-se o sinal identificado.

4 Resultados

Neste capitulo, são apresentados os experimentos computacionais realizados, bem como os resultados obtidos no processo.

Para realização dos testes, foram submetido 3 vídeos a rede neural em diferentes ocasiões, em ambientes ruidosos e ambientes iluminado e focado a configuração do sinal. Analise de eficiência foi calculada através da porcentagem de classificação corretas, ou seja, o numero de sinais corretamente classificados dividido pelo numero total de sinais. A Figura 5 e 8 apresenta um exemplo de funcionamento do algoritmo de classificação e análise em condições adversas. A Figura 11 mostram a execução em condições ideais de iluminação e visualização da mão.

Figura 3: Letra A

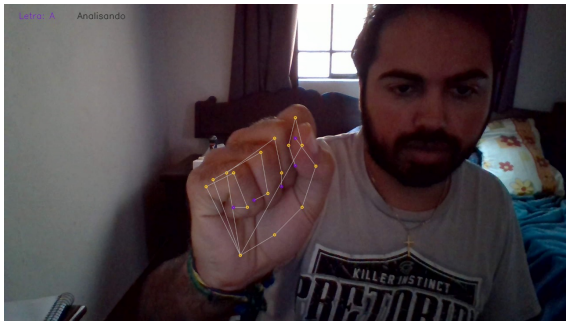

Figura 4: Letra A

Figura 5: Resultados da execução, no reconhecimento da Letra A em diferentes angulações da palma da mão

Figura 6: Letra B

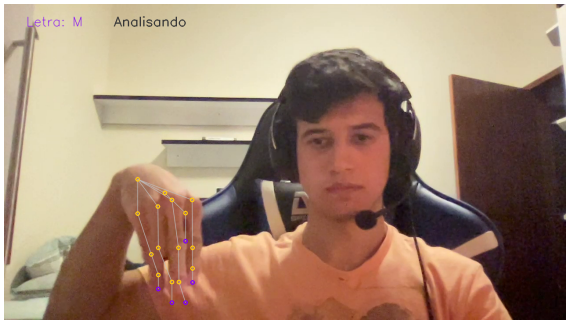

Figura 7: Letra M

Figura 8: Resultados da execução

Os resultados demonstram uma satisfatibilidade razoável na classificação das letras do alfabeto em ambientes poluídos, onde, a iluminação e qualidade de resolução impacta diretamente no resultado e no tempo de execução, possuindo um percentual de 73,68% de assertividade, 14 classificações corretas em 19 sinais totais. Ao vídeo que possuía condições melhores de visualização e caracterização dos sinais, obtivemos um percentual de 83.36%, em 16 classificações corretas em 20 sinais totais. Os principais aspectos identificados em cenários de classificação incorreta estão relacionados à qualidade de resolução da imagem, distancia da mão a câmera, ambiente de baixa luminosidade e velocidade de execução do sinal, dado que é analisado *frame* a *frame* do video.

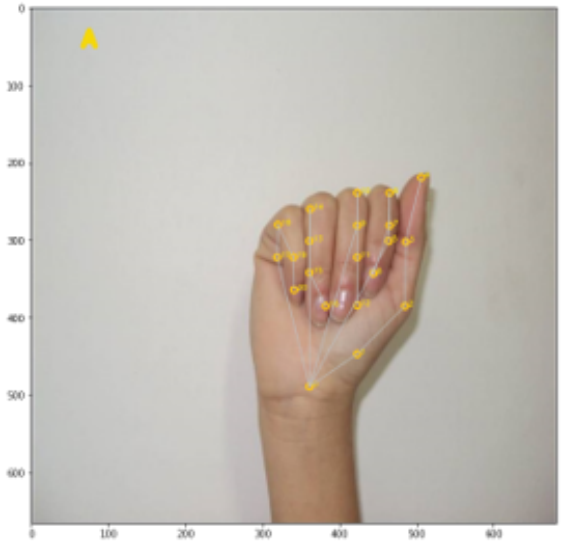

Figura 9: Letra A

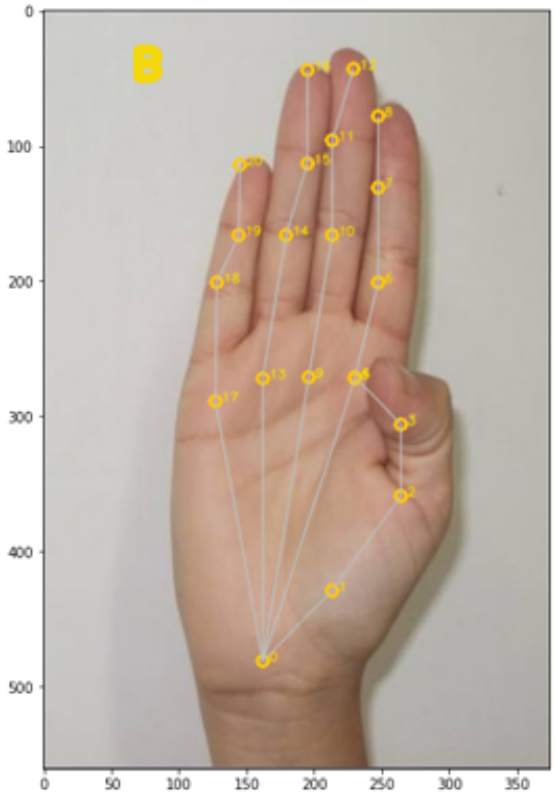

Figura 10: Letra B

Figura 11: Resultados da execução em condições ideais

5 Conclusão

Neste capítulo, são apresentados aspectos conclusivos sobre o trabalho desenvolvido. Este capítulo engloba as conclusões gerais referentes aos resultados da aplicação proposta, bem como as perspectivas de trabalhos futuros.

Por meio do presente trabalho podemos concluir que o uso de tecnologias na inclusão da comunidade surda apresenta ganhos significativos para sua inserção perante a sociedade. A ferramenta proposta e testada neste trabalho apresentou uma média eficácia, com nível médio de acerto, levando em considerações os erros derivados de iluminação, posicionamento ou até a não identificação do sinal. Sistemas baseados na visão computacional dependem de uma grande quantidade de métricas e detalhes para serem benéficos e realizarem suas tarefas de forma correta. Para trabalhos futuros, pretende-se implementar funções de análise e comparação do *frame* anterior ao próximo para solucionar as questões de classificação dos sinais com movimento em suas configurações.

Outro ponto seria identificar e corrigir os defeitos nas imagens com iluminação e posicionamento ruim e também a criação de uma interface gráfica para manipulação da aplicação.

Referências

Amaral, L., Lima, G., Vieira, T., & Vieira, T. (2017). Reconhecimento de gestos estáticos da mao usando a transformada de distância e aplicaçoes em libras. *Universidade Federal de Alagoas*.

Barros Junior, J. D. (2016). Tradução automática de línguas de sinais: do sinal para a escrita.

Caffe. (2012). Retrieved from https://caffe.berkeleyvision.org/

CIVIL, C. (2002). *Lei nº 10.436, de 24 de abril de 2002.* Retrieved from http://www.planalto.gov.br/ccivil_03/leis/2002/L10436.htm

CMU-Perceptual-Computing-Lab. (2021). *openpose/models/hand at master · cmu-perceptual-computing-lab/openpose.* Retrieved from https://github.com/CMU-Perceptual-Computing-Lab/openpose/tree/master/models/hand

Coco 2018 keypoint detection task. (2020). Retrieved from https://cocodataset.org/#keypoints-2018

de Almeida, M. A., Maia, C. X., Sá, A. P., & Mota, F. A. O. (2018). Comparação de distâncias de dissimilaridade em histogramas de imagens para auxílio ao reconhecimento de sinais de libras. *Anais dos Simpósios de Informática do IFNMG-Campus Januária*.

Diana, J. (2017, Jun). *Ossos da mão: função, nomes e localização | anatomia.* Toda Matéria. Retrieved from https://www.todamateria.com.br/ossos-da-mao/

do Nascimento Siqueira, R. (n.d.). Reconhecimento de símbolos de libras.

Escobedo Cárdenas, E. J. (2020). Desenvolvimento de uma abordagem para reconhecimento contínuo da língua brasileira de sinais utilizando imagens dinâmicas e técnicas de aprendizagem profunda.

Koroishi, G. O., & Silva, B. V. L. (2015). Reconhecimento de sinais da libras por visão computacional. *Mecatrone*, *1*(1).

Nandhini, A. S., Roopan, D. S., Shiyaam, S., & Yogesh, S. (2021, may). Sign language recognition using convolutional neural network. *Journal of Physics: Conference Series*, *1916*(1), 012091.

Retrieved from https://doi.org/10.1088/1742-6596/1916/1/012091 doi: 10.1088/1742-6596/1916/1/012091

Ribeiro, D. R. S., & Rodrigues, M. T. A. N. (2017). Tradução de libras por imagem.

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

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1145–1153).

Sridevi, P., Islam, T., Debnath, U., Nazia, N. A., Chakraborty, R., & Shahnaz, C. (2018). Sign language recognition for speech and hearing impaired by image processing in matlab. In *2018 ieee region 10 humanitarian technology conference (r10-htc)* (pp. 1–4).

Voigt, J. F., et al. (2018). Aprendizagem profunda para reconhecimento de gestos da mão usando imagens e esqueletos com aplicações em libras.

WHO, W. H. O. (2021, Apr). *Deafness and hearing loss.* Author. Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss