INTERFACE TECNOLÓGICA

AVALIAÇÃO DE RECONHECIMENTO DE EMOÇÃO DIMENSIONAL COM UMA ABORDAGEM BASEADA EM ATENÇÃO

EVALUATION OF AN ATTENTION BASED DIMENSIONAL EMOTION RECOGNITION

Antonio José da Silva Ferreira – antonio.jsilvaf@gmail.com
Faculdade de Tecnologia de Catanduva (Fatec) – Catanduva – SP – Brasil

Gabriela Regina Soares – gabysoares09@gmail.com
Faculdade de Tecnologia de Catanduva (Fatec) – Catanduva – SP – Brasil

João Baptista Cardia Neto – joao.cardia@fatec.sp.gov.br
Faculdade de Tecnologia de Catanduva (Fatec) – Catanduva – SP – Brasil

DOI: 10.31510/infa.v19i2.1523
Data de submissão: 01/09/2022
Data do aceite: 28/11/2022
Data da publicação: 20/12/2022

RESUMO

O feedback não verbal e o reconhecimento das expressões faciais têm sido área de muita pesquisa nas últimas décadas. As expressões faciais são uma maneira concreta de reconhecer emoções e "ensinar" os computadores a detectar corretamente o que cada expressão facial significa e a qual emoção está ligada. Assim, no âmbito do reconhecimento de imagens, as Redes Neurais Convolucionais (RNC), através de suas camadas sobre os pixels da imagem, facilitam a descoberta de padrões. Dessa forma, através da aplicação de uma RNC com um mecanismo de atenção, o objetivo do presente artigo é decodificar as expressões não verbais presentes no banco de dados utilizado e identificar a quais emoções estão ligadas. Através da análise do CCC (Coeficiente De Correlação De Concordância) e do Erro Quadrático Médio (RMSE) para as dimensões de *valence* e *arousal*, o presente artigo mostra que o método utilizado traz resultados, mas ainda é possível melhorar o aprendizado de máquina.

Palavras-chave: *Valence*. *Arousal*. *Feedback* não verbal. Redes neurais convolucionais.

ABSTRACT

Non-verbal feedback and the recognition of facial expressions have been an area of much research in the last decades. Facial expressions are a concrete way to recognize emotions and "teaching" computers to detect correctly what each facial expression means and to which emotion it is attached. Thus, in the scope of image recognition, Convolutional Neural Networks (CNN), through their layering over image pixels, facilitate pattern discovery. Therefore, through the application a CNN with an attention mechanism, the objective of this paper is to decode the non-verbal expressions present in the used database and identify to which emotion it is linked. Through the analysis of the CCC (Correlation Coefficient of Concordance) and the

Mean Squared Error (RMSE) for the valence and arousal dimensions, this paper shows that the method used brings results, but there is still room for improvement in machine learning.

Keywords: Valence. Arousal. Non-verbal feedback. Convolutional neural networks.

1 INTRODUÇÃO

Estudos relacionados à análise e reconhecimento das expressões faciais receberam grande ênfase nos últimos anos. Visto que as expressões da face são capazes de indicar o estado emocional de um indivíduo e fornecer *feedback* no processo de comunicação, sua decodificação é de suma importância para as interações sociais. De acordo com Freitas-Magalhães (2018, p.37), "a emoção manifesta-se através dos processos quinésicos e cinéticos no âmbito do repertório da comunicação não verbal" e seu reconhecimento através das expressões faciais é um processo complexo.

Outro ponto extremamente relevante para o processo de interação humana é o *feedback* não verbal. Uma evidência disso é a existência de áreas e células cerebrais que são responsáveis pelo mapeamento e reconhecimento desse tipo de contexto. Essas áreas são capazes de codificar as intenções de indivíduos baseando-se parcialmente em sua leitura corporal. Sendo assim, é possível afirmar que diversas partes do corpo contribuem para o *feedback* não verbal durante interações (PAULISTA, 2009).

Diante disso, o presente artigo propõe um estudo acerca dos métodos de reconhecimento das expressões faciais. Assim, através da aplicação de algoritmos que fazem a decodificação das expressões faciais, objetiva-se identificar quais emoções o interlocutor transmite e estimar o *feedback* não verbal.

Para melhor compreensão do tema e dos resultados alcançados, este estudo apresenta alguns trabalhos correlatos no reconhecimento de expressões faciais na subseção 1.1. A segunda seção explicita os conceitos e a importância do feedback não verbal e das redes neurais. Em seguida, é apresentada a metodologia de pesquisa utilizada para a construção deste estudo. A quarta seção, por sua vez, traz os resultados obtidos. Por fim, a quinta seção mostra as conclusões alcançadas. Assim, encaminhamo-nos para as considerações acerca da descrição ora apresentada.

1.1 Trabalhos correlatos

Dado o foco desse artigo, esta subseção aborda trabalhos relacionados ao reconhecimento e decodificação das emoções através de expressões faciais. O artigo de Minaee, Minaei e Abdolrashidi (2021) se concentra na adição de mecanismos de atenção que focam na análise de partes importantes do rosto para o reconhecimento das emoções, já que segundo os autores nem todos os aspectos da face são relevantes para a detecção de uma emoção específica. A proposição dos autores é a de que é possível alcançar resultados favoráveis empregando uma rede convolutiva com menos de dez camadas e utilizando o conceito de atenção (quando o treinamento parte do zero). No caso da arquitetura proposta pelo artigo em questão, esta extração de características relevantes consiste na construção de 4 camadas convolutivas, sendo que após a segunda e a quarta camada há também a inserção de uma camada de *max-pooling* e de uma unidade linear retificada (*rectified linear unit* ou ReLU) como função de ativação. De acordo com as conclusões obtidas, verificou-se que redes neurais com menos de dez camadas podem apresentar (ou até superar) resultados satisfatórios no reconhecimento de emoções quando comparadas a redes mais profundas.

Outro trabalho relacionado ao tema e que corrobora as conclusões obtidas pelo trabalho citado no parágrafo anterior é o de Siqueira, Magg e Wermter (2020). Os autores afirmam que quando consideramos o contexto de aprendizagem profunda, um conjunto de redes muito profundas podem se tornar ineficientes (devido à grande redundância) e dispendiosas. No caso do trabalho proposto pelos autores, foram feitas experiências em conjuntos com representações partilhadas (*Ensembles with Shared Representations* ou ESR) baseadas em redes convolutivas. O objetivo é demonstrar eficiência e escalabilidade considerando um conjunto de dados de expressões faciais em grande escala. Assim, os autores mostram que a redundância e a carga computacional de redes profundas podem ser drasticamente reduzidas, variando o nível de ramificação do ESR sem perda da diversidade e do poder de generalização, importantes para o desempenho do conjunto. No que tange ao reconhecimento das expressões faciais, por exemplo, iniciar a ramificação muito “cedo” (nível 1) ou “muito tarde” pode prejudicar os resultados obtidos. Justamente por isso, os autores avaliam os efeitos da variação do nível de ramificação quantitativa e qualitativamente.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção traz o embasamento teórico para compreensão dos assuntos tratados nas próximas seções. Nele apresenta-se os conceitos e definições sobre *feedback* não verbal e redes

neurais. Além disso, explana-se a importância e as aplicações do reconhecimento de expressões faciais e a teoria das redes neurais convolucionais.

2.1 *Feedback* não verbal

De acordo com Al Tawil (2019), os relacionamentos interpessoais que formam as comunidades são compostos por dois tipos de comunicação: verbal e não verbal. Enquanto a verbal apresenta-se através dos significados das palavras, a não verbal mostra-se além do significado real das expressões e podem revelar elementos importantes ligados às emoções e atitudes. Esses dois tipos de comunicação não são opostos, mas sim complementares em interações presenciais (AL TAWIL, 2019).

Segundo Liu, Calvo e Lim (2016), esse comportamento não verbal pode ser classificado em quatro categorias: cinésicos, que incluem movimentos de cabeça, expressões faciais e gestos; vocálicos, que se ligam ao volume e tom da fala; hápticos, que se relacionam ao tato e ao contato corporal e por fim, proxêmicos que estão ligados às orientações espaciais (como a distância entre um indivíduo e outro). Nesse trabalho, nos atentaremos apenas à categoria dos comportamentos cinésicos, mais especificamente nas expressões faciais, já que de acordo com Paulista (2009, p. 90), estudos demonstram "que a expressão facial das emoções básicas é igualmente expressada em qualquer cultura".

O termo *feedback* pode ser traduzido como "retroalimentação" e diz respeito à resposta oferecida diante do comportamento ou estímulo de um indivíduo com que se mantém uma comunicação ou relacionamento. Hamond, Himonides e Welch (2021) afirmam que de maneira geral, o *feedback* é um componente crucial no que tange ao potencial de mudança na performance individual. Diante da importância do assunto, este artigo aborda a utilização de algoritmos para a interpretação do *feedback* fornecido através de expressões faciais.

2.2 Reconhecimento de expressão facial

Conforme colocado por Cruz (2019, p. 23), "a expressão facial é uma maneira eficaz para reconhecer emoções, sobretudo por não ser uma abordagem intrusiva de coleta de dados quando comparada aos sensores fisiológicos [...]". Assim, o reconhecimento das expressões faciais torna-se um fator fundamental na comunicação social humana e falhas na sua compreensão podem comprometer todo o processo comunicativo.

Um dos pioneiros no estudo das expressões faciais são Ekman e Friesen (1978) que publicaram um estudo sobre as seis expressões faciais básicas, consideradas universais. São

elas: felicidade, tristeza, surpresa, medo, raiva e nojo. Após a publicação desse estudo, outros foram surgindo, inclusive o das Expressões Faciais Compostas (EFCs).

Ainda de acordo com o estudo de Ekman e Friesen (1978), cada emoção que uma pessoa transmite possui um movimento muscular facial equivalente. Os autores chamam cada um desses movimentos de Unidade de Ação (UA). Através das UAs é possível caracterizar, identificar e diferenciar cada emoção básica. Pedro (2013) ainda acrescenta um conceito importante no que tange às expressões faciais: as dimensões de valência (ou *valence*), que dizem respeito à categoria (positiva ou negativa) da expressão gerada por uma determinada emoção e de *arousal* que se trata do nível de ativação, a intensidade da emoção expressada.

As principais formas de extração de características da imagem do indivíduo são as geométricas e as aparentes. Yu *et al.*, (2016) explana que a extração geométrica é utilizada para capturar as deformações causadas pela ativação dos músculos a partir de pontos chave faciais, como altura das sobrancelhas, altura dos lábios, largura do nariz etc. Já a extração aparente, ainda de acordo com Yu *et al.*, (2016), divide o rosto do indivíduo estudado em sub-regiões e analisa quais as deformações faciais causadas pelas movimentações musculares.

Por fim, Cruz (2019) afirma que o processo de aprendizagem de máquina utilizado para o reconhecimento das expressões faciais, está intimamente ligado ao conhecimento das sub-regiões faciais (provenientes do método de extração aparente), aos padrões manifestados por elas e ao estudo das Redes Neurais Convolucionais (RNC).

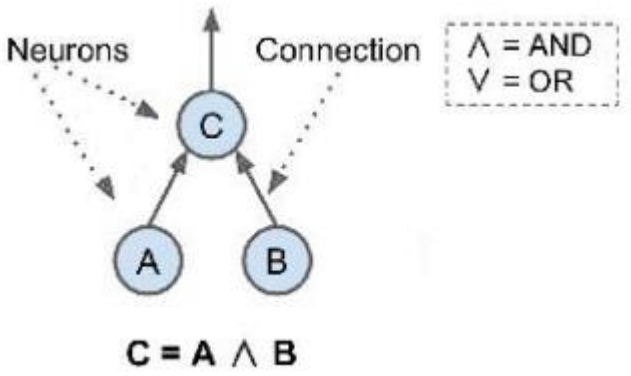
2.3 Redes neurais

Os trabalhos iniciais que sugeriam modelos para aprendizado utilizavam redes que simulavam neurônios no cérebro, portanto chamadas de redes neurais artificiais (RUSSELL E NORVIG, 2021). Ainda, alguns autores entendem que uma rede neural é uma máquina que simula a forma como o cérebro resolve tarefas (HAYKIN, 1998).

Os neurônios artificiais, também conhecidos como unidades, possuem um conjunto de entradas a_j que é ponderada por um peso $w_{i,j}$. A saída de uma unidade é o somatório das entradas multiplicadas por seus pesos, sendo que essa saída pode ser 1 ou 0 dependendo da função de ativação (RUSSELL E NORVIG, 2021). Existem diferentes funções de ativação, sendo logística, *hard threshold* e *rectified linear unit* (ReLU). Um neurônio que possui uma função de ativação com *hard threshold* é conhecido como *Perceptron*. A Figura 1 traz o modelo de um *Perceptron*.

Ainda segundo Haykin (1998) as redes neurais artificiais se assemelham ao cérebro humano pois o conhecimento adquirido é oriundo do ambiente por meio de um processo de aprendizado e a força das conexões entre neurônios, conhecidas como pesos sinápticos, são usadas para armazenar o conhecimento adquirido.

Figura 1 - Modelo de *Perceptron*



Fonte: Adaptado de Géron (2017)

Na Figura 1 é possível visualizar um conjunto de unidades que executa uma lógica E (*AND* em inglês): o neurônio C é ativado apenas quando ambos os neurônios A e B estão ativados, já que um sinal de entrada único não é suficiente para ativá-lo (GÉRON, 2017).

Outra definição dada para o *Perceptron* é a de Norvig e Russell (2021, p.837):

> *Perceptron* é uma rede neural de uma única camada com uma função de ativação explícita (*hard threshold*), popularizado por Frank Rosenblatt (1957). Após uma demonstração em julho de 1958, o jornal The New York Times o descreveu como "o embrião de um computador eletrônico que (a Marinha) espera que seja capaz de andar, falar, ver, escrever, reproduzir-se e ter consciência de sua existência".

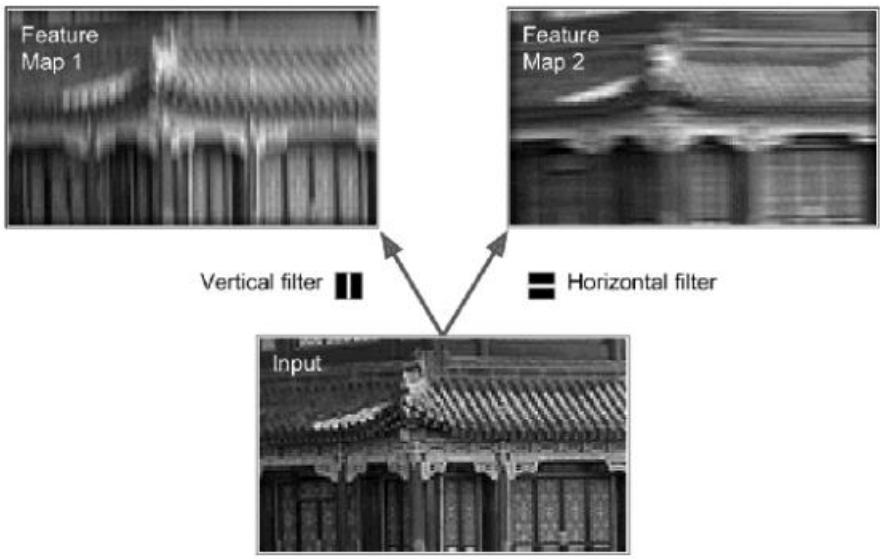
2.3.1 Redes neurais convolucionais (RNC)

De acordo com Skansi (2018), as redes neurais artificiais foram criadas por Lecun e outros em 1998, inspirando-se nas ideias de David H. Hubel e Torsten Weisel apresentadas em um seminário, vencedor do prémio Nobel de 1981 em fisiologia e medicina, que explorava o córtex visual animal. Apesar das RNCs não servirem apenas para reconhecimento de imagem, é através dela que o conceito de visão computacional é desenvolvido atualmente.

Segundo Kattenborn *et al.* (2021), RNCs são baseadas em neurônios que são organizados em camadas e podem, portanto, aprender representações hierárquicas. Os neurônios entre as camadas são conectados por pesos e vieses. A camada inicial é a camada de entrada e a última camada, a de saída. No meio estão as camadas ocultas que transformam as características dos recursos de entrada de uma maneira que correspondam à saída.

Ainda segundo Kattenborn *et al.* (2021), as RNCs são projetadas para aprender características espaciais como bordas, cantos ou formas mais abstratas. A chave para aprender essas características são transformações múltiplas e sucessivas dos dados de entrada (convoluções) em diferentes escalas espaciais (por exemplo, por meio de operações de agrupamento). A Figura 2 ilustra como uma mesma imagem de entrada pode ser usada para reconhecimento de dois padrões diferentes (retas verticais e retas horizontais, da esquerda para a direita), e assim, produzir resultados diferentes.

Figura 2 – Aplicação de filtros para obtenção de mapeamentos de característica



Fonte: Géron (2017)

Como observa-se na Figura 2, há dois resultados diferentes oriundos de uma mesma entrada (*input*). O que distingue as duas saídas é o filtro usado, sendo um filtro de linha vertical na imagem de saída da esquerda e um filtro de linha horizontal na imagem da direita.

3 MATERIAIS E MÉTODOS

Esta seção aborda os procedimentos metodológicos utilizados para a construção deste artigo e quais os materiais empregados, detalhados a seguir.

3.1 Metodologia

Este trabalho apresenta-se com o objetivo de identificar expressões faciais e estimar o *feedback* não verbal (positivo e negativo). Assim, inicialmente fez-se um levantamento bibliográfico, através de estudos e artigos que tratavam sobre a importância do *feedback* não verbal, o reconhecimento de expressões faciais e a supressão das incertezas na sua detecção e a relevância das redes neurais convolucionais no processamento e análise de imagens.

Já que se trata de uma pesquisa de natureza aplicada, após esse embasamento teórico, aplicou-se a RNC contendo um mecanismo de atenção que foi definida em Wen *et al.* (2022), chamada de DAN. Na abordagem do trabalho principal são utilizadas três redes com responsabilidades diferentes, *Feature Clustering Network* (FCN), *Multi-head cross attention networks* (MAN) e *Attention Fusion Network* (AFN). A ideia geral por trás da proposta original é que os mapas atencionais sofram distorções antes do processo de extração de características.

3.2 Materiais

Inicialmente, o trabalho utilizou o banco de imagens AffWild2, composto por inúmeras imagens, sendo cada uma um único frame de um respectivo vídeo. Cada imagem continha seus valores para as dimensões de *valence* e *arousal* pré-determinados.

A rede DAN foi responsável pela análise das imagens e como resultado forneceu um arquivo (formato .csv) com a própria estimativa de *valence* e *arousal* das imagens. Através destes valores estimados e dos valores pré-definidos do banco de imagens, foram aplicadas duas métricas de avaliação de desempenho, o CCC (*Concordance Correlation Coefficient*), que mede a acurácia e a consistência da mesma em outras predições, e o RMSE (Root *Mean Square Error*), que mede o quão distante o valor estimado (resultado do algoritmo) está do valor verdadeiro (valores do banco de imagens). Ambas as métricas têm como resultado um valor de ponto flutuante entre -1 e 1, tendo o CCC um valor mais próximo de 1 como ótimo e RMSE tendo o resultado mais próximo de 0 como um bom resultado.

4 RESULTADOS E DISCUSSÃO

Através da análise do CCC e do RMSE para as dimensões de *valence* e *arousal*, observa-se que os resultados obtidos ainda ficam longe de ser exemplares. No caso do CCC – que deve estar o mais próximo possível de 1 – as médias alcançadas pelo algoritmo foram muito baixas (tanto para *valence* quanto para *arousal*), o que indica resultados considerados pobres. A Tabela 1 apresenta a média dos resultados obtidos.

Tabela 1 – Média dos resultados de CCC e RMSE

Componente emocional	***CCC***	***RMSE***
Valence	0,0048	0,48202
Arousal	-0,0104	0,32081

Fonte: Autoria Própria

Como a Tabela 1 evidencia, no que tange ao RMSE, sabe-se que apesar de quase nunca alcançado na prática, o resultado zero indicaria um ajuste perfeito aos dados. No trabalho em questão, as médias dos resultados para *valence* e *arousal* ficaram próximas de 0,5, que indica algum aprendizado, mas não suficientemente bom.

5 CONSIDERAÇÕES FINAIS

O reconhecimento das expressões faciais compõe sistemas cognitivos que podem ser aplicados em diversas áreas. Assim, este artigo apresenta-se como uma pesquisa de natureza aplicada que objetiva identificar expressões faciais e decodificar aspectos ligados ao *feedback* não verbal.

Inicialmente, este trabalho utilizou imagens do banco de dados AffWild2 e a RNC DAN para detecção de expressões. Logo após, houve a comparação dos resultados corretos com as respostas alcançados pelo algoritmo no reconhecimento das expressões. Através da análise do CCC e do RMSE para as dimensões de *valence* e *arousal*, observa-se que os resultados obtidos ainda ficam longe de ser excelentes. No caso do CCC, as médias alcançadas pelo algoritmo foram muito baixas, o que indica resultados considerados insatisfatórios. No que tange ao RMSE, sabe-se que apesar de quase nunca existir um ajuste perfeito aos dados, as médias alcançadas para *valence* e *arousal* ficaram próximas de 0,5.

Portanto, visto que o presente artigo se limitou à aplicação da RNC DAN para análise do banco de imagens, conclui-se que os resultados obtidos ainda se distanciam dos ideais. Sugere-se que em trabalhos futuros sejam realizados estudos a fim de aumentar a acurácia na detecção das expressões faciais e averiguar a qualidade dos *feedbacks* não verbais.

REFERÊNCIAS

AL TAWIL, R. Nonverbal Communication in Text-Based, Asynchronous Online Education. **The International Review of Research in Open and Distributed Learning**, v. 20, n. 1, 28 fev. 2019.

CRUZ, A. A. da. **Uma abordagem para reconhecimento de emoção por expressão facial baseada em redes neurais de convolução.** 2019. 120 f. Dissertação (Mestrado em Informática) – Universidade Federal do Amazonas, Manaus, 2019. Disponível em: <https://tede.ufam.edu.br/handle/tede/7320>. Acesso em: 04 abr. 2022.

EKMAN, P.; FRIESEN, W.V. **Facial action coding system:** a technique for the measurement of facial movement. Palo Alto, 1978.

FREITAS-MAGALHÃES, A. **O código de Ekman:** o cérebro, a face e a emoção. Porto, Portugal: Editora Escrytos, 2018. 443 p.

GÉRON, A. **Hands-on machine learning with scikit-learn and tensorflow:** Concepts. Tools, and Techniques to build intelligent systems. Editora O'Reilly, 2017.

HAMOND, L.; HIMONIDES, E.; WELCH, G. A natureza do feedback no ensino e na aprendizagem de piano com o uso de tecnologia digital no ensino superior. **Orfeu**, Florianópolis, v. 6, n. 1, p. 01-31, 2021. DOI: https://doi.org/10.5965/2525530406012021e0011. Disponível em: <https://www.revistas.udesc.br/index.php/orfeu/article/view/19928>. Acesso em: 01 abr. 2022.

HAYKIN, S. O. **Neural networks:** a comprehensive foundation. 2ª Edição. Editora Pearson Education, 1998. 842 p.

KATTENBORN, T. *et al*. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 173, p. 24-49, 2021.

LIU, C.; CALVO, R. A.; LIM, R. Improving medical students' awareness of their non-verbal communication through automated non-verbal behavior feedback. **Front. ICT**, 2016. DOI: 10.3389/fict.2016.00011. Disponível em: <https://www.frontiersin.org/articles/10.3389/fict.2016.00011/full>. Acesso em: 01 abr. 2022.

MINAEE, S.; MINAEI, M.; ABDOLRASHIDI, A. **Deep-Emotion:** Facial Expression Recognition Using Attentional Convolutional Network. Sensors (Basel). 2021. Disponível em: <https://paperswithcode.com/paper/deep-emotion-facial-expression-recognition>. Acesso em: 26 ago. 2022.

PAULISTA, G. da P. **Incorporando meta learning:** o papel crítico da expressão não-verbal na interação face a face e na performance de equipes de trabalho. 2009. Tese (Doutorado em Engenharia e Gestão do Conhecimento) – Universidade Federal de Santa Catarina, Florianópolis, 2009. Disponível em: <https://repositorio.ufsc.br/handle/123456789/103237>. Acesso em: 07 mar. 2022.

PEDRO, T. M. J. **Alexitimia e avaliação da valência e arousal de expressões emocionais.** 2013. Dissertação (Mestrado em Psicologia Clínica e da Saúde) – Universidade de Aveiro, 2013. Disponível em: < http://hdl.handle.net/10773/12764>. Acesso em: 04 abr. 2022.

RUSSEL, S.; NORVIG, P. **Artificial Intelligence:** a modern approach. 4ª Edição. Editora Pearson Education, 2021. 1115 p.

SKANSI, S. **Introduction to Deep Learning:** from logical calculus to artificial intelligence. Editora Springer, 2018.

SIQUEIRA, H.; MAGG, S.; WERMTER, S. (2020). **Efficient Facial Feature Learning with Wide Ensemble-Based Convolutional Neural Networks.** Proceedings of the AAAI Conference on Artificial Intelligence. Disponível em: <https://paperswithcode.com/paper/frame-attention-networks-for-facial>. Acesso em: 28 ago. 2022.

YU, *et al*. Customized expression recognition for performance-driven cutout character animation. **Applications of Computer Vision (WACV)**, 2016, IEEE Winter Conference on, p. 1–9.

WEN, Z. *et al*. **Distract Your Attention:** Multi-head Cross Attention Network for Facial Expression Recognition. Disponível em: <https://arxiv.org/pdf/2109.07270v4.pdf>. Acesso em: 28 set. 2022.