Reconhecimento de Emoções Humanas Através de Expressões Faciais

André Felipe Guimarães
16.2.4045
Departamento de Computação
Universidade Federal de Ouro Preto
Ouro Preto, Brasil
andre.guimaraes@aluno.ufop.edu.br

Gustavo Fonseca Rocha
16.2.4437
Departamento de Computação
Universidade Federal de Ouro Preto
Ouro Preto, Brasil
gustavo.rocha@aluno.ufop.edu.br

I. Introdução

De acordo com Ray Birdwhistell, a comunicação humana é transmitida, em maioria, através da linguagem corporal. [1] A emoção é uma condição que pode desencadear diversos tipos de comportamentos, entre eles expressões faciais. [2] De acordo com o estudo de Darwin, a reação do comportamento corporal frente as emoções foram adquiridas como herança genética e por isto o mesmo comportamento é verificado quando observado entre diversas culturas e civilizações, em alguns casos até entre espécies. [2]

II. Objetivo

O objetivo deste trabalho é mimetizar parte desta herança genética através de técnicas de reconhecimento de padrões, afim de compreender melhor o comportamento humano. Para isto o modelo deverá ser capaz de identificar emoções a partir de imagens de face.

III. Trabalhos Relacionados

A. Artigo 1 - Facial Emotion Recognition System A Machine Learning Approach [3]

Utiliza PSO(Optimização por enxame de partículas) [1] transformativo em conjunto com Algoritmo Microgenético (mGA), denominando PSO embutido em mGA, projetado para atingir o acumulo de aspecto.

O modelo apresentado pela Figura 1, demonstra como o modelo do artigo funciona. Inicialmente, o usuário iniciaria o sistema e faria o acesso ao sistema; Entraria com um *input*, uma imagem de uma expressão facial; O modelo transformaria esta imagem utilizando a técnica PSO embutido em mGA, que retornaria uma imagem binária. E por fim, o modelo classificaria entre: Sorriso, Feliz, Surpresa, Exposição de feição animal, Normal.

Este modelo não apresenta métricas de avaliação do modelo.

[1]A optimização por enxame de partículas é um ramo da inteligência artificial também classificado por alguns autores como um ramo da computação evolucionária, que otimiza um problema iterativamente ao tentar melhorar a solução candidata com respeito a uma dada medida de qualidade.

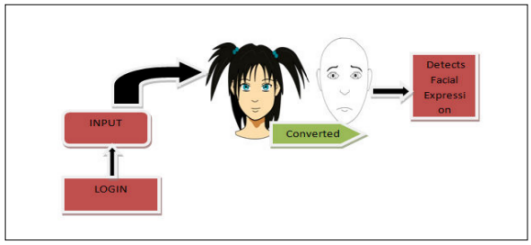

Figura 1. Passos do modelo apresentado pelo artigo. [3]

B. Artigo 2 - Facial Emotion Recognition System through Machine Learning approach [4]

Este artigo se trata de um modelo para reconhecimento automático de emoções a partir expressões faciais através de imagens via *web-cam*. O objetivo é identificar indivíduos estressados, atribuindo-lhes terapia musical para aliviar a condição. As emoções consideradas neste artigo foram: felicidade, tristeza, surpresa, medo, nojo e raiva, que são universalmente aceitas.

O sistema pode ser amplamente categorizado em três estágios: estágio de pré-processamento, estágio de detecção de rosto, estágio de extração de recursos e estágio de classificação de emoções.

A entrada dá imagem é feita a partir de uma imagem obtida por uma *web-cam*. Inicialmente, a etapa de pré-processamento transforma a imagem recebida para binário. É aplicado o algoritmo Viola Jones para reconhecimento do rosto de forma confiável e pouco custosa. Em seguida, como apresentado pela Figura 2, é realizado uma etapa de extração, onde os autores consideram as características mais básicas do rosto, como: ambos olhos, separadamente; nariz; boca. Por fim, o sistema classifica as emoções e reproduz uma música como terapia de acordo com a emoção do indivíduo.

Este artigo não apresentou a tecnologia utilizada para treinamento e classificação das emoção e não apresentou métricas avaliativas referentes ao modelo presente no sistema.

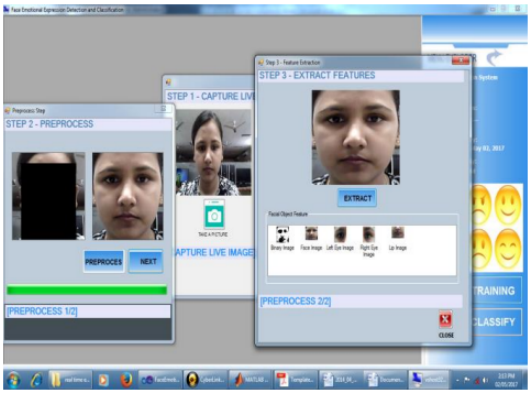

Figura 2. Execução do Sistema apresentado pelo artigo [4]

C. Artigo 3 - Content Based Facial Emotion Recognition Model using Machine Learning Algorithm [5]

Este artigo relatou o uso de um modelo de CNN (Rede Neural Convolucional) que usa a biblioteca TFLearn em cima de TensorFlow, na linguagem Python. O modelo possuía o objetivo de reconhecer emoções a partir expressões faciais em tempo real através de imagens via *web-cam*.

O modelo utilizou 3 camadas convolucionais, sendo 2 delas totalmente conectadas a camada *maxpool* para reduzir o tamanho das imagens. Este modelo teve como objetivo reconhecer 7 emoções no total. Além disso, realizou-se um treinamento com 20.000 fotos do conjunto de dados de um desafio Kaggle, FER-2013.

Com resultados obtidos, as redes A,B e C obtiveram uma acurácia com validação de 63% , 53% e 63% respectivamente. A rede A possuía o menor tamanho, a rede B o maior tamanho das redes e a rede C um tamanho médio. A rede C demonstrou uma curva de aprendizado mais lenta, mas é semelhante à da rede A na precisão final. A rede A demonstrou ser a abordagem mais promissora para este artigo de reconhecimento de emoção. Porém, a rede C possuiu um desempenho 20% melhor no *dataset* Radboud Faces Database (RAFD) em comparação a rede A.

D. Artigo 4 - A Robust Pose & Illumination Invariant Emotion Recognition from Facial Images using Deep Learning for Human-Machine Interface [6]

Utiliza redes neurais convolucionais para identificar 5 tipos de emoções, sendo elas neutro, raiva, felicidade, surpresa e nojo. O algoritmo utilizado é Viola e Jones que é baseado em classificador Haar, ou seja, os valores não são determinados pelos treinamentos, mas sim manualmente.

O modelo foi estruturando utilizando três camadas de convolução, cada camada de convolução foi seguida pela função de ativação ReLu e camadas de *pool* para extrair recursos. Para classificação de emoção, foram utilizadas camadas de saída de classificação totalmente conectado (FC), *softmax* e classificação. No total, foi utilizado 15 camadas para o reconhecimento de emoções a partir de imagens faciais.

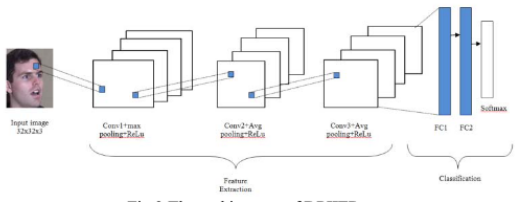

Figura 3. Passos do modelo apresentado pelo artigo. [6]

A saída era um vetor de 5 elementos que corresponde a cada classe de emoção, representando uma distribuição de probabilidade para ocorrência de determinada emoção de acordo com a expressão facial. A emoção que obtivesse a maior probabilidade de todos os rótulos de classe seria escolhida.

IV. Base de Dados

Quando trata-se de algorítimos que utilizam redes neurais, um dos principais pontos de atenção são os dados utilizados. Para este trabalho utilizaremos três bases de dados que formarão uma base de dados robusta.

A base de dados chamada JAFFE (*Japanese Female Facial Expression*), possui um conjunto de dados pequeno mas com muita qualidade. Em seu conteúdo, existe um total de 213 fotos que estão devidamente balanceadas, onde várias mulheres japonesas expressam emoções. A emoções capturadas estão subdivididas em 6 + 1 classes. Sendo elas: Neutro, Raiva, Nojo, Medo, Alegria, Tristeza e Surpresa. As imagens estão representadas em um tamanho de 256x256 *pixels*. [7] Todas as imagens foram incluídas para compor o banco final.

A segunda base de dados que será utilizada será a FER-2013(*Facial Expression Recognition*). Esta por sua vez possui um tamanho maior comparado a JAFFE. Em contrapartida, a base FER não está devidamente balanceada e portanto o algorítimo precisará realizar um pré-processamento neste sentido. Nesta base de dados estão representadas imagens com tamanho 48x48 *pixels*, onde incluem fotos de pessoas diferentes representando algumas emoções, desenhos de rosto e animações com traços de emoções. Esta base de dados também é subdividida entre 6+1 classes de emoções, da mesma forma que o banco JAFFE. [8] Foram selecionadas 547 imagens de cada emoção para compor o banco final.

A terceira base de dados se chama AffectNet. Esta base de dados é composta por imagens que são subdividas entre os seguintes rótulos: Neutro, Alegria, Tristeza, Supresa, Medo, Nojo, Raiva, Desprezo, Nenhum, Incerto, Sem Rosto. As imagens possuem a localização do rosto e cada rosto possui também a localização dos 68 pontos fiduciais. Cada expressão facial possui também um valor que define a intensidade da expressão na imagem. Todas as imagens deste banco possuem 224x224 *pixels* e são coloridas por RGB. [9] Foram selecionadas 4001 imagens das emoções: Neutro, Alegria, Tristeza, Supresa, Medo, Nojo, Raiva.

Todas as imagens selecionadas foram transformadas em imagens preto e branco de tamanho 48x48 *pixels*, é possível ver um exemplo na Figura 4. Para os experimentos, a base de

Figura 4. Exemplo de imagem presente na base de dados. Fonte: Fornecido pelo Autor

dados mesclada foi subdividida como 90% para treino e 10% para testes.

V. Metodologia

Afim de obter maiores acurácias nos modelos de classificação, a tecnologia escolhida foi a redes neurais convolucionais (CNN).

A. CNN

CNN é um algoritmo de classificação com alta acurácia para problemas de reconhecimento de imagem [10]. O objetivo de uma rede convolucional é reconhecer traços e pequenas características que podem definir um objeto quando agrupadas. Para isto, o algoritmo, geralmente é estruturado por uma sequência de camadas, sendo elas: Camada de Convolução, Camada de Agrupamento (*Pooling)* e Camada totalmente conectada (*Fully Connected*. [9]

1) Camada de Convolução: As convoluções funcionam como filtros que enxergam pequenos quadrados e vão passando por toda a imagem captando os traços mais marcantes. O filtro, que também é conhecido por *kernel*, é formado por pesos inicializados aleatoriamente, atualizando-os a cada nova entrada durante o processo de *backpropagation*. A pequena região da entrada onde o filtro é aplicado é chamada de *receptive field*. [10] A Figura 5 ilustra o primeiro passo de convolução para uma imagem de 28x28 dimensões com *receptive field* de área 5x5.

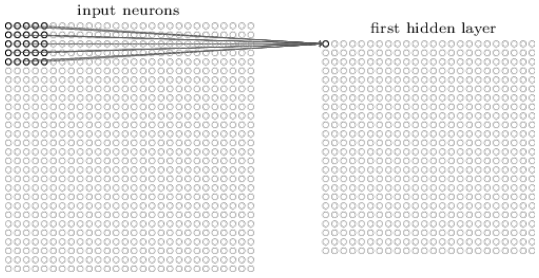

Figura 5. Convolução de imagem de 28x28 dimensões com *receptive field* de área 5x5. [10]

Geralmente após a aplicação de uma camada de convolução, a saída recebe um tratamento através de uma função de ativação. A função de Ativação ReLu é a função mais utilizada. Isto porque é uma função de ativação que possui baixo custo computacional, visto que apenas para remove valores negativos presentes na matriz.

2) Camada de Agregação: A camada de agregação generaliza um conjunto de dados e, assim como a camada de convolução, é escolhido uma unidade de área para transitar ao longo da matriz. A unidade é responsável por resumir os dados da área em um único valor. Como consequência, entrada tem sua dimensionalidade é reduzida em relação a saída. A área método mais utilizado é o *maxpooling*, no qual apenas o maior número da unidade é passado para a saída [10]. A Figura 6 ilustra o primeiro passo de agregação para uma imagem de 4x4 dimensões com uma unidade de área 2x2.

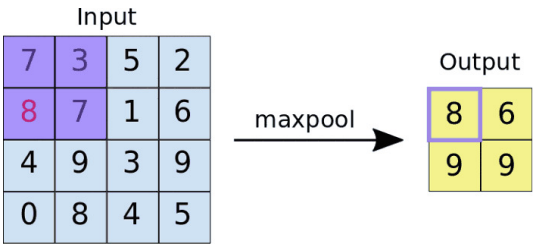

Figura 6. Primeiro passo de agregação para uma imagem de 4x4 dimensões com uma unidade de área 2x2. [10]

3) Camada Totalmente Conectada: Também chamada de camada densa, a camada totalmente conectada é onde se encontra a rede neural que reunirá as informações obtidas pelas outras camadas e fará a classificação. Esta camada geralmente também faz uso da função de ativação ReLu assim como uma função *softmax* para obter as devidas probabilidade de classificação para cada classe. [10] A Figura 7 ilustra uma estrutura comum de um modelo de CNN, sendo a sequência de execução da esquerda para direita.

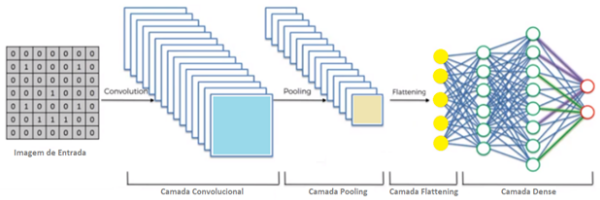

Figura 7. Convolução de imagem de 28x28 dimensões com *receptive field* de área 5x5. [11]

4) Camadas Extras: Além das camadas mais comuns de uma CNN, também existes etapas que podem auxiliar o modelo com problemas de *overfitting* ou preparar as informações antes da camada densa.

A camada *Flatten*, como já ilustrada pela Figura 7, é uma etapa que precede a camada densa e tem como objetivo transformar as matrizes resultantes pela convolução e agregação em um vetor linear [13].

A camada *Dropout* é uma etapa implementada internamente na rede neural densa durante a etapa de treinamento do modelo. Esta etapa consiste em realizar um regularização, onde alguns neurônios são desligados aleatoriamente, juntamente com suas conexões. Esta estratégia evita *overfitting* do modelo de classificação [11].

A camada *Batch Normalization* permite que cada camada da rede faça o aprendizado de forma mais independente. É usado para normalizar a saída das camadas anteriores. O uso desta camada torna-se o treinamento do modelo eficiente e também pode ser usado como regularização para evitar overfitting do modelo. [12]

B. VGG-16

VGG é um modelo de CNN proposto por Karen Simonyan e Andrew Zisserman pela Universidade de Oxford [12]. A arquitetura do modelo foi proposto inicializando com dois blocos compostos por duas camadas de convolução e uma camada de agregação seguido de três blocos compostor por três camadas de convolução e uma camada de agregação, por fim, a camada densa composta por *Flatten* e classificação através de *softmax*. A Figura 8 ilustra a arquitetura proposta para o VGG-16.

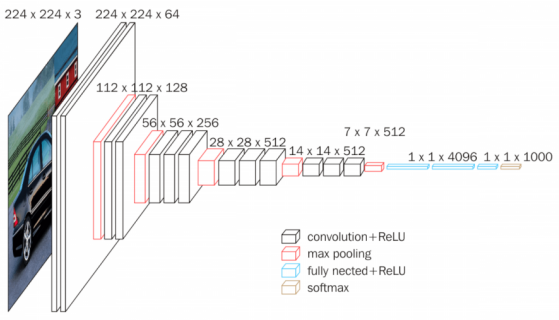

Figura 8. Arquitetura do VGG-16. [13]

C. Modelo Escolhido

O modelo escolhi foi inspirado no formato do VGG, apenas com mudanças com relação a adição de camadas Batch Normalization e Dropout.

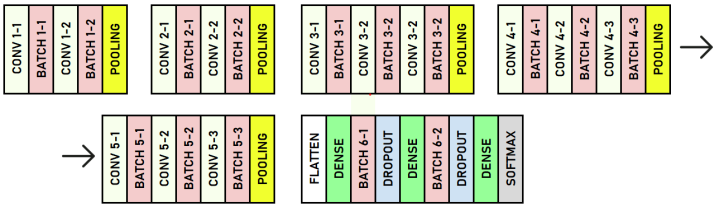

Figura 9. Arquitetura do modelo escolhido. Fonte: Fornecido pelo Autor

Como é possível visualizar na Figura 9, o modelo proposto consiste em duas sequências de duas camadas de convolução seguidas de camadas de *Batch Normalization*, terminando com uma camada de *Pooling*. Em seguida, três sequências de três camadas de convolução seguidas de camadas de *Batch Normalization*, terminando com uma camada de *Pooling*. Por fim, as seguintes camadas: *Flatten*, densa, *Batch Normalization*, *Dropout*, densa, *Batch Normalization*, *Dropout*, densa com função de ativação *Softmax*

D. Aumento de Dados

O aumento de dados é a técnica de aumentar o tamanho dos dados usados para treinar um modelo. Para previsões confiáveis, os modelos de aprendizado profundo geralmente exigem muitos dados de treinamento, que nem sempre estão disponíveis. Portanto, os dados existentes são aumentados a fim de fazer um modelo melhor generalizado. [14]

Para este trabalho, dois modelos foram treinados. Um usando apenas os 90% dos dados separados para treino e o segundo utiliza os mesmos 90%, mas com técnicas de aumento de dados.

As técnicas de aumento de dados escolhidas foram:

- Rotação de até 90 graus
- *Zoom* de até 20%
- Espelho Horizontal
- Translação vertical e horizontal de até 20%
- Variação de luminosidade entre 20% e 80%

VI. Resultados

Os modelos foram executados com otimizador Adam e uma tolerância de melhora de 15 épocas usando como métrica a acurácia com a base de teste.

A. Modelo sem aumento de dados

Realizando análises sobre acurácia obtida pelo algorítimo, com o modelo sem a presença de aumento de dados obtivemos 95.21% de acurácia em treino e 52,98% de acurácia em testes.

Como apresentado na Figura 10, o modelo obteve um *overfitting* durante o treinamento e como não houve melhora para a base de treino, o treinamento foi encerrado com 26 épocas.

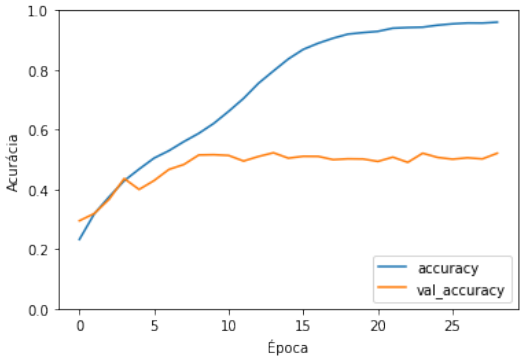

Figura 10. Histórico de evolução do modelo sem aumento de dados. Fonte: Fornecido pelo Autor

A matriz de confusão representada na Figura 11 para o modelo sem o aumento de dados, demonstrou uma assertividade maior para a emoção de Alegria e Nojo. Percebe-se

que o modelo confunde bastante Medo e Surpresa. As piores expressões do modelo foram Raiva e Neutro.

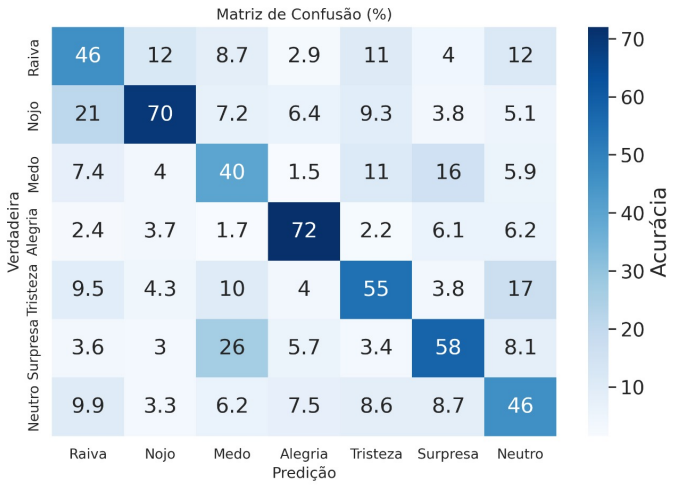

Figura 11. Matriz de confusão do melhor modelo sem aumento de dados. Fonte: Fornecido pelo Autor

B. Modelo com aumento de dados

Realizando análises sobre acurácia obtida pelo algorítimo, com o aumento de dados para o modelos obtivemos 71,48% de acurácia em treino e 70,58%de acurácia em testes. Demonstrando que o aumento de dados impediu o *overfitting* de treino além de aumentar a acurácia do teste.

Como pode ser observado na Figura 12, o modelo treinado com aumento de dados demorou cerca de 350 épocas até que a execução seja encerrada automaticamente.

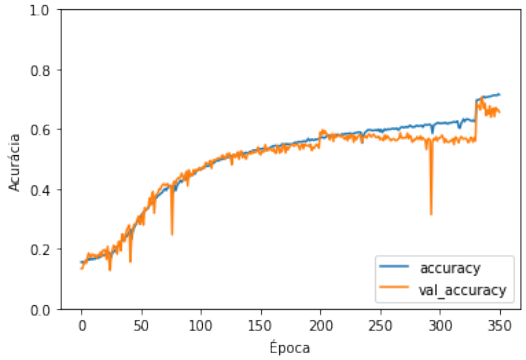

Figura 12. Histórico de evolução do modelo com aumento de dados. Fonte: Fornecido pelo Autor

A matriz de confusão representada na Figura 11 para o modelo com o aumento de dados, demonstrou uma assertividade maior para a emoção de Alegria, Nojo e Raiva. Percebe-se que neste modelo não possuiu tantas confusões entre emoções, inclusive melhorou as confusões do modelo anterior com relação as emoções Medo e Surpresa. A pior expressão do modelo para ser reconhecida foi o Neutro, mas apesar disto todas as emoções obtiveram ao menos 60% de acurácia.

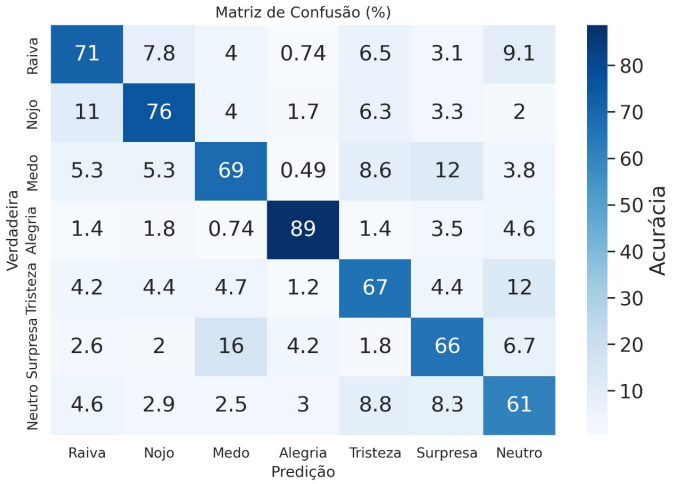

Figura 13. Matriz de confusão do melhor modelo com aumento de dados. Fonte: Fornecido pelo Autor

C. Aplicação

Utilizando o modelo com aumento de dados foi gerado uma aplicação que consistem em capturar expressões faciais de pessoas através de imagens em tempo real.

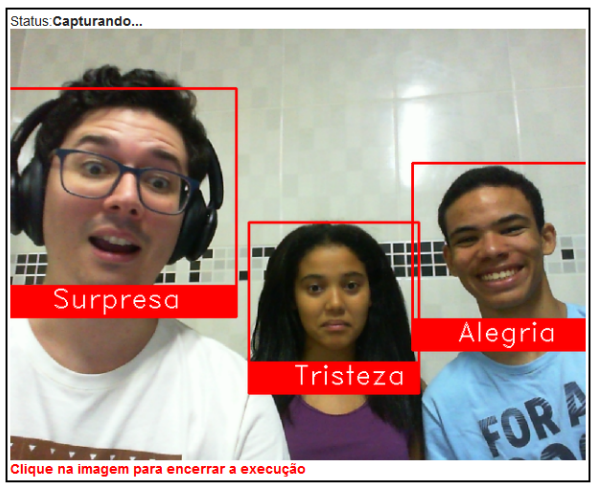

Figura 14. Aplicação desenvolvida usando o modelo como classificador. Fonte: Fornecido pelo Autor

Na Figura 14 é possível ver que a aplicação rastreia todos os rostos da imagens e atribui uma expressão facial a cada uma delas.

VII. Conclusão

Através dos resultados obtidos com os experimentos, pode-se concluir que é recomendável e relevante a utilização de técnicas de aumento de dados para treinamento de classificação expressões faciais o treinamento do modelo.

Apesar do modelo contar com sete classes, a classificação obteve uma acurácia acima de 60% para todas as emoções, chegando até em 89% de acurácia para a classificação de

Alegria, demonstrando que a técnica de redes neurais convolucionais possui bons resultados para reconhecimento de expressões faciais tendo imagens como dados de entrada.

O modelo gerado também possui diversas aplicações práticas sendo uma delas a captura de expressões através de imagens em tempo real.

Como trabalho futuro, pode-se treinar o modelo com outras base de dados e conceber outro modelo que reconheça as expressões através de pontos fiduciais como complementação.

REFERÊNCIAS

[1] BIRDWHISTELL, R.L. Kinesics and context: essays on body motion communication. 4.ed. Philadelphia: UPP (University of Pensylvania Press), 1985.

[2] DARWIN, Charles. The Expression of the Emotions in Man and Animals. Reino Unido, Oxford University Press, 2009.

[3] Ramalingam, V V & Pandian, A & Jayakumar, Lavanya. (2018). Facial Emotion Recognition System – A Machine Learning Approach. Journal of Physics: Conference Series. 1000. 012028. 10.1088/1742-6596/1000/1/012028.

[4] Deshmukh, Renuka & Paygude, Shilpa & Jagtap, Vandana. (2017). Facial Emotion Recognition System through Machine Learning approach. 10.1109/ICCONS.2017.8250725.

[5] Jadhav, Ranjana & Ghadekar, Premanand. (2018). Content Based Facial Emotion Recognition Model using Machine Learning Algorithm. 1-5. 10.1109/ICACAT.2018.8933790.

[6] Palaniswamy, Suja & Saxena, Suchitra. (2019). A Robust Pose & Illumination Invariant Emotion Recognition from Facial Images using Deep Learning for Human-Machine Interface. 1-6. 10.1109/CSITSS47250.2019.9031055.

[7] Lyons, Michael, Kamachi, Miyuki, & Gyoba, Jiro. (1998). The Japanese Female Facial Expression (JAFFE) Dataset [Data set]. Zenodo. Disponível em: https://doi.org/10.5281/zenodo.3451524

[8] Sambare, Manas (2013). FER-2013 Dataset [Data set]. Kaggle. Disponível em: https://www.kaggle.com/msambare/fer2013

[9] A. Mollahosseini; B. Hasani; M. H. Mahoor, ”AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,”in IEEE Transactions on Affective Computing, 2017.

[10] Alves, Gisely. (2018). Entendendo Redes Convolucionais (CNNs). Disponível em: https://medium.com/neuronio-br/entendendo-redes-convolucionais-cnns-d10359f21184 acessado 23/11/2021.

[11] Viceri. (2020). Arquiteturas de Redes Neurais Convolucionais para reconhecimento de imagens. Disponível em: https://viceri.com.br/insights/arquiteturas-de-redes-neurais-convolucionais-para-reconhecimento-de-imagens/ acessado 23/11/2021.

[12] Dwivedi, Rohit (2020). Everything You Should Know About Dropouts And BatchNormalization In CNN. Disponível em: https://analyticsindiamag.com/everything-you-should-know-about-dropouts-and-batchnormalization-in-cnn/.

[13] Hassan, Muneeb ul . (2018). VGG16 – Convolutional Network for Classification and Detection. Disponível em: https://neurohive.io/en/popular-networks/vgg16/ acessado 23/11/2021.

[14] Kumar, Harshit . Data augmentation Techniques. Disponível em: https://iq.opengenus.org/data-augmentation/.