Facial Expression Recognition to Aid Visually Impaired People

João Marcos Silva[1], Romuere Silva[1], Rodrigo Veras[1], Kelson Aires[1], Laurindo Britto Neto[1]
[1]*Department of Computing – Federal University of Piauí*, Teresina, Brazil
{jms, romuere, rveras, kelson, laurindoneto}@ufpi.edu.br

***Abstract*—Facial expression recognition systems can help a visually impaired person to identify the emotions of the person with whom she interacts, assisting in her non-verbal communication. Among the various researches carried out in recent years on recognition of facial expressions, the best results obtained come from methods that use deep learning, mainly with the use of convolutional neural networks. This work presents a literature review on the problem of recognition of facial expressions, through the use of convolutional neural networks and proposes two approaches in which the first one uses pre-trained CNN models together with the Linear SVM classifier that, applied to the bases CK+ and JAFFE data, obtained maximum accuracy of 89.6% and 95.7%, respectively. And in the second approach, a CNN model built from scratch is used with the CK+ and FER2013 databases, which obtained accuracy rates of 85% and 65.8%, respectively.**

***Index Terms*—facial expression recognition, convolutional neural network, deep learning, visually impaired people**

I. Introdução

Expressões faciais compreendem uma importante ferramenta na comunicação não verbal. Segundo Mehrabian e Russell [1], no processo de diálogo interpessoal, as palavras são responsáveis por apenas 7% da informação transmitida, a fala oral por 38%, e as expressões faciais e os movimentos corporais por 55%. Isso torna a interação social uma das grandes dificuldades na vida de uma pessoa com deficiência visual. Um dos problemas mais relevantes enfrentado é a incapacidade de identificar as emoções da pessoa com quem ele está interagindo [2], dado que emoções humanas são refletidas principalmente por meio de expressões faciais. Dentro desse contexto, um sistema de reconhecimento de expressões faciais (*Facial Expression Recognition – FER*) pode auxiliar nesse problema.

Reconhecimento de expressões faciais vem sendo muito pesquisado ultimamente na área de visão computacional. Diversas abordagens e técnicas foram propostas, com intuito de melhorar as taxas de acurácia e resolver problemas que estão presentes em cenários reais, como variação de iluminação, pose da cabeça, oclusão de parte da imagem. Dentre as diversas pesquisas realizadas nos últimos anos sobre reconhecimento de expressões faciais, os melhores resultados advêm de métodos que utilizam *deep learning*, mais precisamente, das Redes Neurais Convolucionais *(Convolutional Neural Network – CNN)* [3], que são provavelmente o modelo de aprendizado profundo mais usado para resolver problemas de visão computacional [4].

Este trabalho tem o apoio da Fundação de Amparo a Pequisa do Piauí, por meio do Edital FAPEPI/MCT/CNPq Nº 007/2018 (PPP) convênio FAPEPI/CNPq.

O objetivo deste trabalho é propor uma abordagem em reconhecimento de expressões faciais para auxiliar pessoas com deficiência visual, baseada em CNN. Tal abordagem será incorporada a um sistema *wearable* [5], [6], usando uma arquitetura cliente-servidor, de forma que economize recursos de hardware como memória, bateria e processamento.

Este trabalho está organizado da seguinte forma: na Seção II é realizada a revisão de literatura, contendo uma breve descrição de cada trabalho e uma discussão sobre os resultados da revisão; na Seção III são descritos os materiais e métodos utilizados, tais como, as bases de dados e a metodologia experimental; a Seção IV apresenta as abordagens propostas; na Seção V, os resultados e discussão; e, finalmente, a Seção VI conclui este trabalho e descreve os trabalhos futuros.

II. Revisão da Literatura

Em pesquisas recentes, Li e Deng [7] apresentaram uma revisão da literatura, em que examinaram o estado da arte sobre reconhecimento de expressões faciais utilizando *deep learning*. Além de redes neurais convolucionais, eles também abordaram outras técnicas de aprendizado profundo, como *Deep Belief Network* (DBN), *Recurrent Neural Network* (RNN), *Generative Adversarial Network* (GAN) etc. A revisão feita neste trabalho será uma atualização e complemento da revisão feita por Li e Deng [7], em que serão abordadas pesquisas a partir do ano de 2018, porém priorizando trabalhos que utilizam métodos baseados em CNN's.

Sajjanhar et al. [8] realizaram experimentos utilizando um modelo CNN construído a partir do zero, e utilizando as CNN's pré-treinadas Inception [9], VGG (VGG-16 e VGG-19) [10] e VGG-Face [11]. Tais experimentos foram realizados para avaliar o desempenho desses modelos sobre o problema de classificação de expressões faciais, utilizando a técnica de Transferência de Aprendizagem [12]. O modelo proposto é treinado e testado com três conjuntos de imagens resultantes do pré-processamento das imagens originais. O resultado é comparado para três bancos de dados diferentes: CK+ [13], JAFFE [14] e FACES [15] e as acurácias foram de 85,19%, 65,17% e 84,38%, respectivamente. Para o experimento utilizando os modelos pré-treinados, que utilizaram as mesmas bases de dados, mostraram que o modelo VGG-19, apesar de ser treinado para reconhecimento de objetos, possui um

resultado melhor que o VGG-Face, que é treinado para reconhecimento de face nas bases JAFFE e FACES.

Wu et al. [16] propuseram um novo método, que visa o reconhecimento de expressões faciais em diferentes poses, com base na detecção de pontos de referência. O modelo consiste em duas redes compartilhadas, no qual a primeira é utilizada para detectar 29 pontos faciais focados, principalmente, nas partes da sobrancelha e boca, pois essas áreas podem representar melhor diferentes expressões. Elas são entrada para a segunda rede, que utiliza os métodos *RoIAlign* e concatenação de mapas de características para o reconhecimento da expressão facial. Para verificar a capacidade do modelo para resolver o problema, foi utilizado o banco de dados CASIA-MFE [16], que contém amostras com diferentes poses. Além disso, foi testado o desempenho em bancos de dados mais populares CK+, MMI [17] e Oulu-CASIA [18]. No banco de dados CASIA-MFE, foi obtida uma acurácia de 95,97%, bastante superior em comparação a outros métodos.

Georgescu et al. [19] combinaram características aprendidas por CNN's e características *handcrafted*, obtidas por meio do modelo *Bag-of-Visual-Words* (BoVW) [20]. O modelo consiste em um *pipeline* formado pelo modelo BoVW, em conjunto com três CNN's ajustadas (*fine-tuning*) dos modelos VGG-Face, VGG-F [21] e VGG-13 [22]. Todas são treinadas utilizando o método *Dense-SparseDense* (DSD) [23], a fim de evitar *overfitting*. As características extraídas de todos os modelos são concatenadas e normalizadas usando *l2-norm*. Eles utilizaram aprendizagem local, com *K-Nearest Neighbors* (K-NN) [24] e *Support Vector Machines* (SVM) [25] para classificação local, pois o classificador Linear SVM [26], na estrutura de aprendizagem local, se torna não-linear. Foram realizados experimentos nos bancos de dados FER2013 [27], FER+ [22] e AffectNet [28], obtendo acurácias máximas de 75,42%, 87,76% e 63,31%, respectivamente. Os resultados foram comparados aos modelos usados de forma isolada, e a abordagem proposta obteve resultados superiores.

Kim et al. [29] propuseram um modelo hierárquico constituído de duas redes neurais convolucionais. A primeira é baseada em aparência e utiliza como entrada imagens pré-processadas com LBP. A segunda foi baseada em características geométricas. Eles demonstraram que o erro de reconhecimento da expressão facial ocorre mais frequentemente na emoção com a segunda maior probabilidade ao usar apenas a rede de aparência. Seguindo essa premissa, o modelo combina as duas saídas de maior probabilidade resultantes da função *Softmax* com a segunda rede, a qual, com base em unidade de ações faciais (AU's), calcula a diferença da imagem com o pico de emoção e da imagem neutra gerada por um *autoencoder*, também proposto na pesquisa, e permite uma reclassificação a fim de melhorar a precisão dos resultados. O modelo foi testado nas bases de dados CK+ e JAFFE, com validação cruzada *10-fold* em ambos, e conseguiram altas taxas de acurácia, 96,46% e 91,27%, respectivamente. O desempenho é comparado com Xie e Hu [30], que obtiveram 93,46% de acurácia na base CK+. É também comparado com a abordagem proposta por Lopes et al. [31], que obtiveram 84,48% de acurácia na base JAFFE.

Ul Haque e Valles [32] construíram um modelo *Deep Convolutional Neural Network* (DCNN) baseado na arquitetura VGG-16. Em relação a trabalhos anteriores, a fim de melhorar a precisão e o desempenho, foram modificados parâmetros do modelo, como a taxa de aprendizagem, o tamanho de lote, o número de épocas e a taxa de *dropout*. O modelo foi testado em versões modificadas do banco de dados KDEF com diferentes condições de iluminação. Além disso, esse banco de dados possui imagens com diferentes poses de cabeça, o que dificulta no reconhecimento. A acurácia máxima obtida com o conjunto de dados original, sem modificações de iluminação, foi de 86,44%, e a acurácia mínima obtida na modificação do conjunto de dados com imagens mais escuras foi de 75,27%. Tais resultados foram superiores a trabalhos anteriores em que atingiram 78,32% e 46,5%, respectivamente.

Zou et al. [33] desenvolveram um modelo CNN bastante simples. A rede possui três camadas convolucionais, duas de *Pooling* [34], uma densa e, por fim, a camada *Softmax* [34] para a classificação das imagens. Utilizaram Normalização de Lote [35] para evitar desaparecimento de gradiente, além da técnica de *Dropout* [36] com taxa de 0,5, a fim de evitar o *overfitting* [37]. Para a realização do experimento utilizaram o banco de dados CK+. Foi feita a detecção facial nas imagens e redimensionamento para o tamanho de 96×96 pixels. O resultado obtido foi de 99,14%. Comparando com os modelos AlexNet [38] e VGG19, o modelo proposto possui uma acurácia 6,9% e 4,07% maior, respectivamente.

Hu et al. [39] propuseram uma arquitetura inspirada no modelo DenseNet [40]. A rede consiste na fusão de características em vários níveis. É constituída de 7 blocos de fusão de características, em que cada bloco consiste em 4 unidades de extração de características de multi-granularidade, que são conectadas entre si por múltiplos caminhos. As vantagens são a melhoria da eficiência do treinamento e melhor representação das características. Para realização dos experimentos utilizaram as bases de dados CK+ e FER2013. Os pré-processamentos realizados foram: detecção facial, equalização de histograma e alinhamento facial. Obtiveram 94,07% no banco de dados CK+ e 65,4% no banco de dados FER2013. Os resultados foram comparados com a rede DenseNet que obteve 92,68% de acurácia na base CK+ e 63,9% na base FER2013.

A. Discussão Sobre a Revisão de Literatura

Nos trabalhos selecionados, os autores buscam, além de melhorar o desempenho das redes e obter altas taxas de acurácia, contornar problemas que ainda são persistentes no reconhecimento de expressões faciais, tais como, a variação de iluminação, diferentes poses de cabeça, variação de intensidade etc., ou seja, condições que facilmente são encontradas em ambientes reais. Eles também buscam evitar situações comuns no contexto de *deep learning*, como o *overfitting*, situação em que o modelo se adequa aos dados de treinamento e possui certa dificuldade de classificar dados novos, ou seja, os dados de teste. E, por fim, procuram contornar o problema

de se ter dados de treinamento insuficientes, pois modelos profundos possuem muitos parâmetros a serem ajustados, exigindo grande quantidade de imagens para que o modelo possa generalizar de forma a classificar corretamente imagens nunca vistas.

Abordagem de Ul Haque e Valles [32] foca no problema de variação de iluminação realizando testes com imagens em diferentes condições de iluminação. Wu et al. [16] visam obter bons resultados, concentrando-se no problema de diferentes poses de cabeça e identificando pontos de referências espaciais. Para o problema de dados insuficientes, uma técnica utilizada por diversos trabalhos, é o processo de *fine-tuning* [41], que consiste em utilizar redes pré-treinadas com os pesos já ajustados e adaptar para o problema em questão. Essas redes pré-treinadas, por exemplo, Inception, VGG e DenseNet, foram treinadas utilizando o banco de dados do concurso *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* [42], constituído de mais de 14 milhões de imagens. Além disso, 54,4% dos trabalhos utilizam a técnica de Aumento de Dados [43], que consistem em, a partir das imagens existentes, realizar alterações como rotações, translações, adição de ruído etc., a fim de criar novas imagens, permitindo que o modelo tenha uma melhor generalização.

Sobre as bases de dados utilizadas, destaca-se o Extended CohnKanade (CK+) por ser a mais utilizada (59% dos trabalhos). As abordagens que realizaram experimentos com essa base, foram as que obtiveram maior taxa de acurácia. Isso se justifica devido o banco possuir imagens controladas e sem oclusão ou variação de pose de cabeça.

III. Materiais e Métodos

Nesta seção serão abordadas, com uma breve descrição, as bases de dados utilizadas e metodologia experimental empregada.

A. Bases de Dados

Nesta seção são descritos os bancos de dados mais utilizados no reconhecimento de expressões faciais. Diversos bancos de dados estão disponíveis, dentre eles existem os constituídos de uma única imagem para cada expressão, como JAFFE, KDEF, FER2013, RAF-DB, AffectNet etc., e bancos formados por sequência de imagens, que representam a transição da expressão neutra para o ápice da expressão do indivíduo: CK+, CASME II, MMI, oulu-CASIA etc. Serão detalhados nesta pesquisa apenas CK+, JAFFE e FER2013, pois eles são os mais amplamente utilizados e estão disponíveis publicamente.

1) Extended CohnKanade (CK+): Extended CohnKanade (CK+) [13] é um dos bancos de dados mais utilizados em FER. Ele contém 593 sequências de imagens de 123 indivíduos, com idades entre de 18 à 50 anos. Essas sequências variam de 10 a 60 *frames*, em que o último *frame* representa o pico de expressão. Das 593 sequências de imagens, apenas 327 são rotuladas com as sete expressões faciais (raiva, desprezo, nojo, felicidade, tristeza e surpresa) baseadas no Sistema de Codificação de Ação Facial (FACS). Por ser uma banco de dados de sequência de imagens, geralmente, para métodos baseados em imagem estática, se utiliza o último *frame* da sequência.

2) Japanese Female Facial Expression (JAFFE): Japanese Female Facial Expression (JAFFE) [14] é um banco de dados que contém 213 imagens de 10 mulheres japonesas. As imagens exibem sete expressões faciais (raiva, nojo, medo, felicidade, tristeza, surpresa e neutro). As fotos foram tiradas no Departamento de Psicologia da Universidade de Kyushu. Por ser um banco de dados que possui pouca quantidade de imagens, ele se torna desafiador.

3) Facial Expression Recognition 2013 (FER2013): O banco de dados Facial Expression Recognition 2013 (FER2013) [27] foi introduzido durante o *workshop ICML 2013 Challenges in Representation Learning*. Foi criado usando a API de pesquisa de imagens do Google utilizando palavras chaves correspondentes às expressões faciais. Por serem obtidas da internet, as imagens possuem condições que dificultam o reconhecimento, como oclusão, diferentes poses de cabeça etc. O banco de dados contém 35.887 imagens, sendo 28.709 imagens de treinamento, 3.589 imagens de validação e 3.589 imagens de teste rotuladas com sete expressões (raiva, nojo, medo, felicidade, tristeza, surpresa e neutro).

B. Metodologia Experimental

Para a realização dos testes foram utilizadas as três bases de dados mais utilizadas na literatura: CK+, JAFFE e FER2013. Devido ao fato das bases CK+ e JAFFE possuírem pouca quantidade de imagens, optou-se por utilizar redes pré-treinadas para os testes. Usar redes pré-treinadas traz vantagens, pois além de não ser necessário possuir grandes quantidades de dados, reduz drasticamente o tempo e o poder computacional exigido. Tais redes servirão como extrator de características, as quais seguirão para o classificador Linear SVM [26]. As redes escolhidas foram: VGG-16, treinada com imagens do banco de dados VGG-Face [11]; VGG-16, VGG-19, DenseNet, Inception e MobileNet [44], treinadas com imagens do banco de dados ImageNet [42]; e, por fim, foram realizados experimentos com um modelo CNN treinado do zero.

1) Experimentos com modelos pré-treinados: Neste experimento foram utilizadas as bases CK+ e JAFFE. Não foi possível realizar testes na base FER2013, devido a limitação de memória RAM da plataforma Google Colab. Como a base CK+ é constituída de sequências de imagens, apenas a última imagem de cada sequência é utilizada, pois é o *frame* em que se encontra o ápice da expressão.

O Modelo VGG-16, disponível na biblioteca keras-vggface, foi treinado usando o banco de dados VGG-Face, que contém 2,6 milhões de imagens de rostos. Os outros modelos utilizados foram treinados com imagens do banco de dados ImageNet, que consiste em mais de 14 milhões de imagens. Porém, são imagens de milhares de objetos e cenas diferentes, ao contrário do VGG-Face que contém apenas rostos. As camadas totalmente conectadas, que servem para classificação, são removidas e, então, a rede é utilizada como extrator de características.

TABELA I
DESEMPENHO DOS MODELOS PRÉ-TREINADOS (PRIMEIRA ABORDAGEM).

Banco de Dados	Modelo	Pesos	Acurácia (%)
CK+	VGG-16	VGG-Face	**89,6**
	VGG-16	ImageNet	81,9
	VGG-19		83,6
	DenseNet		84,4
	Inception		78,8
	MobileNet		81,6

Banco de Dados	Modelo	Pesos	Acurácia (%)
JAFFE	VGG-16	VGG-Face	**95,7**
	VGG-16	ImageNet	90
	VGG-19		93,4
	DenseNet		89,7
	Inception		86,2
	MobileNet		93,3

2) Experimentos com modelo CNN: Além da utilização dos modelos pré-treinados, também foi desenvolvida uma CNN para ser treinada do zero. Diferente da utilização dos modelos pré-treinados, essa rede é responsável pela extração de características e pela classificação das imagens. Neste experimento foram utilizadas as bases de dados CK+ e FER2013. Da mesma forma do experimento anterior, na base CK+, é utilizado a última imagem de cada sequência que representa o pico da expressão.

IV. ABORDAGENS PROPOSTAS

A primeira abordagem proposta (Abordagem 1) consiste na utilização de modelos CNN pré-treinados. A arquitetura proposta pode ser vista na Fig. 1. As camadas totalmente conectadas do modelo pré-treinado, que são utilizadas para classificação, são removidas, pois o modelo em questão irá ser utilizado apenas como extrator de características. As características seguem para o classificador Linear SVM [26], que será responsável pela classificação das imagens em expressões faciais. Os pré-processamentos utilizados em ambas bases foram a detecção facial, por meio do algoritmo Viola-Jones [45], e redimensionamento das imagens para o tamanho de 224×224 pixels (tamanho de entrada padrão dos modelos pré-treinados). O tamanho original das imagens da base CK+ é de 640×490 pixels, e da base JAFFE é de 256×256 pixels. Com intuito de melhor avaliar a capacidade de generalização, é utilizada a técnica de Validação Cruzada *K-fold* [46], no qual a base é dividida em K partes, em que uma parte é utilizada para teste e o restante para treino, repetindo o processo K vezes alternando o conjunto de testes. Neste trabalho o valor de K escolhido foi 10.

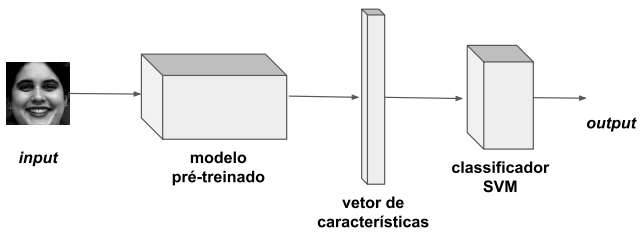

Fig. 1. Arquitetura usada na primeira abordagem, utilizando modelos pré-treinados.

Para a segunda abordagem proposta (Abordagem 2) foi desenvolvido um modelo CNN do zero. A arquitetura do modelo proposto pode ser vista na Fig. 2. Ela é constituída de quatro blocos convolucionais e três camadas totalmente conectadas. Cada bloco inclui duas camadas convolucionais. Após a primeira camada convolucional é adicionada uma camada de normalização de lote, seguida da camada de *Max Pooling* [34], e adicionado a técnica de *Dropout* com taxa de 0,3 para evitar o *overfitting*. Cada camada convolucional utiliza a função de ativação ReLU (*Rectified Linear Units*) [34] para evitar o desaparecimento do gradiente [47]. Logo após isso, as características seguem para as camadas totalmente conectadas. E, por fim, a camada *Softmax*, com sete saídas, é utilizada para a classificação das expressões faciais, em que cada saída é a probabilidade da imagem pertencer a determinada classe.

Na base CK+ foram feitos os seguintes pré-processamentos: detecção facial, redimensionamento para o tamanho de 96×96 pixels e conversão para escala de cinza. Como o treinamento de uma CNN do zero exige uma grande quantidade de dados para uma melhor generalização e para evitar *overfitting*, também foi aplicado a técnica de Aumento de Dados, no qual aplicou-se as seguintes perturbações nas imagens originais: rotação, escala (*zoom*) e deslocamento (*shift*). Foi utilizada Validação Cruzada *10-fold*. Na base FER2013 não foram feitos nenhum tipo de pré-processamento. O método de validação utilizado nessa base foi diferente dos demais, para seguir a mesma metodologia de validação dos trabalhos pesquisados. Nesse caso foram utilizadas, das 35.887 imagens, 28.709 para treino, 3.589 para validação e 3.589 para teste.

V. RESULTADOS E DISCUSSÃO

A Tabela I mostra os resultados dos diferentes modelos para cada base de dados. Em relação aos experimentos com a Abordagem 1 é possível observar pela Tabela I que o modelo VGG-16, pré-treinado com as imagens da base VGG-Face, possui maior taxa de acurácia que os demais em ambas as bases, devido ao fato dessa base ser constituída apenas de imagens de rostos. Na base JAFFE foi obtida acurácia de 95,7%, superando os trabalhos da revisão de literatura que a utilizaram.

No segundo experimento na base CK+, usando a Abordagem 2, foi realizado uma série de pré-processamentos. Esses pré-processamentos tem o intuito de reduzir a complexidade das características a serem aprendidas, além de reduzir o custo

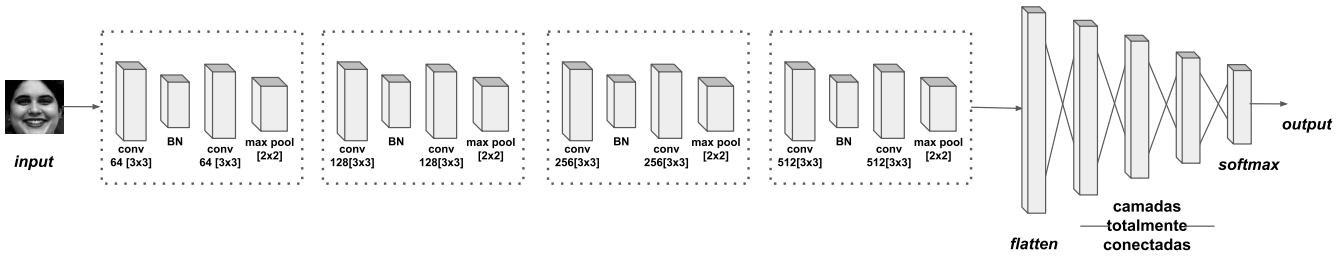

Fig. 2. Arquitetura do modelo CNN utilizado na segunda abordagem.

computacional e o tempo de treinamento. Na base FER2013 nenhum pré-processamento foi utilizado, pois as imagens já estão em escala de cinza e possuem tamanho de 48×48 pixels. Essa base possui certas condições que dificultam o reconhecimento, como diferentes posições de cabeça, oclusões etc. Por esse motivo, alcançar altas taxas de acurácia nessa base é um grande desafio. Como resultado, foi obtido uma taxa de acurácia de 85% na base CK+ e 65,8% na base FER2013.

A Tabela II mostra a comparação dos resultados entre as abordagens propostas e os trabalhos da revisão que usam a mesma base de dados, no caso CK+, JAFFE e FER2013, e a mesma metodologia experimental. É possível observar que para a base de dados CK+, a acurácia obtida é superior apenas à abordagem desenvolvida por Sajjanhar et al. [8]. Porém, a abordagem proposta supera os demais na base JAFFE. Na base FER2013, o modelo desenvolvido foi superior ao trabalho de Hu et al. [39].

TABELA II
COMPARAÇÃO DAS ABORDAGENS PROPOSTAS COM TRABALHOS DA REVISÃO.

Base de Dados	Trabalho	Pré-Processamento		Acurácia (%)
		Detecção Facial	Aumento de Dados	
CK+	[Sajjanhar et al. 2018]	✓		7 classes 85,19
	[Wu et al. 2018]		✓	7 classes 98,22
	[Kim et al. 2019]	✓	✓	6 classes 96,5
	[Zou et al. 2019]	✓	✓	7 classes **99,14**
	[Hu et al. 2020]	✓		7 classes 94,07
	Abordagem 1	✓		7 classes 89,6
	Abordagem 2	✓	✓	7 classes 85
JAFFE	[Sajjanhar et al. 2018]	✓		6 classes 65,17
	[Kim et al. 2019]	✓	✓	6 classes 91,3
	Abordagem 1	✓		7 classes **95,7**
FER2013	[Georgescu et al. 2019]		✓	7 classes **75,42**
	[Hu et al. 2020]		✓	7 classes 65,4
	Abordagem 2			7 classes 65,8

VI. CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho foram propostas duas abordagens. A Abordagem 1 que utilizou modelos pré-treinados, CNN VGG-16 treinado com a base VGGFace, para extração de características e o classificador Linear SVM para a classificação das imagens. Na Abordagem 2 um modelo CNN é desenvolvido e treinado do zero. Foi possível observar que a utilização de modelos pré-treinados trazem mais vantagens e são capazes de obter maior taxa de acurácia em relação a modelos treinados do zero. O modelo MobileNet foi utilizado no primeiro experimento visando trabalhos futuros, pois foi desenvolvida para ser utilizada em aplicações móveis, possuindo baixo tamanho, apenas 16 megabytes. Para trabalhos futuros, pretende-se fazer melhorias no modelo CNN e utilizar outros tipos de pré-processamentos nas bases de dados. Além disso, serão feitas adaptações das abordagens propostas a fim de viabilizar a implementação e utilização em dispositivos *wearable*, que é o objetivo inicial desta pesquisa, e análises em relação ao consumo de recursos de hardware.

REFERÊNCIAS

[1] A. Mehrabian and J. A. Russell, "An approach to environmental psychology," *MIT*, 1974.

[2] A. Ashok and J. John, "Facial expression recognition system for visually impaired," *In: International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI 2018)*, pp. 244–250, 2018.

[3] H. Hakim and A. Fadhil, "Survey: Convolution neural networks in object detection," *Journal of Physics: Conference Series*, vol. 1804, pp. 22–23, 10 2020.

[4] M. A. Ponti, L. S. F. Ribeiro, T. S. Nazare, T. Bui, and J. Collomosse, "Everything you wanted to know about deep learning for computer vision but were afraid to ask," *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, 2017.

[5] L. Britto Neto, V. R. M. L. Maike, F. L. Koch, M. C. C. Baranauskas, A. Rocha, and S. Goldenstein, "A wearable face recognition system built into a smartwatch and the visually impaired user," in *ICEIS*, INSTICC. SciTePress, 2015, pp. 5–12.

[6] L. Britto Neto, F. Grijalva, V. R. M. L. Maike, L. C. Martini, D. Florencio, M. C. C. Baranauskas, A. Rocha, and S. Goldenstein, "A kinect-based wearable face recognition system to aid visually impaired users," *IEEE THMS*, vol. 47, no. 1, pp. 52–64, 2017.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *Computer Vision and Pattern Recognition*, 2018.

[8] A. Sajjanhar, Z. Wu, and Q. Wen, "Deep learning models for facial expression recognition," *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 2018.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 2818–2826, 2016.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *British machine vision conference*, 2015.

[12] D. Silver and K. Bennett, "Guest editor's introduction: Special issue on inductive transfer learning," *Machine Learning*, vol. 73, pp. 215–220, 12 2008.

[13] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete expression dataset for action unit and emotion-specified expression," *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010)*, pp. 94–101, 2010.

[14] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," *3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, 1998.

[15] N. C. Ebner, M. Riediger, and U. Lindenberger, "Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behavior research methods*, pp. 351–362, 2010.

[16] W. Wu, Y. Yin, Y. Wang, X. Wang, and D. Xu, "Facial expression recognition for different pose faces based on special landmark detection," *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018.

[17] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," *in Proc. IEEE Int. Workshop on EMOTION: Corpora for Research on Emotion and Affect*, 2010.

[18] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and vision computing*, p. 607–619, 2011.

[19] M. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64 827–64 836, 2019.

[20] Sivic and Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477 vol.2.

[21] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *BMVC 2014 - Proceedings of the British Machine Vision Conference 2014*, 05 2014.

[22] E. Barsoum, C. Zhang, C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," pp. 279–283, 08 2016.

[23] S. Han, J. Pool, S. Narang, H. Mao, S. Tang, E. Elsen, B. Catanzaro, J. Tran, and W. J. Dally, "DSD: regularizing deep neural networks with dense-sparse-dense training flow," *CoRR*, vol. abs/1607.04381, 2016. [Online]. Available: http://arxiv.org/abs/1607.04381

[24] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds., vol. 17. MIT Press, 2005, pp. 513–520. [Online]. Available: https://proceedings.neurips.cc/paper/2004/file/42fe880812925e520249e808937738d2-Paper.pdf

[25] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[26] R. E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," in *JMLR*, 2008.

[27] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," *in International Conference on Neural Information Processing*, p. 117–124, 2013.

[28] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, Jan 2017.

[29] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE Access*, 2019.

[30] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*, 2019.

[31] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit*, p. 610–628, 2017.

[32] M. I. Ul Haque and D. Valles, "Facial expression recognition using dcnn and development of an ios app for children with asd to enhance communication abilities," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2019, pp. 0476–0482.

[33] J. Zou, X. Cao, S. Zhang, and B. Ge, "A facial expression recognition based on improved convolutional neural network," in *2019 IEEE International Conference of Intelligent Applied Systems on Engineering (ICIASE)*, 2019, pp. 301–304.

[34] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[37] D. M. Hawkins, "The problem of overfitting," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, 2004, pMID: 14741005. [Online]. Available: https://doi.org/10.1021/ci0342472

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, May 2017. [Online]. Available: https://doi.org/10.1145/3065386

[39] Q. Hu, C. Wu, J. Chi, X. Yu, and H. Wang, "Multi-level feature fusion facial expression recognition network," in *2020 Chinese Control And Decision Conference (CCDC)*, 2020, pp. 5267–5272.

[40] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[41] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019–1034, 2015.

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

[43] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, pp. 1–48, 2019.

[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[45] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 05 2004.

[46] P. Refaeilzadeh, L. Tang, and H. Liu, *Cross-Validation*. Boston, MA: Springer US, 2009, pp. 532–538.

[47] R. A. Virrey and L. C. De Silva, "Convolutional neural networks for facial emotion recognition towards the development of automatic pain quantifier," in *7th Brunei International Conference on Engineering and Technology 2018 (BICET 2018)*, 2018, pp. 1–4.