# Desenvolvimento de uma Base de Dados de Sinais de Libras para Aprendizado de Máquina: Estudo de Caso com CNN 3D ⋆

**Giulia Z. de Castro** [*] **Rúbia R. Guerra** [***]
**Moises M. de Assis** [***] **Tamires M. Rezende** [*]
**Gabriela T. B. de Almeida** [***] **Sílvia G. M. Almeida** [**]
**Cristiano L. de Castro** [***] **Frederico G. Guimarães** [***]

[*] *Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil, (e-mail: giuliaz@ufmg.br, tamiresrezende@ufmg.br).*
[**] *Instituto Federal de Minas Gerais - Campus Ouro Preto, Ouro Preto, Minas Gerais, Brasil, (e-mail: silvia.almeida@ifmg.edu.br)*
[***] *Machine Intelligence and Data Science (MINDS) Laboratory, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil, (e-mail: rubia-rg@ufmg.br, moisesmendes@ufmg.br, crislcastro@ufmg.br, fredericoguimaraes@ufmg.br)*

**Abstract:** A recurrent problem in Brazilian Sign Language (Libras) recognition is the absence of a robust dataset that allows the validation of different methodologies. This work presents a new dataset for Libras and its respective recording procedure. The first available version contains 20 signs, recorded 5 times by 10 different signers, making up 1000 recordings. A study case with the new data utilizing a 3D Convolutional Neural Network for sign recognition is also presented, employing summarization and data augmentation techniques. The network implemented achieved an average accuracy of 72,6%.

**Resumo**: Um dos problemas enfrentados nos trabalhos de reconhecimento de sinais de Libras é a ausência de uma base de dados robusta, que permita a validação de diferentes metodologias. Dessa forma, este trabalho apresenta uma nova base de dados da Língua Brasileira de Sinais e seu protocolo de gravação. A primeira versão da base contém 20 sinais, gravados 5 vezes, por cada um de 10 sinalizadores, totalizando 1000 gravações. É apresentado também um estudo de caso da aplicação de uma Rede Neural Convolucional 3D para a tarefa de classificação dos sinais, utilizando técnicas de sumarização e data augmentation. A rede implementada alcançou um resultado médio de 72,6% de acerto.

*Keywords:* Libras, sign recognition, dataset, recording protocol, 3DCNN.

*Palavras-chaves:* Libras, reconhecimento de sinais, base de dados, protocolo de gravação, CNN 3D.

## 1. INTRODUÇÃO

A língua de sinais é uma forma de comunicação visual-motora utilizada pela comunidade surda. Assim como as línguas orais, as línguas de sinais são únicas para cada cultura, apresentando estruturas gramaticais próprias. No Brasil, ela é chamada de Língua Brasileira de Sinais (Libras) e é a segunda língua oficial do país desde a publicação da Lei nº 10.436 em 2002.

O problema de identificação automática de sinais [1] das línguas de sinais pode ser considerado como uma aplicação específica do problema mais geral de reconhecimento de gestos, visto que as expressões não-manuais não são necessárias em grande parte dos sinais (Capovilla, 2017). Geralmente, o *framework* padrão de reconhecimento de gestos consiste na extração de atributos espaço-temporais de quadros de vídeo, seguida pela modelagem da dinâmica intra-quadros por meio de classificadores, como *Support Vector Machines* (SVM) (Rautaray e Agrawal, 2015), *Hidden Markov Model* (HMM) (Kumar et al., 2017) e Redes Neurais Artificiais (John et al., 2016).

Um dos grandes desafios em reconhecimento de Libras é a disponibilidade de bases de dados representativas que possibilitem a validação de novas metodologias. Até o momento, observa-se que a maioria dos trabalhos no assunto criaram suas próprias bases (Filho et al., 2017; Escobedo-Cardenas e Camara-Chavez, 2015; Almeida et al., 2014), o que dificulta a comparação das técnicas utilizadas na tarefa de reconhecimento. Uma das poucas bases disponí-

⋆ O presente trabalho foi realizado com o apoio financeiro da CAPES e CNPq.

[1] Menor unidade da língua de sinais, composta pelo movimento das mãos, corpo e expressões faciais.

veis publicamente é a de Amaral et al. (2019), contendo informações de profundidade (Figura 1) para 10 sinais.

Figura 1. Exemplo de imagem disponibilizada por Amaral et al. (2019)

Em relação ao registro de sinais de Libras, estudos anteriores utilizam sensores multimodais, como, por exemplo, o sensor RGB-D (Almeida et al., 2014; Escobedo-Cardenas e Camara-Chavez, 2015; Filho et al., 2017). Outras propostas abordam o problema por meio de vestimentas especiais, como luvas e sensores do tipo *wearable* (Kawamoto et al., 2018). Estes têm como objetivo contornar limitações causadas pela variação de iluminação, facilitar a segmentação das regiões de interesse para a detecção do sinal ou extrair atributos relativos à trajetória das mãos. O presente estudo busca contribuir com uma base de sinais padronizada, disponibilizada gratuitamente e em um cenário que permite adaptações por parte do usuário. Ela contém, inicialmente, 20 sinais gravados sistematicamente 5 vezes por 10 sinalizadores distintos, totalizando 1000 amostras e encontra-se disponível em Almeida et al. (2019).

Procurando estabelecer um patamar de comparação para trabalhos futuros aplicados a esta nova base de dados, decidiu-se aplicar um modelo de aprendizado baseado em redes neurais como um estudo de caso. Trabalhos propondo a utilização da CNN 2D em tarefas de classificação envolvendo imagens vêm ganhando atenção (Rawat e Wang, 2017), especialmente após os resultados obtidos por Krizhevsky et al. (2012) no dataset *ImageNet*. Abordagens de CNN 2D em vídeo, contudo, geralmente são aplicadas a cada quadro individual e não consideram a informação de movimento contida em múltiplos quadros contíguos. Uma forma de efetivamente incluir a correlação temporal interquadro no modelo é a utilização de convoluções 3D nas camadas convolucionais da CNN (Tran et al., 2014). Sendo assim, inspirado nas arquiteturas de CNN 3D propostas em Molchanov et al. (2015) e Zhang et al. (2017), este estudo apresenta uma abordagem para reconhecimento de sinais de Libras incorporando a implementação de uma CNN 3D ao *framework* padrão de reconhecimento de gestos.

O artigo está estruturado da seguinte forma: a seção 2 apresenta o protocolo de criação da base de sinais de Libras. Posteriormente a seção 3 descreve a arquitetura da CNN utilizada para classificação dos sinais. Por fim, os resultados são apresentados e discutidos nas seções 4 e 5, respectivamente.

## 2. BASE DE DADOS DE SINAIS DE LIBRAS

A construção ou a escolha de uma base de dados é uma etapa fundamental em qualquer problema de reconhecimento de padrões. Quando se trata de sinais de Libras, este assunto se torna desafiador, pois a maioria dos trabalhos criam as próprias bases de dados para validar suas metodologias. Realizando uma breve pesquisa bibliográfica, verificou-se alguns desses trabalhos. A Tabela 1 sintetiza as características de cada um.

Tabela 1. Bases de dados de língua de sinais.

| Base | Ano | Reconhecimento de | Nº de amostras |
|---|---|---|---|
| Athitsos et al. (2008) | 2008 | Sinais da ASL* | 3800 |
| Li (2017) | 2012 | Gestos manuais de ASL | 336 |
| Conly et al. (2013) | 2013 | Sinais da ASL | 1113 |
| Almeida (2014) | 2014 | Sinais da Libras | 170 |
| Rezende (2016) | 2016 | Expressões Faciais da Libras | 100 |

*Língua Americana de Sinais.

Tendo em vista que a Libras tem mais de 10 mil verbetes, percebe-se que há uma carência na área quando se trata de uma base com sinais de Libras padronizados e que sejam disponibilizados em um formato que permita a validação de sistemas de classificação computacional de forma robusta. Dessa forma, o presente trabalho desenvolveu um protocolo de gravação com base nos estudos de Ruffieux et al. (2014), Almeida (2014) e Rezende (2016). O protocolo aborda a escolha dos sinais e quem os executará, os sensores e *softwares* utilizados para aquisição dos vídeos, o cenário das gravações e a estrutura dos dados disponibilizados.

### *2.1 Seleção dos sinais*

Com o auxílio de uma especialista em Libras foram selecionados 20 sinais da língua com base na diversidade das características dos parâmetros fonológicos da Libras. Estes parâmetros referem-se às unidades formacionais de um sinal e são caracterizados por: configuração da mão [2], ponto de articulação [3], movimento das mãos, orientação da palma da mão e expressões não-manuais. Partindo desses critérios, os sinais selecionados foram: acontecer, aluno, amarelo, América, aproveitar, bala, banco, banheiro, barulho, cinco, conhecer, espelho, esquina, filho, maçã, medo, ruim, sapo, vacina e vontade. Cada um deles foi gravado 5 vezes por cada um dos 10 sinalizadores [4].

### *2.2 Sinalizadores*

Dentre os sinalizadores há homens e mulheres, surdos e ouvintes, sem distinção de vestimenta e raça, e com conhecimento variando de básico a avançado na Libras. Sugeriu-se aos sinalizadores que permanecessem olhando para a câmera em posição de descanso antes e após a execução do sinal, para marcar o início e o fim de cada gravação. Além disso, em todas as gravações a posição do sinalizador é fixa, em pé no centro do vídeo, e ele inicia e finaliza o sinal com as mãos sobre as pernas.

### *2.3 Cenário de gravação*

No estúdio de gravação, o sensor ficou em posição fixa, gravando o movimento corporal superior, a expressão facial e o movimento manual. O arranjo do cenário foi disposto como apresentado na Figura 2. As gravações dos sinais ocorreram em um estúdio com boa iluminação e com plano

[2] Forma assumida pela mão na articulação do sinal.
[3] Área do corpo em que o sinal é articulado.
[4] Quem executa o sinal.

de fundo constante feito de tecido Chroma Key [5] . Este plano permite ao usuário da base remover ou adicionar diferentes fundos, sendo uma possível técnica para testar o desempenho de algoritmos de reconhecimento que usam padrões visuais.

Figura 2. Cenário de gravação.

### *2.4 Estrutura dos dados*

Com o intuito de disponibilizar os vídeos dos sinais no formato MP4, utilizou-se uma câmera digital profissional [6] , que possui uma taxa de gravação de 30 fps [7] . Cada quadro dos vídeos possui 1920 × 1080 *pixels*. A base de dados inicialmente proposta possui 1000 amostras [8] . Entretanto, houve o comprometimento de 5 gravações de um sinal [9] , que foram descartadas, totalizando 995 amostras.

## 3. ARQUITETURA E PLANEJAMENTO EXPERIMENTAL

Neste trabalho foi realizado um pré-processamento dos dados e classificação por meio de uma CNN 3D. O pré-processamento consistiu em reduzir informações redundantes nos vídeos de cada gravação e aplicar uma ferramenta para aumentar o número de amostras. A primeira ação padroniza o número de imagens que representam cada gravação e a segunda fornece maior insumo de dados à CNN, que precisa de um grande volume para seu treinamento. A classificação dos sinais da base de dados é realizada pelos procedimentos esquematizados na Figura 3 .

### *3.1 Pré-processamento*

Os vídeos da base de dados descrita foram submetidos a dois processos antes de serem passados à CNN 3D: sumarização e *data augmentation*.

A sumarização de vídeo foi aplicada neste trabalho com o objetivo de eliminar informações redundantes e uniformizar as gravações para terem o mesmo tamanho. No caso dos sinais da base utilizada, o número de quadros de cada sinal varia devido às diferentes velocidades de gravação dos sinalizadores, o que gera, dentre outros, quadros sequenciais muito semelhantes. Além disso, a sumarização escolhe um número pré-definido de quadros por vídeo, o que os torna padronizados e garante que os vetores de características da rede neural terão o mesmo tamanho para todas as amostras. A técnica de sumarização utilizada baseia-se no Problema da Diversidade Máxima (PDM) (Kuo et al., 1993), um problema de otimização que busca encontrar um conjunto de elementos para os quais a diversidade entre eles seja máxima. A solução adotada para o PDM foi a de Almeida et al. (2015), que resolve o problema por meio de uma estratégia evolutiva. Para este estudo escolheu-se utilizar uma sumarização de 12 quadros e, após sumarizar, foram excluídos o primeiro e o último quadros, pois verificou-se que neles a pessoa estava parada, isto é, caracterizam o início e o fim das gravações. Portanto, ao final da sumarização e do corte, obteve-se 10 quadros por vídeo. Este valor teve como referência os trabalhos de Rezende (2016) e Almeida (2014), utilizando o dobro de quadros pra obter mais detalhes da execução dos sinais.

*Data augmentation* refere-se a estratégias utilizadas para aumentar o volume de dados. Elas têm sido aplicadas no treinamento de CNNs por diversos autores para evitar o efeito de *overffiting* [10] , pois tornam o conjunto de dados mais diverso (Krizhevsky et al., 2012; Pigou et al., 2014; Simonyan e Zisserman, 2014; Molchanov et al., 2015). Uma forma de aumento dos dados é a espacial, que inclui as operações de translação, espelhamento horizontal e redimensionamento das imagens (Krizhevsky et al., 2012; Simonyan e Zisserman, 2014). Há também o aumento temporal dos dados, que é geralmente aplicado a vídeos e envolve translação temporal, escalonamento da duração da sequência e deformação no domínio do tempo (Pigou et al., 2014; Molchanov et al., 2015).

Neste trabalho, foram aplicadas tanto estratégias temporais quanto espaciais para aumentar os dados de treinamento. Inicialmente tem-se 995 vídeos da base, dos quais 746 são utilizados para o treinamento e 249 para teste, numa proporção de 75%–25% por classe.

Aplica-se, primeiramente, um deslocamento temporal nos 10 quadros obtidos pela sumarização. Isso é feito somando-se um valor aleatório entre 1 e 4 à posição de cada um desses dez quadros de cada vídeo, dobrando o número de vídeos de treino para 1492. Realiza-se também o espelhamento horizontal e *zoom* dos vídeos originais, resultando em um conjunto com 2984 amostras. Portanto, os dados de treinamento da CNN 3D são esses 2984 vídeos e os de teste são os 249 vídeos citados anteriormente. Todos os quadros foram cortados de 1920 × 1080 *pixels* para 1080 × 1080 *pixels* e redimensionados para 224 × 224 *pixels*, com o intuito de remover uma parte do *background* das imagens e reduzir a quantidade de informação a ser processada. Por fim, os dados de treinamento foram subtraídos da média do conjunto, com o objetivo de normalizar os dados e aumentar a velocidade de aprendizagem.

### *3.2 Arquitetura*

Neste estudo de caso, as CNNs 3D foram utilizadas com o objetivo de fornecer uma linha de base para estudos futuros em reconhecimento de sinais de Libras. Essa escolha foi motivada pelos resultados obtidos em estudos anteriores, indicando a habilidade das redes convolucionais tridimensionais em captar características espaço-temporais (Tran et al., 2014).

[5] Técnica utilizada para posicionar uma imagem sobre outra através do anulamento de uma cor sólida, como o verde claro.
[6] Canon EOS Rebel t5i.
[7] Imagens/quadros por segundo (*frames per second*).
[8] 20 sinais × 5 gravações × 10 sinalizadores.
[9] Sinal: Filho, Sinalizador: 4.
[10] Sobreajuste do modelo aos dados.

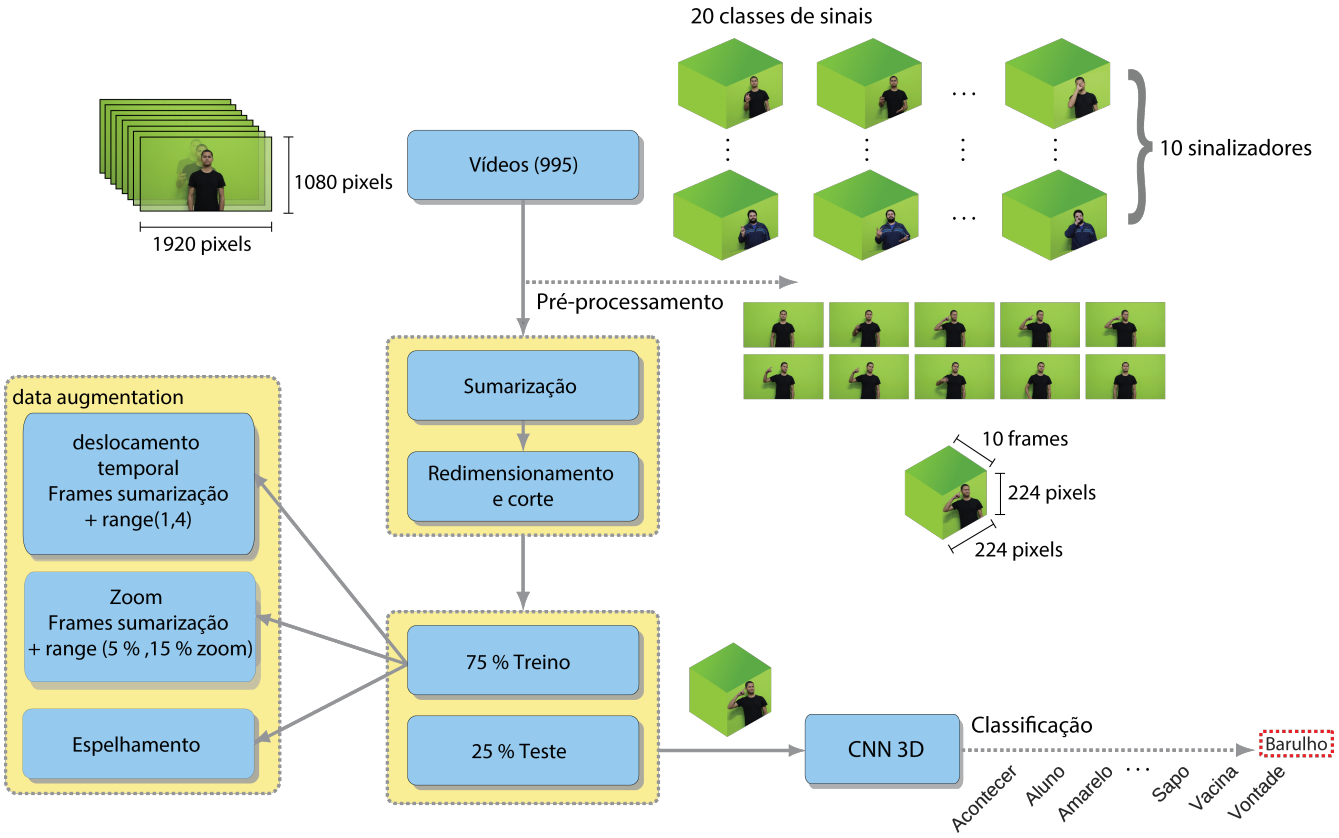


Figura 3. Fluxograma.

A CNN recebe como entrada um volume composto pelos 10 *frames* retornados do pré-processamento, em RGB (dimensões $10 \times 224 \times 224 \times 3$), resultando em um volume de saída que permite que a informação temporal do sinal seja capturada.

A arquitetura da rede consiste em 4 camadas convolucionais, sendo cada uma seguida por uma função de ativação ReLU e uma camada de *max pooling*. As camadas convolucionais possuem 4, 8, 16, e 32 filtros em profundidade, respectivamente. Ao final da rede são utilizadas 2 camadas totalmente conectadas e uma função de ativação *softmax*, que funciona como um classificador. A saída da função de ativação consiste em um vetor contendo a probabilidade de cada um dos 20 sinais corresponderem a uma determinada classe.

Os filtros de convolução utilizados possuem dimensões ($3 \times 3 \times 3$), cujo desempenho em tarefas de análise de vídeos foi avaliado por Tran et al. (2014). As camadas de *max pooling*, que também realizam operações em profundidade, possuem *kernels* ($1 \times 2 \times 2$) e ($2 \times 2 \times 2$), na primeira e demais camadas, respectivamente. A arquitetura da rede é sintetizada na Tabela 2.

Para o treinamento da rede, foi definida uma taxa de aprendizado inicial de 0,001, que foi ajustada conforme o desempenho da mesma. Assim, caso a perda de validação não apresentasse uma melhoria após 3 épocas consecutivas, a taxa foi reduzida por um fator de 10. Foram utilizados lotes de tamanho 128 e o treinamento foi interrompido após 50 épocas. Para evitar *overfitting*, foi empregado o

Tabela 2. Arquitetura.

| Descrição | Saída |
|---|---|
| Volume de entrada | $10 \times 224 \times 224 \times 3$ |
| Conv3D | $8 \times 222 \times 222 \times 4$ |
| MaxPool3D | $8 \times 111 \times 111 \times 4$ |
| Conv3D | $8 \times 111 \times 111 \times 8$ |
| MaxPool3D | $4 \times 55 \times 55 \times 8$ |
| Conv3D | $4 \times 55 \times 55 \times 16$ |
| MaxPool3D | $2 \times 27 \times 27 \times 16$ |
| Conv3D | $2 \times 27 \times 27 \times 32$ |
| MaxPool3D | $1 \times 13 \times 13 \times 32$ |
| Flatten | 5408 |
| Fully Connected | 128 |
| Dropout | 128 |
| Fully Connected | 20 |

*Dropout*, uma técnica de regularização que elimina alguns neurônios ocultos da rede temporariamente.

O experimento foi realizado 10 vezes, sendo obtidas as métricas de desempenho médias.

## 4. RESULTADOS E DISCUSSÕES

A figura 4 apresenta a matriz de confusão obtida após 10 iterações do algoritmo de classificação utilizando a CNN 3D. Foram 2984 amostras de treinamento e 249 de teste, alcançando um resultado médio de 72,6% de acerto. Essas iterações garantem uma aleatoriedade no conjunto de treino e de teste, fazendo com que ora a amostra participe do grupo de treinamento, ora do grupo de teste.

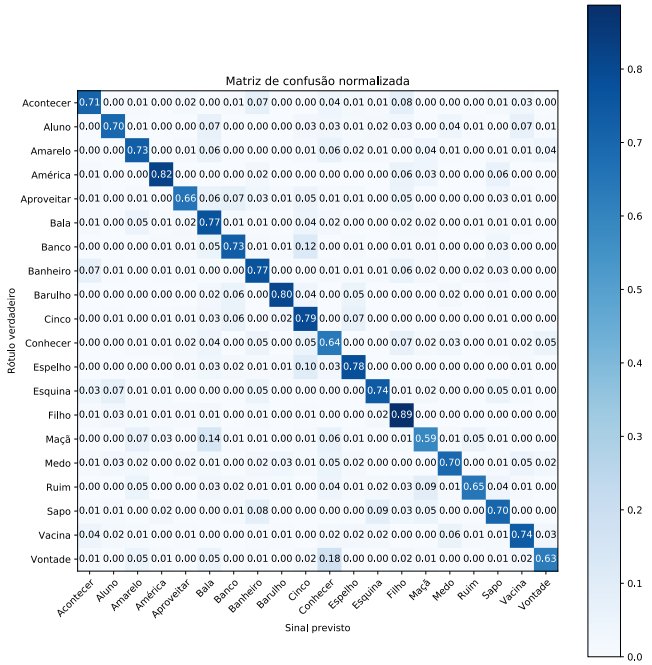


Figura 4. Matriz de confusão normalizada obtida pela média de 10 iterações.

Entre os sinais que apresentaram a menor acurácia, estão "vontade" e "maçã". Observou-se que, em média, 18 % das observações referentes ao sinal "vontade" foram classificadas erroneamente como "conhecer". Esses sinais são representados nas figuras 5 e 6, respectivamente. Percebe-se que, em ambos os casos, o ponto de articulação é o mesmo, isto é, na região em torno do queixo. Assim, sugere-se que a rede neural convolucional foi capaz de aprender as representações relativas ao movimento, mas ainda seria necessário captar outros parâmetros, como a configuração das mãos, para conseguir distinguir entre esses dois sinais. O mesmo acontece com o sinal "maçã" (Figura 7), nas quais 14% das observações foram confundidas com o sinal "bala" (Figura 8). Contudo, conforme a Figura 9, considerando-se os três melhores resultados em cada classe, observa-se que 88,4% das observações do sinal "vontade" e 86,4% do sinal "maçã" apresentaram alta probabilidade de serem corretamente identificados em relação aos seus respectivos rótulos.

Figura 5. Sinal: Vontade.

Figura 6. Sinal: Conhecer.

Vale ressaltar, ainda, que outras ferramentas de pré-processamento podem melhorar o resultado obtido, bem como outra arquitetura de CNN, ou a utilização de uma LSTM. Aplicando a topologia proposta, obteve-se resultados iniciais (Figura 9) com a base de dados a ser disponibilizada para a comunidade científica, com o intuito de ampliar aplicações na área de Libras e abrir espaço para a criação de um tradutor automático da língua.

Figura 7. Sinal: Maçã.

Figura 8. Sinal: Bala.

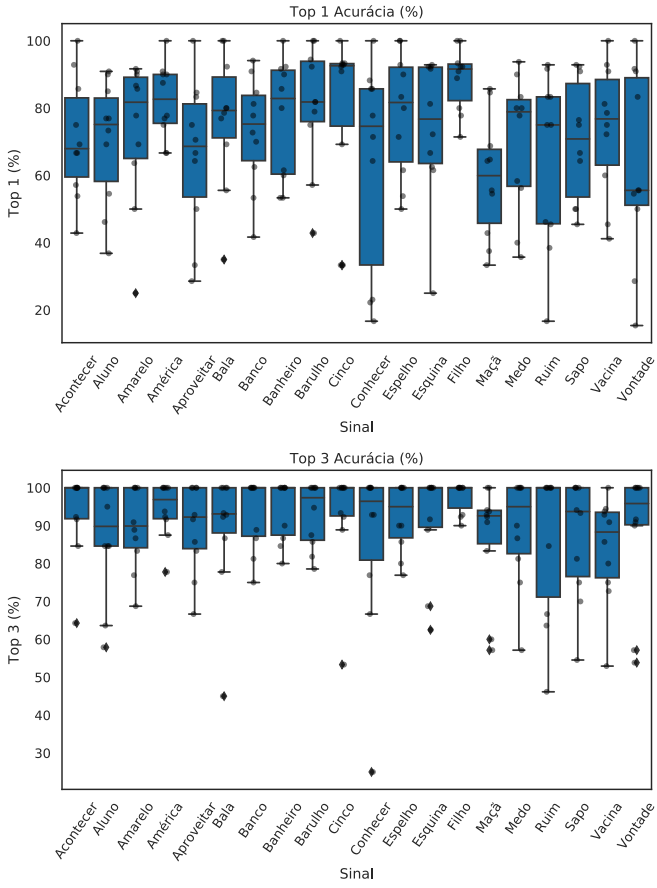


Figura 9. Top 1 e Top 3.

## 5. CONCLUSÕES

Este trabalho apresentou um novo conjunto de dados de vídeos para aplicações em reconhecimento da Língua Bra-

sileira de Sinais. Utilizou-se uma CNN 3D para capturar informações inter-quadros e estabelecer um patamar de comparação para trabalhos futuros que apliquem seus métodos na base disponibilizada. Acredita-se que a base de dados apresentada possa contribuir de forma expressiva para o desenvolvimento de novas aplicações em reconhecimento de sinais de Libras e temas correlatos.

## REFERÊNCIAS


Almeida, S.G.M. (2014). *Extração de Características em Reconhecimento de Parâmetros Fonológicos da Língua Brasileira de Sinais utilizando Sensores RGB-D*. Ph.D. thesis, Universidade Federal de Minas Gerais, Programa de Pós Graduação em Engenharia Elétrica, Belo Horizonte, Minas Gerais, Brasil.

Almeida, S.G.M., Guimarães, F.G., e Ramírez, J.A. (2015). Um método para sumarização de vídeos baseado no problema da diversidade máxima e em algoritmos evolucionários. In *XII Simpósio Brasileiro de Automação Inteligente (SBAI)*, 1298 – 1303. Natal, Rio Grande do Norte, Brasil.

Almeida, S.G.M., Guimarães, F.G., Rezende, T.M., Almeida, G.T.B., e Toffolo, A.C.R. (2019). Libras-20 dataset. `https://doi.org/10.5281/zenodo.2667329`. doi:10.5281/zenodo.2667329.

Almeida, S.G.M., Guimarães, F.G., e Ramírez, J.A. (2014). Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors. *Expert Systems with Applications*, 41(16), 7259–7271.

Amaral, L., Júnior, G.L.N., Vieira, T., e Vieira, T. (2019). Evaluating deep models for dynamic brazilian sign language recognition. In R. Vera-Rodriguez, J. Fierrez, e A. Morales (eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 930–937. Springer International Publishing, Cham.

Athitsos, V., Neidle, C., e Sclaroff, S. (2008). American sign language lexicon video dataset (asllvd). URL `http://vlm1.uta.edu/~athitsos/asl_lexicon/`.

Capovilla, F.C. (2017). *Dicionário da Língua de Sinais do Brasil. A Libras em Suas Mãos - 3 Volumes*. Edusp.

Conly, C., Doliotis, P., Jangyodsuk, P., Alonzo, R., e Athitsos, V. (2013). Toward a 3d body part detection video dataset and hand tracking benchmark. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '13, 2:1–2:6. ACM, New York, NY, USA. doi:10.1145/2504335.2504337. URL `http://doi.acm.org/10.1145/2504335.2504337`.

Escobedo-Cardenas, E. e Camara-Chavez, G. (2015). A robust gesture recognition using hand local data and skeleton trajectory. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE. doi:10.1109/icip.2015.7350998. URL `https://doi.org/10.1109/icip.2015.7350998`.

Filho, C.F.F.C., de Souza, R.S., dos Santos, J.R., dos Santos, B.L., e Costa, M.G.F. (2017). A fully automatic method for recognizing hand configurations of brazilian sign language. *Research on Biomedical Engineering*, 33(1), 78–89. doi:10.1590/2446-4740.03816. URL `https://doi.org/10.1590/2446-4740.03816`.

John, V., Boyali, A., Mita, S., Imanishi, M., e Sanma, N. (2016). Deep learning-based fast hand gesture recognition using representative frames. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE. doi:10.1109/dicta.2016.7797030. URL `https://doi.org/10.1109/dicta.2016.7797030`.

Kawamoto, A., Bertolini, D., e Barreto, M. (2018). A dataset for electromyography-based dactylology recognition. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. doi:10.1109/smc.2018.00408. URL `https://doi.org/10.1109/smc.2018.00408`.

Krizhevsky, A., Sutskever, I., e Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Kumar, P., Gauba, H., Roy, P.P., e Dogra, D.P. (2017). Coupled hmm-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, 86, 1–8.

Kuo, C.C., Glover, F., e Dhir, K.S. (1993). Analyzing and modeling the maximum diversity problem by zero-one programing. *Decision Sciences*, 24(6), 1171–1185.

Li, W. (2017). Webpage of dr wanqing li. URL `http://www.uow.edu.au/~wanqing/#MSRAction3DDatasets`.

Molchanov, P., Gupta, S., Kim, K., e Kautz, J. (2015). Hand gesture recognition with 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1–7.

Pigou, L., Dieleman, S., Kindermans, P.J., e Schrauwen, B. (2014). Sign language recognition using convolutional neural networks. In *European Conference on Computer Vision*, 572–578. Springer.

Rautaray, S.S. e Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1), 1–54.

Rawat, W. e Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9), 2352–2449.

Rezende, T.M. (2016). *Aplicação de Técnicas de Inteligência Computacional para Análise da Expressão Facial em Reconhecimento de Sinais de Libras*. Master's thesis, Universidade Federal de Minas Gerais, Programa de Pós Graduação em Engenharia Elétrica, Belo Horizonte, Minas Gerais, Brasil.

Ruffieux, S., Lalanne, D., Mugellini, E., e Khaled, O.A. (2014). A survey of datasets for human gesture recognition. In *International Conference on Human-Computer Interaction*, 337–348. Springer.

Simonyan, K. e Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, 568–576.

Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., e Paluri, M. (2014). C3D: generic features for video analysis. *CoRR*, abs/1412.0767. URL `http://arxiv.org/abs/1412.0767`.

Zhang, L., Zhu, G., Shen, P., Song, J., Afaq Shah, S., e Bennamoun, M. (2017). Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 3120–3128.