Redes Neurais Artificiais Aplicadas ao Reconhecimento de Expressões Faciais Negativas na Língua Brasileira de Sinais

Ana Gabriela P. e Silva, Emanuel O. da Silva, Giovana de Lucca, Letícia C. Passos, Matheus M. Matos, Elloá B. Guedes

[1]Núcleo de Computação
Escola Superior de Tecnologia
Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

{agps, eos, gol, lcps, mmmt}.eng@uea.edu.br, ebgcosta@uea.edu.br

Abstract. *Facial expressions play an important role in Brazilian sign language, a gesture-visual language used by people with hearing disabilities. Aiming at collaborating with solutions to the automatic recognition of this language, facial expressions must also be considered. In this work, several artificial neural networks were trained and tested for negative facial expressions classification, whose detection in previous works had low performance. The results obtained indicate a* 63% *percentual improvement in this task when compared to existing works, with advances also in detection of conditional and binary facial expressions.*

Resumo. *As expressões faciais possuem um papel importante na Língua Brasileira de Sinais, a qual é uma linguagem gesto-visual utilizada por pessoas com deficiência auditiva. Na tentativa de elaborar soluções que colaborem para o reconhecimento automático desta língua, as expressões faciais também precisam ser consideradas. Neste trabalho, diferentes redes neurais foram treinadas e testadas para a classificação de expressões faciais gramaticais negativas, cujo detecção em trabalhos prévios da literatura foi identificado como tendo baixo desempenho. Os resultados obtidos indicam uma melhoria percentual de* 63% *nesta tarefa em relação aos trabalhos existentes, com ganhos também na detecção de expressões faciais condicionais e binárias.*

1. Introdução

A *Língua Brasileira de Sinais* (Libras) é uma linguagem gesto-visual segundo a qual uma mensagem é emitida por meio de movimentos nas mãos, corpo e face e recebida pela visão [de Fátima Brecailo 2012]. Esta língua é um caminho para a abertura social de pessoas surdas, surdo-cegas e com outros tipos de deficiência, tendo sido reconhecida pela Lei Federal 10.436 de 24/04/2002 [Flores et al. 2012].

Em Libras, cada palavra é representada por um sinal e carrega níveis linguísticos diferentes, tais como fonologia, morfologia, sintaxe e semântica. Além disso, as expressões faciais e corporais corroboram para o entendimento da informação. As expressões faciais, em particular, podem ser afetivas ou gramaticais. Quando possuem caráter gramatical, as chamadas *expressões faciais gramaticais*, conferem informação

gramatical a uma sentença expressa em sinais, complementando o seu sentido e podendo ser de nove diferentes tipos gerais [Quadros and Karnopp 2004].

O reconhecimento automático da língua de sinais é uma importante área de pesquisa que tem como objetivo atenuar os obstáculos impostos no dia a dia das pessoas surdas e/ou com deficiência auditiva e aumentar a integração destas pessoas na sociedade majoritariamente ouvinte [Teodoro 2015]. Para elaboração de soluções neste domínio, é essencial, portanto, considerar também a análise das expressões faciais gramaticais. O trabalho de Freitas et al., por exemplo, já considerou esta perspectiva [de Almeida Freitas et al. 2014b]. Utilizando redes neurais artificiais, os autores conceberam um modelo capaz de classificar a expressão facial emitida, obtendo resultados satisfatórios para algumas expressões, mas com baixo desempenho na detecção de expressões faciais negativas, com F-*score* em torno de 0.45.

Considerando as limitações identificadas para detectar expressões faciais gramaticais negativas na literatura, este trabalho se propôs a explorar outras redes neurais para o problema identificado, objetivando a concepção de modelos com melhor desempenho na correta classificação deste tipo de expressão. Um dimensionamento das redes neurais para este cenário permitiu a identificação de 110 redes neurais adequadas, das quais 10 obtiveram melhores resultados na etapa de testes. Considerando a métrica de F-*score*, foi possível identificar uma rede neural com duas camadas ocultas com desempenho igual a 0.73, um incremento percentual de 63% na classificação correta de expressões negativas quando comparado ao estado da arte.

Para apresentar os resultados obtidos, este trabalho está organizado como segue. Os conceitos fundamentais sobre expressões faciais gramaticais, com ênfase nas expressões negativas, são mostrados na Seção 1.1. Uma visão geral do conjunto de dados utilizado encontra-se detalhada na Seção 2. A metodologia utilizada para conceber as diferentes redes neurais para este problema pode ser vista na Seção 3. Os resultados e a discussão são apresentados na Seção 4. Por fim, as considerações finais e sugestões de trabalhos futuros são mostrados na Seção 5.

1.1. Expressões Faciais Gramaticais

A Libras é uma língua de modalidade de comunicaçãoo gestual-visual porque utiliza, como canal ou meio de comunicação, movimentos gestuais e expressões faciais que são percebidos pela visão. A formação de sinais nesta língua envolve a combinação de diferentes parâmetros, a citar: (1) configuração das mãos; (2) pontos de articulação; (3) movimento; (4) orientação e (5) expressão facial e/ou corporal [Ramos 2004].

No tocante às expressões faciais, as mesmas estão relacionadas à estruturas nos níveis da morfologia e sintaxe, sendo obrigatórias em determinados contextos [Quadros and Karnopp 2004]. No nível morfológico, as expressões faciais correspondem ao grau de intensidade de um adjetivo ou ao grau de tamanho para um substantivo. Já a nível da sintaxe, as expressões faciais determinam o tipo da estrutura da sentença, a qual pode ser de nove tipos diferentes, conforme ilustrado na Tabela 1.

Para realizar estas expressões faciais devem ser utilizados o movimento da cabeça, a direção do olhar, a elevação das sobrancelhas, o franzir da testa e até mesmo o movimentos dos lábios, conforme ilustrado na Figura 1. Cada expressão facial possui características bem definidas e pode aparecer mais de uma vez em uma única sentença. É

Tabela 1: Tipos de estrutura de sentenças quanto à sintaxe das expressões faciais e exemplos correspondentes. Elaborado a partir de: [Quadros and Karnopp 2004, Ramos 2004, Sousa 2010]

Tipo de Sentença	Exemplo
Afirmativa	*Eu irei à escola.*
Negativa	*Não gosto de chocolate.*
Condicional	*Se chegarmos no horário iremos ao cinema.*
Interrogativa com pronome interrogativo	*Quando é o seu aniversário?*
Interrogativa binária (sim/não)	*Você vai ao shopping hoje?*
Interrogativa de dúvida	*Você tem certeza que esse lápis é seu?*
Relativa	*O carro que quebrou está na oficina.*
Tópico	*Cores, eu gosto de vermelho.*
Foco	*O almoço foi risoto. Não, o almoço foi arroz.*

importante ressaltar que a omissão das expressões faciais pode retirar o sentido de uma determinada frase emitida na língua brasileira de sinais [Quadros and Karnopp 2004].

(a) Negativa **(b) Afirmativa** **(c) Interrogativa**

Figura 1: Exemplos de Expressões Faciais Gramaticais. Imagens: [Sousa 2010].

As expressões faciais gramaticais negativas podem ser indicada de duas formas: (1) com o movimento para os lados, porém ressalta-se que este movimento não é obrigatório em Libras e refere-se principalmente às questões discursivas; ou (2) por meio da modificação do contorno da boca, juntamente com o abaixamento das sobrancelhas e levemente da cabeça. Esta segunda forma, em particular, é considerada obrigatória para enfatizar a negação por estar relacionada ao aspecto sintático [Sousa 2010].

Dadas as características particulares das expressões negativas e a complexidade envolvida na emissão das mesmas, o reconhecimento automático deste tipo de expressão facial tem se mostrado um desafio. Além do que foi exposto, é essencial considerar também as variações nas expressões faciais emitidas por diferentes pessoas e ainda a co-ocorrência destas expressões com outros elementos da língua brasileira de sinais, que podem resultar em oclusão da face. Estes dois aspectos acentuam as dificuldades na concepção de métodos automáticos de reconhecimento [de Almeida Freitas et al. 2014b].

O reconhecimento automático da configuração das mãos em Libras já foi

considerado por alguns trabalhos na literatura [Teodoro 2015, Porfirio 2013], e dada a importância das expressões faciais gramaticais, mais recentemente este aspecto também foi endereçado. Freitas et al. inicialmente conceberam um conjunto de dados, intitulado *Grammatical Facial Expressions Data Set*, contendo 100 coordenadas de diferentes posições faciais obtidas a partir de 225 vídeos gravados com um sensor de movimentos enquanto um sujeito emitia diferentes sentenças em Libras [de Almeida Freitas et al. 2014a].

Utilizando apenas 17 coordenadas das diferentes posições faciais no *dataset* concebido, os autores treinaram e testaram uma única rede neural artificial para o problema de classificação da expressão facial gramatical correspondente. A rede proposta pelos autores, do tipo *multilayer perceptron* com 10 neurônios na camada oculta, foi treinada com diferentes taxas de aprendizado. De acordo com os resultados obtidos, a melhor expressão facial detectada foi a de foco, ao passo que as expressões negativas tiveram F*-score* de 45%, indicando uma baixa sensitividade do modelo proposto, o qual classifica menos expressões negativas do que deveria [de Almeida Freitas et al. 2014b]. A melhoria desta métrica de desempenho é um dos objetivos centrais deste trabalho.

2. Visão Geral do Conjunto de Dados

O *dataset* utilizado para a realização deste trabalho, *Grammatical Facial Expressions Data Set*, é composto de exemplos de expressões faciais gramaticais anotadas de maneira supervisionada a partir de 225 vídeos nos quais dois sujeitos emitem sentenças em Libras [de Almeida Freitas et al. 2014a]. A partir de cada vídeo, os autores do *dataset* fizeram uma amostragem em *frames* nos quais foram marcadas 100 coordenadas (x, y, z) indicando altura, largura e profundidade de diferentes pontos de referência na face do sujeito. Um atributo alvo denota qual expressão gramatical está sendo emitida no respectivo *frame*, que pode ser um dos 9 tipos de expressões faciais existentes ou até mesmo nenhum dos tipos considerados (expressão neutra). Em resumo, cada exemplo neste *dataset* é composto de 300 atributos preditores e um atributo alvo. O *dataset* pode ser obtido gratuitamente e é uma iniciativa dos autores na tentativa de estimular novos trabalhos que enderecem a detecção de expressões faciais gramaticais [de Almeida Freitas et al. 2014a].

No escopo deste trabalho, decidiu-se utilizar os mesmos 17 pontos na face adotados no trabalho de Freitas et al. [de Almeida Freitas et al. 2014b], localizados nas seguintes regiões da face: olhos esquerdo e direito (2 pontos em cada), sobrancelhas esquerda e direita, íris esquerda e direita, linhas acima da sobrancelha esquerda e direita, nariz, boca (2 pontos), ponta do nariz e contorno da face. Um pré-processamento eliminou os demais pontos do *dataset*, ignorou os dados do sujeito que emitia os sinais (que poderia ser de dois tipos distintos) e descartou os dados do *timestamp*, por não influenciarem na obtenção dos resultados.

Uma análise preliminar do conjunto de dados permitiu identificar que os atributos correspondentes às coordenadas z (referentes à profundidade) poderiam ser descartados. Conforme ilustrado na Figura 2, além de possuírem uma variação muito pequena, a existência de *outliers* forneceu evidências sobre eventuais erros de medição.

Considerando os atributos preditores disponíveis, foi analisada a estatística descritiva dos mesmos, detalhada na Tabela 2. É interessante notar que os atributos preditores possuem distribuições oblíquas, o que pode ser evidenciado pela não coincidência entre

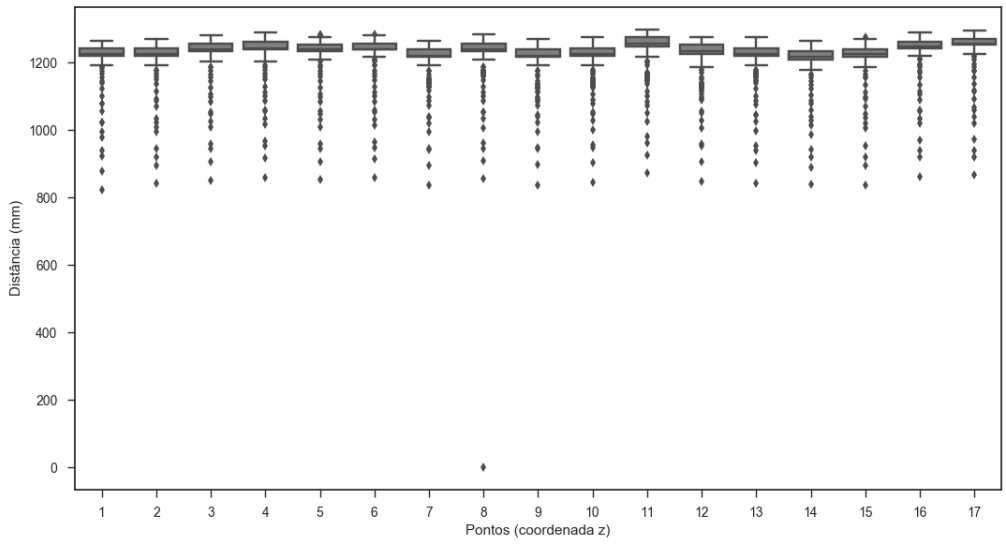

Figura 2: Boxplot da distribuição das coordenadas de profundidade referentes aos 17 pontos escolhidos.

média e mediana, e que o desvio padrão é significativo, indicando dispersão.

Tabela 2: Estatística descritiva dos 17 pontos selecionados. Os símbolos $\overline{\mathbf{x}}$,$\tilde{\mathbf{x}}$ e $\sigma_{\mathbf{x}}$ denotam a média, mediana e desvio padrão da coordenada x, respectivamente. A notação é análoga para a coordenada y.

Posição na Face	$\overline{\mathbf{x}}$	$\tilde{\mathbf{x}}$	$\sigma_{\mathbf{x}}$	$\overline{\mathbf{y}}$	$\tilde{\mathbf{y}}$	$\sigma_{\mathbf{y}}$
Olho esquerdo	303.44	304.93	11.33	218.93	229.06	15.13
Olho direito	332.18	333.15	11.23	218.49	227.43	13.83
Sobrancelha esquerda	300.66	301.33	12.53	211.23	221.23	15.11
Sobrancelha direita	334.37	333.63	11.62	210.90	218.82	13.59
Nariz	317.62	317.82	12.76	228.28	232.96	17.57
Boca	318.39	319.05	14.18	245.03	255.32	18.67
Contorno da face	319.83	319.82	27.16	245.97	243.08	22.02
Íris esquerda	303.44	306.49	10.69	218.93	229.05	15.07
Íris direita	332.18	334.14	10.56	218.49	228.07	13.76
Ponta do nariz	317.64	319.62	11.49	231.87	241.99	17.23
Linha acima da sobrancelha esquerda	229.75	300.41	13.46	207.75	217.39	14.79
Linha acima da sobrancelha direita	335.13	335.04	12.08	207.07	214.91	13.17

Concluída a caracterização e análise do *dataset* utilizado, partiu-se então para o dimensionamento, treino e teste das redes neurais, cuja metodologia de realização encontra-se descrita na seção a seguir.

3. Materiais e Métodos

As redes neurais adotadas neste trabalho foram as redes *feedforward multilayer perceptron*, treinadas com o algoritmo *backpropagation* implementado pelo método do gradiente

descendente estocástico. A camada de entrada das redes neurais propostas deveria constar de 34 neurônios, relativos ao atributos preditivos disponíveis no *dataset* e descritos na seção anterior. A camada de saída possuía apenas um neurônio, responsável por indicar se os dados da expressão facial fornecida como entrada eram relativos a uma expressão negativa ou não. Pode-se verificar, portanto, que a tarefa de aprendizado considerada foi uma tarefa de classificação.

Para propor diferentes redes neurais para este cenário, considerou-se a utilização de redes neurais com uma ou duas camadas ocultas, em virtude de serem aproximadoras universais de qualquer função [Haykin 2009]. Quanto ao número de neurônios nessas camadas, embora não haja uma maneira analítica de precisar este valor, considerou-se a regra da pirâmide geométrica:

$$N_h = \alpha \cdot \sqrt{N_i \times N_o}, \tag{1}$$

em que N_h é o número de neurônios na camada oculta, que se deseja determinar; α é uma constante que assume valores no intervalo $0.5 \leq \alpha \leq 2$; N_i é o número de neurônios na camada de entrada ($N_i = 34$); e N_o é o número de neurônios na camada oculta. Como resultado, observou-se que N_h residia no intervalo de 3 a 12. Levando isto em consideração, foram geradas diferentes arquiteturas de redes neurais considerando a distribuição desta quantidade de neurônios em até duas camadas, resultando em 110 redes neurais adequadas ao problema em questão. Todas estas redes possuíam função de ativação tangente hiperbólica e taxa de aprendizado adaptativa igual a 0.01

Os exemplos disponíveis no conjunto de dados foram utilizados para treinar e testar as redes, considerando uma partição de 70% para treino e de 30% para testes. Da quantidade disponível para treino, 10% dos exemplos foram reservados para validação e prevenção de *overfitting*. Os atributos de entrada foram normalizados antes de sua apresentação às redes. Também foram consideradas 200 execuções do treino e teste das redes neurais com apresentação aleatória dos exemplos, visando diminuir algum viés introduzido pelas escolhas randômicas dos pesos iniciais.

Em virtude do problema de identificação das expressões faciais gramaticais negativas ter sido endereçado como um problema de classificação binária, a métrica de desempenho F*-score* foi adotada para comparar as diferentes redes neurais obtidas e elencar as que possuíam melhor performance perante o conjunto de testes. O F*-score* é derivado a partir de duas outras métricas, a *precisão* e a *revocação*, definidas como a fração de instâncias recuperadas que são relevantes e a fração de instâncias relevantes que são recuperadas, respectivamente. A Eq. (2) detalha a obtenção desta métrica:

$$\mathbf{F}\text{-Score} = 2 \cdot \frac{\text{precisão} \times \text{revocação}}{\text{precisão} + \text{revocação}}. \tag{2}$$

A métrica *F-score* é obtida pela média harmônica da precisão e da revocação, assumindo valores no intervalo $[0, 1]$. Quanto maior o seu valor, melhor o resultado obtido.

4. Resultados e Discussão

Considerando a metodologia definida, as 110 redes neurais identificadas para o contexto deste trabalho foram implementadas, treinadas e testadas com ferramental da lingua-

gem de programação Python, nos quais o *framework* `sci-kit learn`[1] e a biblioteca `pandas`[2] possuíram papel central.

Considerando a métrica de desempenho adotada, o desempenho médio das redes em termos de F-*score* foi de 0.66, em que o menor valor observado foi de 0.52 e o máximo foi de 0.73. Selecionando as 10 redes com maior F-*score* neste cenário, tem-se os resultados detalhados ilustrados na Tabela 3.

Tabela 3: Arquiteturas e métricas obtidas das 10 redes neurais com maior F-*score*.

Arquitetura	F-score	Precisão	Revocação
(34,12,10,1)	0.732806	0.732806	0.704569
(34,12,9,1)	0.720248	0.755333	0.692118
(34,11,12,1)	0.719625	0.755280	0.689983
(34,12,12,1)	0.719400	0.765068	0.683065
(34,12,7,1)	0.718982	0.751323	0.694665
(34,11,9,1)	0.718333	0.751099	0.695151
(34,11,10,1)	0.717882	0.717882	0.717882
(34,11,11,1)	0.715041	0.753240	0.686006
(34,12,8,1)	0.713988	0.713988	0.682642
(34,12,11,1)	0.713128	0.755166	0.679593

De acordo com os resultados obtidos, é possível verificar que as redes neurais que melhor endereçaram a identificação das expressões faciais negativas possuíam 2 camadas ocultas. Em particular, a rede *multilayer perceptron* com duas camadas ocultas, tendo a primeira 12 neurônios e a segunda 10 neurônios, foi a que obteve melhor performance neste cenário. Em comparação com o trabalho de Freitas et al. v, cuja rede neural proposta obteve F-*score* igual a 0.45, é possível observar um incremento percentual de 63% na classificação.

Após a identificação da rede neural (34-12-10-1) com melhor desempenho para a detecção de expressões faciais negativas dentre o conjunto de redes propostas no cenário de testes considerado, um questão em aberto que emergiu consistia em saber qual o desempenho desta rede perante a classificação das demais expressões faciais.

Para responder à pergunta remanescente, foram consideradas as demais expressões faciais gramaticais de Libras e o exemplos existentes no *dataset* original para treino e testes, respeitando as partições anteriormente consideradas. A única diferença significativa residiu na camada de saída, cuja função de ativação passou a ser `softmax`, adequada para cenários de classificação com múltiplas classes. Após um novo treinamento, os resultados obtidos do F-*score* foram agrupados por expressão facial, e comparados com os resultados atualmente obtidos pela literatura [de Almeida Freitas et al. 2014b], ilustrados na Figura 3.

Como é possível observar, embora tenha havido pequenas melhoras e perdas em algumas expressões faciais, a rede neural proposta neste trabalho apresenta desempenho superior na classificação das expressões faciais negativas, condicionais e binárias.

[1] `http://scikit-learn.org/`
[2] `http://pandas.pydata.org/`

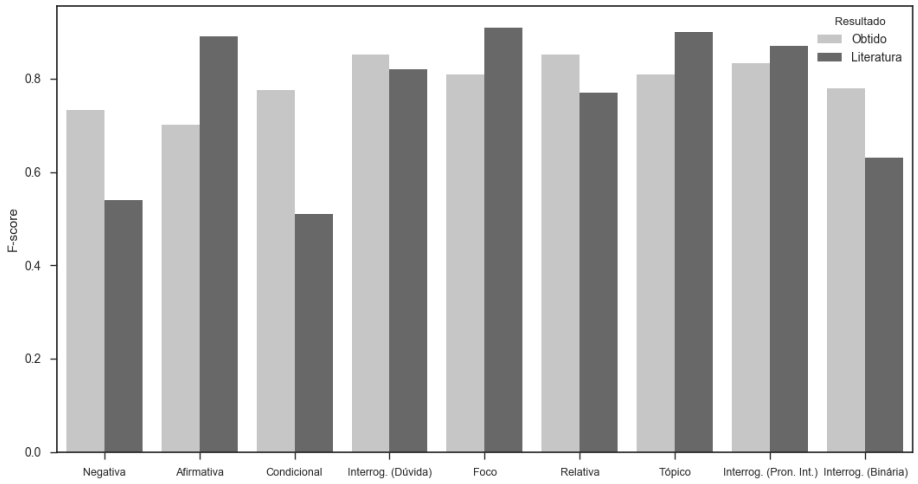

Figura 3: Comparativo do F-*score* para a rede (34-12-10-1) na classificação de todas as expressões faciais gramaticais em contraste com os resultados da literatura.

5. Considerações Finais

Este trabalho teve por objetivo propor, treinar, testar e identificar redes neurais que melhor endereçassem a identificação de expressões faciais negativas em Libras. Para tanto, foram propostas 110 redes, das quais 10 foram selecionadas como tendo resultados satisfatórios. Uma análise dos resultados permitiu identificar a rede com arquitetura (34-12-10-1) como tendo melhor desempenho, fornecendo um ganho percentual de 63% na classificação correta de expressões negativas quando comparada ao estado da arte. Posteriormente, esta mesma rede foi testada para as demais expressões faciais e, para algumas demas, também houve uma melhoria no desempenho na tarefa de detecção.

Os resultados obtidos neste trabalho corroboram para o reconhecimento automático da Língua Brasileira de Sinais e evidenciam as redes neurais artificiais como um modelo da Aprendizagem de Máquina adequado para o cenário em questão. Além da melhoria de performance obtida, outro aspecto que merece ser ressaltado na solução proposta neste trabalho é a eliminação da coordenada de profundidade, o que pode colaborar no reconhecimento de Libras a partir de vídeos bidimensionais, como aqueles capturados com auxílio de *smartphones*. Isto pode colaborar no desenvolvimento de novas soluções acessíveis ao grande público.

Em trabalhos futuros almeja-se relacionar a expressão facial com o movimento das mãos, colaborando para o reconhecimento automático de sentenças completas. Além disso, sugere-se a investigação de técnicas para anotação automática dos pontos faciais a partir de vídeos, evitando que esta tarefa seja realizada de maneira supervisionada.

Agradecimentos

Os autores Emanuel Oliveira da Silva, Letícia C. Passos e Matheus Miranda Matos agradecem o apoio financeiro provido pela Universidade do Estado do Amazonas e pela Fundação de Amparo à Pesquisa do Amazonas por meio do Programa de Apoio à Iniciação Científica.

Referências

de Almeida Freitas, F., Barbosa, F. V., and Peres, S. M. (2014a). Grammatical facial expressions data set. `https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions`. Acessado em 11 de setembro de 2017.

de Almeida Freitas, F., Barbosa, F. V., and Peres, S. M. (2014b). Grammatical facial expressions recognition with machine learning. In *International Florida Artificial Intelligence Research Society Conference*, pages 180–185.

de Fátima Brecailo, S. (2012). Expressão facial e corporal na comunicação em Libras. IMAP. Disponível em `goo.gl/Lw5KYD`. Acessado em 11 de setembro de 2017.

Flores, E. M., Barbosa, J. L. V., and Rigo, S. J. (2012). Um estudo de técnicas aplicadas ao reconhecimento da língua de sinais: novas possibilidades de inclusão digital. *Revista Novas Tecnologias na Educação*, 10(3):1–10.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson, Nova Jersey, 3 edition.

Porfirio, A. J. (2013). Reconhecimento das configurações de mão da Libras a partir de malhas 3D. Master's thesis, Universidade Federal do Paraná, Curitiba.

Quadros, R. M. and Karnopp, L. B. (2004). *Língua de sinais brasileira: estudos lingüísticos*. Artmed Editora, Porto Alegre. Capítulo 4.

Ramos, C. R. (2004). LIBRAS: a língua de sinais dos surdos brasileiros. *Revista Virtual de Cultura Surda e diversidade*, (4):1–15.

Sousa, D. V. C. (2010). Um olhar sobre os aspectos linguísticos da língua brasileira de sinais. *Littera Online*, 1(2):88–100.

Teodoro, B. T. (2015). Sistema de reconhecimento automático de língua brasileira de sinais. Master's thesis, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo.