Um sistema de reconhecimento de sinais em Libras usando CNN e LSTM

Gabriel dos S. L. Stefano*, Wesley L. Passos†, Jonathan N. Gois*, Gabriel M. Araujo*, Amaro A. de Lima*.

***Abstract*— In Brazil, there are around 10 million hard of hearing and deaf people. However, the majority of Brazilians are not fluent in the Brazilian Sign Language (Libras). Many members of the hearing impaired and deaf community have communication issues in everyday life situations. Technological solutions can aid in mitigating this problem. This work proposes a semi-supervised method to identify and classify signals in Libras from Youtube videos. The routine starts by segmenting the videos through a measure of movement intensity. We catalog the video segments according to the corresponding subtitles, which we extract by employing Optical Character Recognition (OCR) if embedded. It composes an ad-hoc dataset that we use to train a Libras recognition system. A Convolutional Neural Network (CNN) performs feature extraction frame-by-frame, and a Recurrent Neural Network (RNN) models the time correlation between the features, thus classifying the signal. The proposed method can achieve accuracy up to 61.6% in the ad-hoc dataset used in this work.**

***Keywords*— Libras, Computer Vision, Convolutional Neural Network, Recurrent Neural Network.**

I. Introdução

Dados do censo demográfico realizado no Brasil em 2010 pelo Instituto Brasileiro de Geografia e Estatística (IBGE), apontam que 9,7 milhões de brasileiros (5,1% da população brasileira na época) possuíam algum grau de deficiência auditiva [1]. Destes, 2,1 milhões eram considerados surdos.

Um dos desafios relacionados aos graus mais severos de surdez é a dificuldade de estabelecimento da comunicação plena entre o surdo e a sociedade. Estudos da *World Federation of the Deaf (WFD)* indicam que cerca de 80% dos surdos do mundo possuem problemas de compreensão nas línguas escritas de seus respectivos países [2]. Para grande parte dessas pessoas, a língua de sinais é a utilizada para comunicação. No caso do Brasil, a maior parte da comunidade surda utiliza a Língua Brasileira de Sinais (Libras).

A Libras não é uma simples transcrição ou representação gestual da Língua Portuguesa. Ela é considerada uma língua natural completamente desenvolvida, já que atende a todos os critérios linguísticos necessários. Isso significa que ela possui estruturas gramaticais próprias e, além disso, da mesma forma que em estruturas orais de comunicação, possui regionalismos, gírias e particularidades intrínsecas de sua natureza visuoespacial [3]. Nas línguas de sinais, os campos visuais e espaciais são imprescindíveis, já que os sinais são produzidos a partir das combinações da forma, expressão facial, movimento das mãos e ponto no corpo onde esses sinais são feitos [4]. Por esse motivo, técnicas relacionadas à visão computacional estão entre os estudos que ganharam relevância nos últimos anos, apresentando resultados interessantes. Além disso, sistemas de reconhecimento de linguagem de sinais baseados em visão computacional tendem a ser mais baratos, por não requerer dispositivos dedicados como braceletes ou luvas, e menos intrusivos.

*Centro Federal de Educação Tecnológica, RJ, Brasil. †PEE/COPPE/UFRJ, Rio de Janeiro, Brasil. E-mails: gabriel.stefano@aluno.cefet-rj.br, wesley.passos@smt.ufrj.br, {jonhatan.gois, gabriel.araujo, amaro.lima}@cefet-rj.br

A. Trabalhos relacionados

Gameiro et al. [5] propôs uma evolução natural do trabalho em [6], que usou apenas uma parte o conjunto de dado agora conhecido como CEFET/RJ-Libras. O trabalho descreve detalhadamente o conjunto de dados, composto de 24 sinais dinâmicos de libras, executados por 20 indivíduos, totalizando 547 sequências de vídeo. Em seguida avalia diferentes tipos de extratores de características baseados no vídeo residual (composto pela diferença entre quadros), bem como os classificadores *K-Nearest Neighbor* (*K*-NN) e *Random Forest* (RF). O método obteve acurácia de 65,81%, o que pode ser considerado promissor, dado o baixo custo computacional do método.

Um trabalho recente descrito em [7] desenvolveu um conjunto de dados baseado no conceito de pares mínimos para reconhecimento de gestos isolados, chamado de LIBRAS-UFOP. O conjunto de dados contém 56 sinais divididos em 4 categorias. Os sinais foram executados por 5 indivíduos com aproximadamente 10,86 repetições por interprete, totalizando 3.040 sequências de dados. Essas sequências, adquiridas por um Microsoft Kinect V1, são compostas vídeos RGB-D e esqueleto. O sistema de reconhecimento proposto por eles utiliza as informações RGB-D, esqueleto e face em 3 Redes Neurais Profundas (DNN) em paralelo. As entradas das redes são características extraídas dos vídeos empregando *Rank Pooling* e as saídas são fundidas por média para gerar uma classificação única. Essa abordagem atinge 74,25% de taxa de acerto. Vale mencionar que este mesmo sistema obteve uma acurácia de 61,25% utilizando somente dados RGB.

Rezende et al. [8] também desenvolveu um conjunto de dados conhecido como MINDS-Libras. O conjunto consiste em 20 sinais repetidos 5 vezes por 12 interpretes, totalizando 1.200 sequências de dados gravados simultaneamente por um Microsoft Kinect v2 e uma DSLR da Canon. Os autores alegam que alguns vídeos foram perdidos e a versão final do conjunto de dados possui 1.155 sequências. O sistema de reconhecimento em [8] emprega Redes Convolucionais tridimensionais (3D-CNN) de forma similar à utilizada em Castro et al. [9]. Os autores também avaliaram diferentes abordagens com imagens em nível de cinza, Histogramas de

Gradientes Orientados (HOG), fluxo óptico, *data augmentation* e uso combinado das informações RGB e RGB-D. O melhor resultado foi obtido utilizando apenas os dados RGB convertidos em níveis de cinza com *data augmentation*, com uma acurácia média de 93,3%.

Um sistema de reconhecimento de sinais Libras baseado em modelos 3D e projeções 2D das mãos está descrito em Porfirio et al. [10]. Para avaliar o sistema, os autores elaboraram um conjunto de dados, conhecido como LIBRAS-HC-RGBDS, com 61 configurações de mão executadas duas vezes por 5 indivíduos. Cada sinal é representado por uma sequência de vídeo RGB-D obtida com um Microsoft Kinect.

O trabalho em [11] utiliza *Hidden Markov Models* (HMM) para classificar sinais de Libras em sequências de vídeo utilizando e tipos diferentes de extratores de características: 1) Conjunto básico, formado pela combinação das seguintes medidas extraídas diretamente dos vídeos: distância e ângulo das mãos em relação ao centro da face, direção do movimento, velocidade, aceleração, alongamento e compactação das mãos e ângulo do maior eixo de cada mão; 2) Saídas de uma CNN para 25 configurações de mão; e 3) Um conjunto misto, onde as características anteriores são concatenadas. Para avaliar o método, os autores desenvolveram um pequeno conjunto de dados composto por 15 sinais de Libras repetidos pelo mesmo indivíduo utilizando luvas amarelas em um ambiente controlado, totalizando 518 sequências de vídeo. Utilizando o conjunto misto de características, os autores obtiveram uma acurácia de 100%.

Todos os métodos e conjuntos de dados descritos nessa seção estão resumidos na Tabela I. Embora recentes e muito promissores, esses trabalhos possuem uma coisa em comum: foram validados em um conjunto de dados hermético e com poucos sinais. Nada se pode afirmar sobre o seu desempenho quando submetidos em sequências obtidas em cenários não controlados e em um ambiente mais realista, com um vocabulário maior.

B. Contribuições e organização do trabalho

Como pode ser observado na Seção I-A, uso de visão computacional para reconhecimento de sinais em libras não é novidade. É possível encontrar na literatura uma quantidade enorme de trabalhos que resolvem de maneira satisfatória o reconhecimento de gestos estáticos. Entretanto, o reconhecimento de sinais dinâmicos em sequências de vídeo é um problema em aberto. A maioria dos conjuntos de dados disponíveis é hermético e, portanto, nada se pode dizer sobre o uso dos sistemas desenvolvidos a partir desses conjuntos em cenários práticos reais e com um vocabulário extenso. Sem uma base escalável e não controlada é muito difícil obter um sistema. Entendemos que o uso de conjunto de dados *in-the-wild* é um passo na direção de resolver esse problema. Nesse trabalho, nós constituímos um conjunto de dados *ad-hoc*, formado por muitas sequências de vídeo obtidas em sites de *streamimg*. A metodologia usada na composição de um conjunto de dados dese tipo é a primeira contribuição deste trabalho. A segunda contribuição é o sistema de reconhecimento desenvolvido, que consiste em duas etapas: i) extração de características quadro-a-quadro através de uma CNN; e ii) modelagem da correlação temporal entre as medidas e classificação utilizando uma rede neural recorrente (RNN), mais precisamente uma *Long Short-Term Memory* (LSTM).

Este artigo está organizada da seguinte forma. A metodologia utilizada configuração de um conjunto de dados *ad-hoc* e não controlado está Seção II. A Seção III contém uma descrição detalhada do sistema de reconhecimento desenvolvido. Resultados e discussões na Seção IV. Conclusões e perspectivas futuras podem ser vistas na Seção V.

II. Base de dados

Para validar o sistema de reconhecimento proposto, constituímos um conjunto de dados *ad-hoc* e escalável utilizando os seguintes critérios: 1) Estar disponível em serviços de *streaming* gratuitos; 2) Possuir legenda, embutida ou não; 3) Grande variabilidade de iluminação, resolução, foco, posicionamento da câmera, sinalizador, postura do sinalizador ou qualquer outra padronização. Como resultado, temos um conjunto de dados desafiador com características *in-the-wild*, conforme pode ser observado na Figura 1.

Cada sinal, ou item lexical, e o seu respectivo rótulo foi extraído de modo semi-supervisionado da seguinte forma. Todos os quadros de cada vídeo foram convertidos em níveis de cinza e binarizados utilizando o método de Otsu [12] adaptativo. Em seguida computamos o vídeo residual a partir da diferença absoluta entre quadros sucessivos. A intensidade de movimento ao longo do vídeo foi obtida a partir da quantidade pixels não nulos no vídeo residual. Espera-se que entre, entre sinais sucessivos, o sinalizador faça uma pequena pausa e produza um vale no sinal de movimento. Como isso varia bastante entre os vídeos, o usuário define empiricamente um limiar que segmenta cada vídeo. Em cada trecho segmentado, o rótulo é extraído a partir da legenda. Nos casos onde a legenda está embutida, aplicamos o OCR disponível na biblioteca *pytesseract*. Utilizando essa metodologia, uma quantidade muito pequena de itens segmentados precisou ser ajustado manualmente (incluindo ou removendo quadros desnecessários). Dos vídeos obtidos, foram sinalizados 777 sinais simples e 707 destes foram corretamente validados, isso corresponde a 91,0% do total. Foram encontradas 377 legendas no intervalo identificado como sinal e destas 319 delas foram extraídas corretamente, o que corresponde a 84,6% do total.

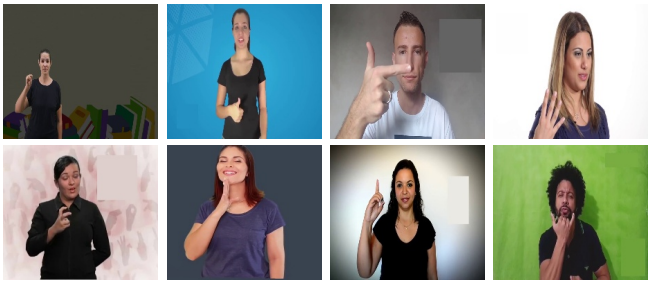

Fig. 1: Amostras de quadros de vídeos com sinalização em Libras disponível gratuitamente na internet.

Os números e principais características da base resultante foram os seguintes:

TABELA I: Síntese das referências apresentadas na Subseção I-A. As abreviações utilizadas são: Ref.: Referência, Dset.: Conjunto de Dados, NC: Número de Classes, NI: Número de intérpretes, NA: Número de amostras, NL: Nível de linguagem (I: isolado ou S: gestos sequenciais), Eq.: equipamento utilizado, TC: Técnica de Classificação, Acc: Acurácia e, por último, observações sobre o trabalho (PI: Palavras isoladas, PS: Palavras sequenciais, ESQL: Informação sobre o esqueleto).

Ref.	Ano	Dset	NC	NI	NA	NL	Eq	TC	Acc	Observações
[10]	2013	LIBRAS-HC-RGBDS	61	5	610	I	Kinect	SVM	98,36%	representações 2D e 3D
[11]	2019	Breda et al., 2019	15	1	518	I/Seq	Câmera+luvas	HMM	100%	PI e PS
[5]	2020	CEFET/RJ-Libras	24	20	547	I	Câmera	RF	65,81%	resíduo quadrático & 3-*fold*
[8]	2021	MINDS-Libras	20	12	1.155	I	Câmera+Kinect	CNN 3D	93,30%	somente RGB
[7]	2021	LIBRAS-UFOP	56	5	3.040	I	Kinect	CNN	74,25%	RGB-D, ESQ e face

- No total são 110 sinais diferentes, no qual 105 são dinâmicos e apenas 5 são estáticos.
- 54 ocorrências para a palavra: Prazer.
- 52 ocorrências para as palavras: Oi e Noite.
- 40 ocorrências para as palavras: Tarde e Tchau.
- 39 ocorrências para a expressão: Tudo bem.
- 38 ocorrências para as palavras: Desculpa e Conhecer.
- 31 ocorrências para a expressão: De nada.
- 29 ocorrências para a expressão: Por favor.
- 28 ocorrências para a palavra: Obrigado.
- 16 ocorrências para a palavra: Manhã.
- 3 ocorrências para as palavras e expressões: Boa tarde, bom dia.
- 2 ocorrências para as palavras e expressões: Você, dia, bom, bem, abrir, sim, sonhar, números, boa noite, de nada, com licença.
- 1 ocorrência para as palavras e expressões: Surdo, ouvinte, saber, amanhã, libras, não saber, aprender, ok, ano, beijos, tudo, até, todos, lá, meu/minha, quer, segundo, treinar, calar, não entender, olá, escola, onça, padaria, avó, taxista, eu, ele/ela, nós, vós, nosso/nossa, seu/sua, comigo, começar, não quer, não ver, falar, entender, minuto, hora, aqui, feriado, fechar, acostumar, brincar, bater, bater (porta), bater (surra), acordar, despertar, dormir, imaginar, ficar, entrar, sair, reunir, reunião, sono, não, nome, curso, coisas, calor, frio, boa manhã, boa madrugada, boa semana, mais ou menos, talvez, madrugada, cedo, trovão, vento, chuva, raio, neblina, furacão, neve, amarelo, azul, branco, marrom, rosa, preto, lilás, verde, vermelho, laranja, roxo.
- 67 pessoas sinalizaram usando a mão direita e apareceram com o corpo inteiro no enquadramento.
- 26 pessoas sinalizaram usando as duas mãos e apareceram com o corpo inteiro no enquadramento.
- 18 pessoas sinalizaram usando a mão esquerda e apareceram com o corpo inteiro no enquadramento.

Para a prototipação do sistema de reconhecimento projetado, um conjunto das 11 palavras mais frequentes foi separado. Na Tabela II estão as palavras escolhidas, juntamente com a informação do total de sinalizações, total de intérpretes e média de quadros por símbolo.

TABELA II: Detalhes sobre as classes, onde TS: total de sinalizações, TI: total de intérpretes e MQ: média de quadros por sinal.

Sinal	TS	TI	MQ
Com licença	41	35	56,29
Conhecer	38	25	37,21
De nada	31	31	93,32
Desculpa	38	32	77,60
Noite	52	37	55,94
Obrigado	28	22	65,89
Oi	52	47	56,36
Prazer	54	37	42,07
Tarde	40	28	45,70
Tchau	40	39	71,42
Tudo bem	39	37	63,12

III. Sistema de reconhecimento

Nesse trabalho foi desenvolvido um sistema de reconhecimento de gestos em libras capaz de incorporar tanto as características temporais dos sistemas dinâmicos quanto a diversidade de um conjunto de dados *in-the-wild*. O sistema na utilização de uma CNN para fazer a extração de características espaciais, seguida de uma RNN, para estimar as relações temporais entre as imagens. Todo o sistema foi desenvolvido em *Python* com auxílio da biblioteca *Keras* [13].

A. Preparação do conjunto de dados

Como explicado na Seção II, 11 classes foram escolhidas por serem as com mais amostras e maior variabilidade, representando bem um conjunto de dados desafiador. A quantidade de imagens relativas a cada sinalização não é constante. Por exemplo, em uma mesma classe podem existir representações com 20 quadros e outras com 120. Isso se deve às características do próprio intérprete, da natureza do vídeo, da resolução de gravação, entre outras. Para criar uniformidade na quantidade de dados sem perder informações relevantes, foi selecionado um conjunto de 18 quadros uniformemente amostrados para cada interpretação. Por fim, 70% das amostras foram utilizadas no treinamento e as demais 30% foram reservadas para o conjunto de teste. É importante ressaltar que o conjunto de imagens, de classes iguais ou diferentes, sinalizados pelo mesmo intérprete não está simultaneamente no grupo de treino e teste.

B. Extração de características utilizando CNN

Para fazer a extração de características foram utilizadas diferentes arquiteturas de CNN, e os resultados na classificação foram apurados para analisar seus desempenhos. As arquiteturas testadas foram: VGG16, VGG19 e *Inception*V3. Em todas elas foram utilizados os pesos da "*imagenet*". A

técnica "*global average pooling*" será aplicada na saída do último bloco convolucional, gerando um vetor na saída.

C. Reconhecimento dos gestos utilizando RNN

Para cada vídeo, 18 quadros uniformemente amostrados foram colocados na entrada da CNN, produzindo um vetor de 18 características. Essas características alimentam a LSTM que estabelece a relação temporal entre elas. Foram testados arranjos a partir de uma estrutura básica: Bloco LSTM com *dropout* intrínseco, seguido de um bloco *Dense* com função de ativação *ReLu*, uma camada de *dropout* e por fim outro bloco *Dense* com função de ativação *softmax*. Para essa estrutura e seus arranjos, foram utilizadas 1024, 2048 e 3072 camadas de LSTM, com valores de *dropout* intrínsecos de 30% e 50%. O bloco *Dense* permaneceu com 512 camadas em todos os arranjos. A camada *dropout* acompanhou o valor de referência utilizado no bloco LSTM e no último bloco *Dense* foram utilizadas 11 camadas, de acordo com o número de classes utilizadas.

D. Análise de Desempenho

O desempenho dos diferentes arranjos de redes neurais convolucionais e recorrentes foram analisados de acordo com a acurácia conjunto de treino e do conjunto de teste. Para demonstrar que o sistema de reconhecimento desenvolvido não está condicionado e nem ajustado exclusivamente para funcionar de acordo com os parâmetros da base de dados de sinais em Libras constituída, os melhor arranjo foi avaliado em uma base de dados da língua Argentina de sinais, chamada de LSA64. Ela possui 3.200 vídeos, com 10 intérpretes executando 5 repetições de 64 sinais diferentes. Os intérpretes gravaram as sinalizações vestindo roupas pretas e luvas coloridas, e os vídeos compartilham o mesmo cenário de fundo branco [14]. Para manter a mesma quantidade de classes, também foram utilizados 11 sinais do conjunto Argentino, sendo eles: "Azul claro", "Brilhante", "Cores", "Desenhista", "Homem", "Inimigo", "Longe", "Opaco", "Nascer", "Verde" e "Vermelho". Cada uma delas representada por 35 vídeos utilizados como referência para treino e 15 vídeos para validação.

IV. Resultados

No total, foi analisado um conjunto de 11 palavras e expressões correspondendo a 26.668 imagens. O número de intérpretes variou entre 22 e 47 pessoas, e o de ocorrências entre 31 e 52.

O desempenho dos diferentes arranjos foi analisado de acordo com a acurácia e perda do conjunto de treino e do conjunto de validação. Para manter a isonomia dos resultados, foram utilizadas 600 épocas de treino em todas as análises. A métrica definida para a avaliação global de desempenho do sistema de reconhecimento foi a acurácia do conjunto de validação. A acurácia corresponde à proporção de predições corretas em relação ao tamanho do conjunto de dados.

Nas próximas seções serão comparados e discutidos os desempenhos das configurações VGG16, VGG19 e InceptionV3, todas seguidas de LSTM. Na avaliação foram testados, além das diferentes configurações de redes, percentuais de *dropout* de 30% e 50%, e número de neurônios da rede LSTM de 1024, 2048 e 3072.

TABELA III: Melhores resultados de acurácia de teste obtidos para cada configuração. As linhas representam as estruturas VGG16+LSTM, VGG19+LSTM e *Inception*V3+LSTM com *dropout* de 30% e 50% e números de neurônios na rede LSTM de 1024, 2048 e 3072. Tudo aplicado a base de dados de Libras.

		Número de neurônios		
Estrutura	Dropout	1024	2048	3072
VGG16+LSTM	30%	41,5%	47%	**51,4%**
	50%	47%	50%	44%
VGG19+LSTM	30%	52%	**55,8%**	52,5%
	50%	37%	43%	50,5%
*Inception*V3+LSTM	30%	60%	61%	50,5%
	50%	57%	59,5%	**62,8%**

A. VGG16 e LSTM

A quantidade de neurônios utilizados no LSTM impactou diretamente na velocidade de convergência do treino, que em todos os arranjos chegou aos 100%. Apesar disso, os resultados de acurácia do conjunto de teste mantiveram-se estáveis na faixa entre 30% e 50%, atingindo o valor máximo de 51,4% na estrutura com 3072 neurônios. Considerando os arranjos com *dropout* de 50%, a convergência dos dados de treino aconteceu de forma mais lenta do que nos arranjos com *dropout* de 30% e os resultados de acurácia do conjunto de validação oscilaram em uma faixa levemente mais baixa, entre 25% e 50%, com o máximo de 49,5% na arquitetura de 2048 neurônios LSTM. Conforme pode ser observado na Tabela III, a estrutura que apresentou o melhor resultado no conjunto de teste para a composição VGG16 e LSTM foi a de 3072 neurônios com 30% de *dropout*.

B. VGG19 e LSTM

O comportamento desses arranjos foi similar aos dos arranjos da VGG16 com *dropout* de 30%. O treino convergiu em todas as estruturas e a acurácia dos dados de validação oscilou na faixa entre 30% e 55%, atingindo o valor de 55,8% na configuração com 2048 neurônios LSTM. Os arranjos com *dropout* de 50% apresentaram valores de acurácia para o conjunto de treino compreendidos entre 30% e 50% na maior parte das épocas, atingindo o pico de 51,2% na estrutura com 3072 neurônios LSTM. Para a composição VGG19 e LSTM, a estrutura que apresentou os melhores resultados foi a de 2048 neurônios com 30% de *dropout*, conforme pode ser observado na Tabela III.

C. InceptionV3 e LSTM

A convergência dos valores de acurácia aconteceu por volta das 100 épocas nos três arranjos experimentados, com *dropout* de 30%. Os resultados referentes ao conjunto de teste oscilaram em torno de 55%, atingindo mais de 60% nas três estruturas, com o pico de 61,6% de acerto na estrutura com 3072 neurônios. No arranjo com 1024 neurônios LSTM a convergência do conjunto de treino aconteceu tardiamente quando comparada com os arranjos de 2048 e 3072 neurônios. Os resultados referentes à acurácia no conjunto de teste oscilaram em torno de 50% após a estabilização, atingindo

um valor máximo de 62,8% na estrutura com 3072 neurônio. Para os arranjos formados a partir da composição *Inception*V3 e LSTM, a estrutura que apresentou os melhores resultados foi a de 3072 neurônios LSTM com 50% de *dropout*, como pode ser visto na Tabela III.

D. Avaliação no conjunto de dados LSA64

O conjunto de dados constituído é bastante desafiador, por conta da sua diversidade. Com o intuito de verificar que o modelo proposto é capaz de ter resultados satisfatórios em outras condições, o melhor modelo de cada arranjo foi avaliado em um conjunto de dados da língua Argentina de sinais. Para tanto, foi utilizado um subconjunto da LSA64, conforme descrito na Seção III-D. No arranjo com VGG16 e LSTM, 3072 neurônios e dropout de 30%, o modelo proposto atingiu 97,4% de acurácia no conjunto de teste. Já no conjunto VGG19 e LSTM, 2048 neurônios e dropout de 30%, a acurácia foi de 92,2%. Por outro lado, no arranjo com InceptionV3 e LSTM, 3072 neurônios e dropout de 50%, a acurácia no conjunto de teste foi de 93,3%. Todos esses resultados, bem como uma comparação com os resultados obtidos no conjunto de dados de Libras, estão condensados na Tabela IV. Conforme descrito na Seção III-D, o conjunto LSA64 é bastante uniforme, uma vez que os intérpretes utilizam a mesma padronização de vestuário e de pose, com plano de fundo plano e iluminação constante. É esperado que o modelo proposto tenha um desempenho maior neste cenário.

TABELA IV: Melhores resultados obtidos pelas estruturas propostas no banco de dados de Libras criado e na base Argentina de sinais, LSA64.

Estrutura	Acc.	
	Libras	LSA64
VGG16+LSTM	51,4%	**97,6%**
VGG19+LSTM	55,8%	92,2%
InceptionV3+LSTM	**62,8%**	93,3%

V. Conclusão

Dada a população de deficientes auditivos no Brasil, um sistema eficiente para reconhecimento de sinais em Libras é essencial. Muitos trabalhos na literatura [5]–[11] apresentam bases de dados em ambiente controlado que são utilizados para a detecção.

Neste trabalho, propomos um método para constituir um conjunto de dados mais próximo de cenários reais, com grande variação de interprete, pose e cenário. Foram utilizados 11 sinais provenientes de vídeos disponíveis em portais de *streaming*. É importante notar que os vídeos disponíveis apresentam classes naturalmente desbalanceadas. Além disso, cada intérprete demanda um tempo distinto para a apresentação do sinal.

Pela variedade espaço-temporal das características dos vídeos, o sistema proposto para o reconhecimento de sinais em Libras é a combinação de uma rede neural convolucional e de uma rede recorrente. Os melhores resultados foram obtidos utilizando o conjunto InceptionV3 com LSTM, com uma acurácia de 62,8%. Como o conjunto de teste constituído é bastante desafiador, em caráter comparativo, a rede foi treinada e testada em um subconjunto na base LSA64 [14]. A conjunto de dados argentino também contém 11 classes, porém os dados são gravados em ambiente controlado, com apenas 10 intérpretes diferentes, todos eles com vestimentas na cor preta, com enquadramento na gravação semelhantes para cada sinalizador e mesmo cenário de fundo. Na configuração Inception V3 com LSTM o método proposto atingiu 93,3% no subconjunto de teste da LSA64. Entretanto, o melhor resultado para o conjunto de dados Argentino, foi obtido com a configuração VGG19 com LSTM: 97,6% de acurácia. Por fim, conclui-se que o sistema de reconhecimento proposto possui resultados promissores na base extremamente desafiadora utilizada. Esses são reforçados pela aplicação em outra língua de sinais com excelentes resultados.

Referências

[1] Instituto Brasileiro de Geografia e Estatística (IBGE), "Censo demográfico 2010 - características gerais da população, religião e pessoas com deficiência," 2010, acessado em 14 de junho de 2020. [Online]. Available: https://biblioteca.ibge.gov.br/visualizacao/periodicos/94/cd_2010_religiao_deficiencia.pdf

[2] World Federation of the Deaf (WFD), "Position paper regarding the united nations convention onthe rights of people with disabilities," acessado em 15 de junho de 2020. [Online]. Available: http://www.un.org/esa/socdev/enable/rights/contrib-wfd.htm

[3] R. M. de Quadros and G. Perlin, *Estudos Surdos III*. Rio de Janeiro, Brasil: Editora Arara Azul, 2007.

[4] D. Otsuka, "Língua brasileira de sinais (Libras)," 2010, acessado em 24 de maio de 2020. [Online]. Available: https://www.infoescola.com/comunicacao/lingua-brasileira-de-sinais-libras/

[5] P. V. Gameiro, W. L. Passos, G. M. Araujo, A. A. Lima, J. N. Gois, and A. R. Corbo, "A brazilian sign language video database for automatic recognition," in *17th IEEE Latin American Robotics Symposium*, Natal, BR, Nov. 2020, pp. 1–6.

[6] C. H. A. Monteiro, L. F. I. Pecoraro, A. T. Lacerda, A. R. Corbo, and G. M. Araujo, "Um sistema de baixo custo para reconhecimento de gestos em libras utilizando visão computacional," in *XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, Santarém, Brazil, Aug. 2016, pp. 349–352.

[7] L. R. Cerna, E. J. E. Cardenas, D. G. Miranda, D. Menotti, and G. Camara-Chavez, "A multimodal libras-ufop brazilian sign language dataset of minimal pairs using a microsoft kinect sensor," *Expert Systems with Applications*, vol. 167, no. 114179, pp. 1–14, Oct. 2021.

[8] T. Rezende, S. Almeida, and F. Guimarães, "Development and validation of a brazilian sign language database for human gesture recognition," *Neural Computing and Applications*, vol. 1, pp. 1–19, Mar. 2021.

[9] G. Z. de Castro, R. R. Guerra, M. M. de Assis, T. M. Rezende, G. T. B. de Almeida, S. Almeida, C. L. de Castro, and F. G. G. aes, "Desenvolvimento de uma base de dados de sinais de Libras para aprendizado de máquina: Estudo de caso com CNN 3D," in *XIV Brazilian Intelligent Automation Symposium (SBAI)*, Ouro Preto, Brazil, Oct. 2019, pp. 2116–2121.

[10] A. J. Porfirio, K. L. Wiggers, L. E. S. Oliveira, and D. Weingaertner, "Libras sign language hand configuration recognition based on 3d meshes," in *IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, UK, Oct. 2013, pp. 1588–1593.

[11] V. M. Breda and D. Silva, "Reconhecimento de gestos em vídeos utilizando modelos ocultos de Markov e redes neurais convolucionais aplicado a Libras," in *XXXVII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, Petrópolis, Brazil, Sep. 2019, pp. 1–5.

[12] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[13] F. Chollet, "About keras," 2017, acessado em 14 de junho de 2020. [Online]. Available: https://keras.io/about

[14] F. Ronchetti, F. Quiroga, C. Estrebou, L. Lanzarini, and A. Rosete, "LSA64: A dataset of argentinian sign language," *XXII Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016.