UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE CIÊNCIA DA COMPUTAÇÃO

THOMAS DILLAN BALTAZAR MENDONÇA

SISTEMA DE RECONHECIMENTO DE EXPRESSÕES FACIAIS PARA CLASSIFICAÇÃO DE EMOÇÕES DE USUÁRIOS EM SISTEMAS COMPUTACIONAIS

RUSSAS

2018

THOMAS DILLAN BALTAZAR MENDONÇA

SISTEMA DE RECONHECIMENTO DE EXPRESSÕES FACIAIS PARA CLASSIFICAÇÃO DE EMOÇÕES DE USUÁRIOS EM SISTEMAS COMPUTACIONAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Ms. Daniel Márcio Batista de Siqueira

RUSSAS
2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M497s Mendonça, Thomas Dillan Baltazar.
Sistema de reconhecimento de expressões faciais para classificação de emoções de usuários em sistemas computacionais / Thomas Dillan Baltazar Mendonça. – 2018.
33 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Ciência da Computação, Russas, 2018.
Orientação: Prof. Me. Daniel Márcio Batista de Siqueira .

1. Reconhecimento Facial. 2. Reconhecimento de Expressões. 3. Redes Neurais Convolucionais. I. Título.

CDD 005

THOMAS DILLAN BALTAZAR MENDONÇA

SISTEMA DE RECONHECIMENTO DE EXPRESSÕES FACIAIS PARA CLASSIFICAÇÃO DE EMOÇÕES DE USUÁRIOS EM SISTEMAS COMPUTACIONAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em: ___/___/_____

BANCA EXAMINADORA

_____

Prof. Ms. Daniel Márcio Batista de Siqueira (Orientador)

Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Markos Oliveira Freitas

Universidade Federal do Ceará (UFC)

_____

Prof. Ms. Alex Lima Silva

Universidade Federal do Ceará (UFC)

RUSSAS
2018

AGRADECIMENTOS

Agradeço, em primeiro lugar, a toda minha família que, com muito carinho e apoio, não mediu esforços para que eu chegasse até esta etapa da minha vida, em especial ao meu pai, Francival Mendonça de Lima, minha mãe, Ana Cristina Baltazar de Lima, e ao meu irmão David Allan Baltazar Mendonça.

A todos os grandes amigos que tive o prazer de conviver durante a graduação, em especial Isaac Rahel, Carlos Victor, Vinícius Almeida, Elis Ionara, Afonso Matheus, Marcos Alencar, Erik Almeida, Hugo Venâncio, Igor Mendes, Marcos Paulo, Marília Cristina, Paloma Bispo, Sabrina Oliveira e Tágila Lima.

Agradeço também a todos os professores que me acompanharam antes da graduação, em especial à Professora Flávia Fernanda, que muito me ensinou durante o ensino médio, e a todos os professores que me acompanharam durante graduação, em especial ao Professor Daniel Márcio Batista de Siqueira, que me apoiou no desenvolvimento deste trabalho.

RESUMO

No processo de desenvolvimento de *software*, o *design* de *interfaces* tem, nas últimas décadas, ganhado muita força e vem se tornando uma das principais tarefas desse processo, sempre buscando interfaces que alcancem uma interação humano-computador o mais “amigável” possível. Para alcançar bons níveis de usabilidade. É comum registrar, por meio de vídeo, a interação do usuário com o sistema durante um teste de usabilidade, uma análise de afetividade entre outras aplicações. Porém, o uso de técnicas de processamento de imagens nos registros gerados para diminuir o tempo da análise ainda é pouco explorado, principalmente se tratando do reconhecimento de faces e expressões faciais. O presente trabalho propõe um sistema automatizado capaz de receber os insumos gerados a partir da interação do usuário com um sistema computacional, tais como imagens e vídeos dos usuários durante interação com o sistema, e aplicar a técnica de reconhecimento de face e de expressões faciais. Para isso, as expressões foram categorizadas em alegria, desgosto, raiva e surpresa. Para a classificação das expressões foi utilizada uma Rede Neural Convolucional utilizando o conjunto de dados aumentado da base JAFFE. Obtendo-se bons resultados ao atingir um acerto de 62% considerando o baixo número de imagens utilizadas para treino diante da complexidade de se classificar expressões faciais.

Palavras-chave: Reconhecimento Facial, Reconhecimento de Expressões, Redes Neurais Convolucionais

ABSTRACT

In the software development process, interface design has gained a lot of strength in the last decades and has become one of the main tasks of this process, always seeking interfaces that achieve a human-computer interaction as "friendly" as possible. To achieve good levels of usability. It is common to record, through video, the user's interaction with the system during a usability test, an affectivity analysis among other applications. However, the use of image processing techniques in the records generated to reduce the time of the analysis is still little explored, especially when dealing with the recognition of faces and facial expressions. The present work proposes an automated system capable of receiving the inputs generated from user interaction with a computer system, such as images and videos of users during interaction with the system, and apply the face recognition technique and facial expressions. For this, the expressions were categorized in joy, disgust, anger and surprise. For the classification of the expressions a Convolutional Neural Network was used using the increased data set of the JAFFE base. Good results were achieved by reaching a 62% accuracy considering the low number of images used for training given the complexity of classifying facial expressions.

Keywords: Facial recognition, Expressions Recognition, Convolutional Neural Networks

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

IHC	Interação Humano-Computador
ISO	*International Organization for Standardization*
AVA	*Ambiente Virtual de Aprendizagem*
VC	Visão Computacional
PDI	Processamento Digital de Imagens
FACS	*Facial Action Coding System*
JAFFE	*Japanese Female Facial Expression*
CNN	*Convolutional Neural Network*
RNAs	Redes Neurais Artificiais

SUMÁRIO

1 INTRODUÇÃO

No processo de desenvolvimento de *software*, o *design* de interfaces tem uma grande importância e vem se tornando uma das principais tarefas desse processo. Nas últimas décadas com o a popularização dos computadores tem se buscado interfaces que alcancem uma interação humano-computador o mais “amigável” possível (CYBIS; BETION; FAUST, 2010). A interface é parte de um sistema computacional com o qual uma pessoa entra em contato; seja de forma física, perceptiva ou conceitual (MORAN, 1981).

Buscando auxiliar a análise de problemas nas interfaces que podem impactar de forma negativa a interação do usuário com o sistema, é comum registrar, por meio de vídeo, a interação do usuário com o sistema durante um teste de usabilidade, uma análise de afetividade entre outras aplicações. Porém, o uso de técnicas de processamento de imagens nos registros gerados para diminuir o tempo da análise ainda é pouco explorado, principalmente se tratando do reconhecimento de faces e expressões faciais.

A Visão Computacional (VC) é a ciência responsável pela forma como a máquina enxerga o ambiente a sua volta. Através da extração das informações significativas de câmeras de vídeos, escâneres e outros dispositivos (CHIU e RASKAR, 2009). O reconhecimento de objetos é uma área principal da VC, reconhecer faces não é uma tarefa trivial para o computador, devido aos rostos serem estímulos visuais complexos e multidimensionais, tornando o reconhecimento uma tarefa de alto nível (AGARWAL et al., 2010). Existem diversas formas de reconhecer faces como: extração de vetores de características das partes básicas do rosto baseado; e Análise de componentes principais.

As Redes Neurais Convolucionais (CNN) por ter como inspiração o sistema visual humano, e também por ser caracterizada como um tipo particular de rede neural profunda muito mais fácil de ser treinada, vem sendo largamente adotadas pela comunidade de Visão Computacional para o reconhecimento de expressões faciais. O que caracteriza esse tipo de rede é ser composta basicamente de camadas convolucionais, que processam as entradas considerando campos receptivos locais.

O presente trabalho propõe um sistema automatizado capaz de receber os insumos gerados a partir dos testes de usabilidade, tais como imagens e vídeos dos usuários durante a interação com o sistema, e aplicar técnica de reconhecimento de face e de expressões faciais. As expressões são categorizadas por alegria, desgosto, raiva e surpresa, tal como definido pelo Modelo de Emoções Básicas proposto por Ekman (1999). Para a classificação das

expressões foi utilizada uma Rede Neural Convolucional utilizando o conjunto de dados gerado com o aumentado de imagens da base JAFFE, aplicada a um módulo que torna possível inferir a emoção sentida pelo usuário através da captura da tela, associando o sentimento com o elemento de interação a problemas de usabilidade ou experiência do usuário.

1.1 Objetivos

1.1.1 Objetivo geral

Desenvolver uma ferramenta de apoio aos testes de usabilidade, que utiliza de técnicas de reconhecimento de expressões faciais e associar essas expressões à interação do usuário com sistema em dado momento.

1.1.2 Objetivos específicos

- Desenvolver um algoritmo capaz de inferir emoções como alegria, desgosto, raiva, surpresa.
- Encontrar formas para associar as expressões ao seu contexto de uso de forma a facilitar a análise.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo descreve os conceitos relacionados a área da pesquisa. A Seção 4.1 define a área Visão Computacional. A Seção 4.2 aborda o reconhecimento de faces. A Seção 4.3 trata de redes neurais artificiais. A Seção 4.4 define convolução. A Seção 4.5 define as redes neural convolucionais. A Seção 4.6 é apresentado o que são as camadas de *pooling.* A Seção 4.7 aborda as camadas não-lineares. E por fim a Seção 4.8 abordando dropo*ut*

2.1 Visão Computacional

A visão computacional (VC) é a ciência responsável pela forma como a máquina enxerga o ambiente a sua volta. Por meio da extração das informações significativas de câmeras de vídeos, escâneres e outros dispositivos (CHIU e RASKAR, 2009). A VC estuda métodos para aquisição, processamento, análise e compreensão de imagens, fornecendo ao computador informações precisas para execução de tarefas.

O reconhecimento de objetos, aprendizagem de máquina, detecção de eventos, restauração de imagens são todas áreas relacionadas à visão computacional. A entrada dos algoritmos de visão computacional são uma ou mais imagens digitais. A análise de imagem (também chamado de compreensão de imagem) relaciona-se inteiramente aos conceitos de Processamento Digital de Imagens (PDI). Não há limites claros no contínuo desde o processamento da imagem em uma extremidade até a visão computacional na outra (GONZALEZ; WOODS, 2008).

Gonzalez e Woods (2008) dividem em três os tipos de processos informatizados nesse contínuo:

- Processos de baixo nível, que envolvem operações primitivas como o pré-processamento de imagens para reduzir o ruído, aprimoramento de contraste e nitidez da imagem: caracterizado pelo fato de que ambas as entradas e saídas são imagens.
- Processos de médio nível, que envolvem tarefas como segmentação, particionamento de uma imagem em regiões ou objetos, e descrição desses objetos para reduzi-los a uma forma adequada para processamento em computadores e classificação de objetos individuais: caracterizado pelo fato de que suas entradas geralmente são imagens, mas suas saídas são atributos extraídos dessas imagens.

- Processamento de nível superior, que envolvem o “fazer sentido” de um conjunto de objetos reconhecidos, como na análise de imagens, e, no extremo do contínuo, desempenhar as funções cognitivas normalmente associadas à visão.

2.2 Reconhecimento facial

O reconhecimento de faces pode aparentar ser uma tarefa simples, visto que os seres humanos realizam essa tarefa de corriqueiramente sem esforço aparente todos os dias (STAN; ANIL, 2011). Mesmo com a interferência de mudanças no estímulo visual devido a expressões, envelhecimento ou distrações, como óculos, barba, mudanças de cabelo, o ser humano pode reconhecer milhares de rostos ao longo dos anos. Porém o desenvolvimento de um modelo computacional capaz de reconhecer faces não é uma tarefa trivial, devido aos rostos serem estímulos visuais complexos e multidimensionais, tornando o reconhecimento uma tarefa de VC de alto nível (AGARWAL et al., 2010).

Para Stan e Anil (2011), com o aumento do número de computadores com maior capacidade de processamento e de baixo custo, o interesse no processamento digital de imagens aumentou, sendo incorporado em diversas aplicações como autenticação biométrica e vigilância. Todos tendo o reconhecimento de face como uma parte primordial da aplicação.

Existem dois métodos básicos para reconhecimento de rosto descritos em Agarwal (2010). O primeiro método é baseado na extração de vetores de características das partes básicas de um rosto, como olhos, nariz, boca e queixo, com a ajuda de modelos deformáveis, em seguida as principais informações são coletadas e convertidas em um vetor de recursos, O outro método é baseado na análise de componentes principais. Nesse método, as informações que melhor descrevem uma face são derivadas da imagem inteira da face.

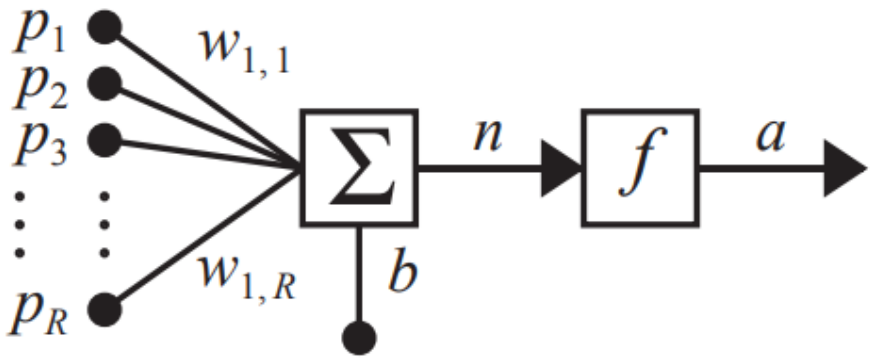
2.3 Redes Neurais Artificiais

As redes neurais artificiais (RNAs) progrediram com o decorrer do tempo através de inovações conceituais e desenvolvimentos de implementação, no entanto, essa evolução não ocorreu de forma constante. Os trabalhos iniciais que abordavam Redes Neurais ocorreram entre os séculos XIX e XX. Tais trabalhos consistiam principalmente de trabalhos interdisciplinares em física, psicologia e neurofisiologia, enfatizando teorias gerais de aprendizagem, visão, condicionamento e não incluindo modelos matemáticos específicos de operação de neurônios (DEMUTH; BEALE; JESS; HAGAN, 2014).

Segundo Demuth et al. (2014) o cérebro humano é estruturado como um computador paralelo, complexo e não-linear, com capacidade de organizar suas estruturas básicas, os neurônios, para realizar os cálculos de forma rápida e paralela. A visão moderna das redes neurais começou nos anos 1940 onde McCulloch e Pitts (1943), mostraram que as RNAs poderiam, em princípio, computar qualquer cálculo aritmético e lógico. Este trabalho frequentemente é reconhecido como a origem do campo de pesquisa das redes neurais artificiais (DEMUTH; BEALE; JESS; HAGAN, 2014).

A primeira aplicação prática para RNAs surgiu no final dos anos 50, com a invenção da arquitetura perceptron e seu algoritmo de aprendizado por Frank Rosenblatt (1958). Um neurônio artificial é a unidade fundamental básica de uma rede neural artificial. Como pode ser visto na Figura 1, o perceptron é definido por três elementos básicos: um conjunto de sinapses p_i definidas como um peso $W_{1,i}$; um somador *b* responsável pela adição do resultado da multiplicação dos sinais de entrada pelas sinapses do neurônio; e uma função de ativação *f*, que define a amplitude do sinal de saída a um valor finito *a* (DEMUTH; BEALE; JESS; HAGAN, 2014).

Figura 1: Representação de um Perceptron



Fonte: (DEMUTH; BEALE; JESS; HAGAN, 2014)

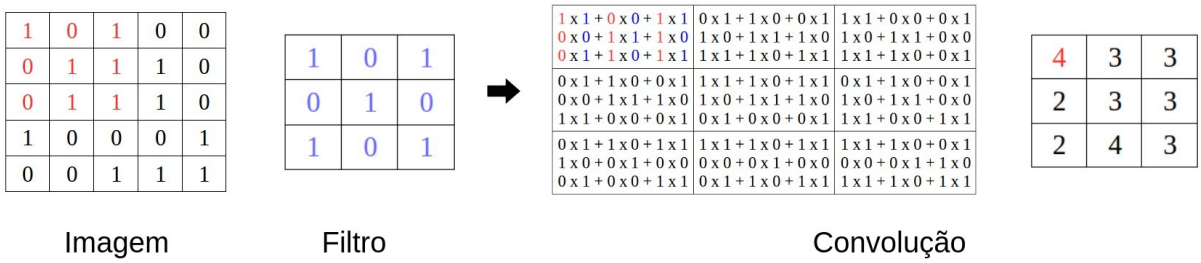
2.4 Convolução

Convolução é uma operação matemática entre duas funções *f* e *g*, produzindo uma terceira função, podendo ser interpretada como uma função modificada de *f*. Quando se fala de processamento de imagens digitais a convolução pode ser aplicada para detecção de bordas, suavização de imagem, extração de atributos, entre outras aplicações.

Quando aplicada a imagens a convolução pode ser vista como o somatório da multiplicação de cada elemento da imagem, junto com seus vizinhos locais, pelos elementos da

matriz que representa o filtro de convolução, geralmente reduzindo o tamanho da imagem original para o tamanho do filtro utilizado como é mostrado na Figura 2.

Figura 2: Exemplo de aplicação da convolução em uma imagem

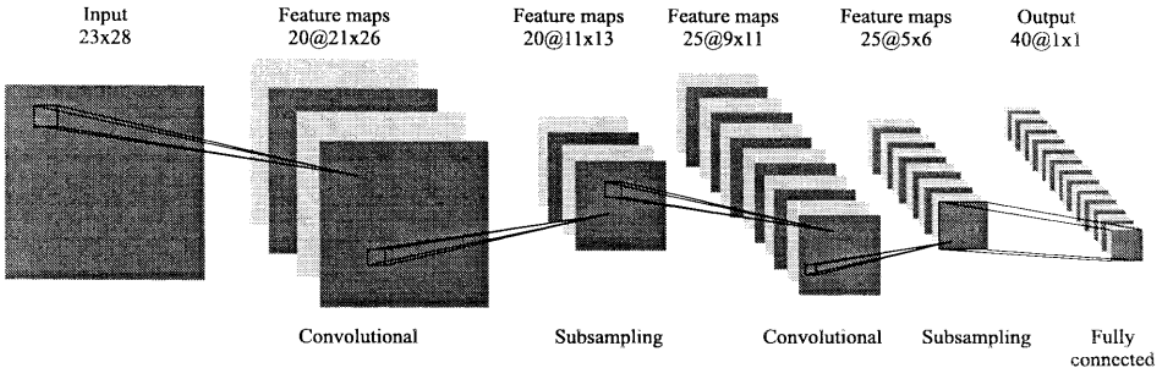


Fonte: Próprio autor

2.5 Rede Neural Convolucional

Uma Rede Neural Convolucional ou Convolutional Neural Network (CNN) é uma variação das redes de Perceptrons de Múltiplas Camadas, tendo sido inspirada no processo biológico de processamentos de dados visuais. Consiste em múltiplas partes com funções diferentes. Pode-se afirmar que principal aplicação das CNNs é o processamento de informações visuais, em particular, imagens, pois a convolução permite filtrar as imagens considerando sua estrutura bidimensional, como mostra a Figura 3.

Figura 3: Representação de uma CNN típica



Fonte: (LAWRENCE et al., 1997)

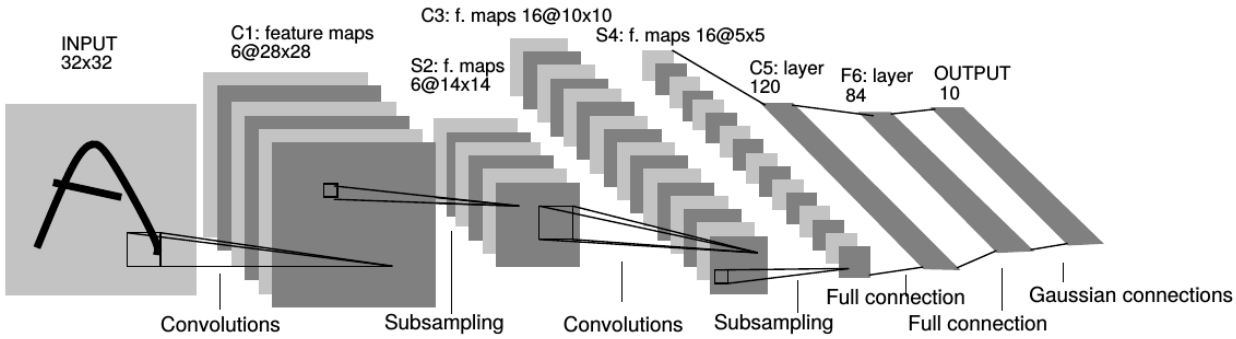
CNNs têm como inspiração o sistema visual humano, e foram o primeiro notável sucesso de treinamento em que múltiplas camadas de uma hierarquia foram treinadas de maneira robusta. O que caracteriza esse tipo de rede é ser composta basicamente de camadas convolucionais, que processam as entradas considerando campos receptivos locais. Também é

caracterizada como um tipo particular de rede neural profunda muito mais fácil de ser treinada. Por esses motivos, tem sido largamente adotadas recentemente pela comunidade de Visão Computacional (LECUN; BENGIO; HINTON, 2015)

Fukushima (1980) propôs os primeiros modelos utilizando tais conectividades locais entre neurônios e em transformações em imagens organizadas de maneira hierárquica. Já na área de reconhecimento de imagens, LeCun foi responsável por atingir o estado da arte em várias tarefas usando Redes Convolucionais (LECUN et al., 1990).

LeCun (1989) propôs uma CNN como pode ser visto na Figura 4, organizada em dois tipos de camadas, camadas convolucionais e camada de *subsampling*. Cada camada possui sua própria estrutura topográfica. A primeira LeNET recebe como entrada uma imagem em escala de cinza no tamanho 32 x 32 *pixels* que passa pela primeira camada de convolução com 6 filtros convolucionais de tamanho 5 x 5, a segunda camada é responsável pelo agrupamento médio das camadas anteriores, o processo repete de forma semelhante por mais uma vez e passa por uma camada convolucional totalmente conectada com 120 mapas de características. Por fim, uma camada om 84 neurônios totalmente conectada com a camada de saída *softmax* totalmente com 10 valores possíveis correspondentes aos dígitos de 0 a 9.

Figura 4: LeNET Arquitetura de CNN proposta por LeCun em 1989



Fonte: (LECUN et al., 1998)

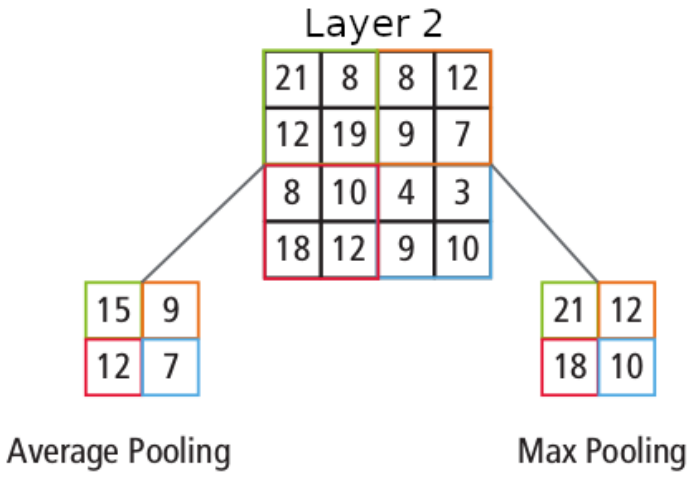
Na arquitetura LeNET o número de neurônios em cada camada da rede é diferente, e cada neurônio possui seu próprio conjunto de parâmetros, e são associados com os neurônios de um mapa retangular da camada anterior. O mesmo conjunto de parâmetros porém em uma diferente região, é associado com neurônios de diferentes localizações. Ao empilhar camadas múltiplas e diferentes em uma CNN, arquiteturas complexas são construídas para problemas de classificação. Quatro tipos de camadas são mais comuns: camadas de convolução, camadas de pool / subamostragem, camadas não lineares e camadas conectadas.

2.6 Camada de pooling

A camada de *pooling* trata cada característica mapeada pelas camadas de convolução, quando aplicado o filtro de convolução, de forma separada. Na sua instância mais simples, calcula os valores médios de uma vizinhança em cada mapa de características. Isso resulta em uma saída de resolução reduzida que é robusta para pequenas variações na localização de recursos na camada anterior, se tornando resistente a ruídos e distorções (LECUN; KAVUKCUOGLU; FARABET, 2010)

Há duas maneiras de fazer o *pooling*: pool máximo e pool médio. Em ambos os casos, a entrada é dividida em espaços bidimensionais não sobrepostos. Por exemplo, na Figura 5, a camada 2 é a camada onde será aplicada o *pooling*. Para o pool médio, a média dos quatro valores na região são calculados. Para o pool máximo, o valor máximo dos quatro valores é selecionado (SAMER; RISHI; ROWER, 2015).

Figura 5: Representação de pooling médio e pooling máximo



Fonte: (SAMER; RISHI; ROWER, 2015)

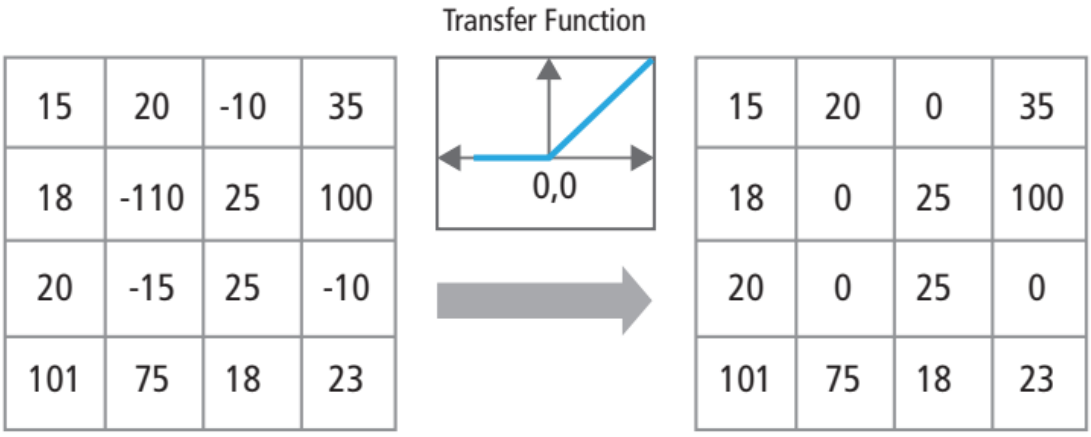
Em algumas versões recentes de CNN, o *pooling* também reúne recursos semelhantes no mesmo local, além do mesmo recurso em locais próximos. O treinamento supervisionado é realizado usando uma forma de descendente de gradiente estocástico para minimizar a discrepância entre a saída desejada e a saída real da rede. Todos os coeficientes de todos os filtros em todas as camadas são atualizados simultaneamente pelo procedimento de aprendizado (SAMER; RISHI; ROWER, 2015).

2.7 Camada não-linear

Samer, Rishi e Rower (2015) argumentam que as redes neurais em geral e as CNNs, em particular, contam com uma função de ativação não-linear para sinalizar identificação de recursos prováveis em cada camada oculta. As CNNs podem usar uma variedade de funções específicas como a Unidade Linear Retificada (ReLU), para implementar eficientemente este disparo não-linear.

Um ReLU implementa a função y = max (x, 0), portanto os tamanhos de entrada e saída dessa camada são os mesmos como pode ser visto na Figura 6. Isso aumenta as propriedades não-lineares da função de decisão e da rede global sem afetar os campos receptivos da camada de convolução. Em comparação com as outras funções não lineares usadas em CNNs (por exemplo, tangente hiperbólica, absoluto de tangente hiperbólica e sigmóide), a vantagem de um ReLU é que a rede treina muitas vezes mais rápido (SAMER; RISHI; ROWER, 2015).

Figura 6: Representação da função ReLU



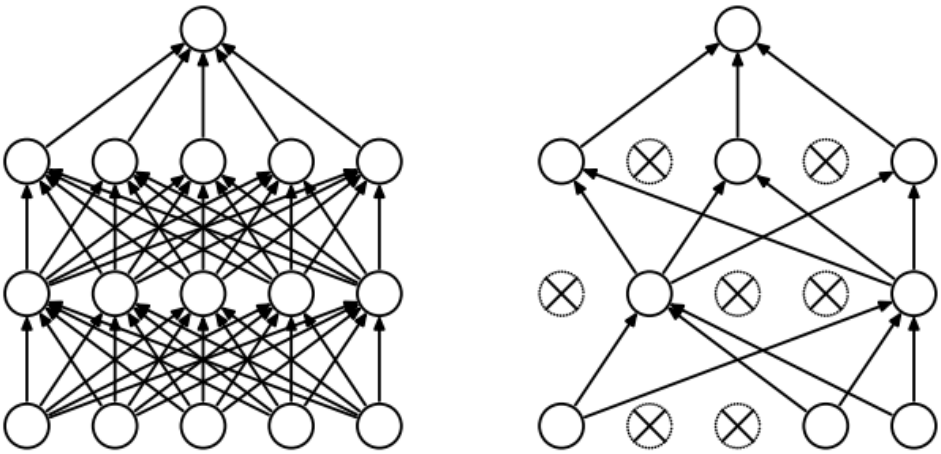
Fonte: (SAMER; RISHI; ROWER, 2015)

2.8 Camada de dropout

As redes neurais profundas contêm múltiplas camadas completamente conectadas não lineares e isso as torna modelos muito robustos e treinar muitas arquiteturas diferentes é difícil, pois encontrar parâmetros ótimos para cada arquitetura é uma tarefa muito difícil e treinar cada grande rede exige muito custo computacional. Além disso, grandes redes normalmente requerem grandes quantidades de dados de treinamento e pode não haver dados suficientes disponíveis para treinar redes diferentes em subconjuntos diferentes (SRIVASTAVA et al., 2014).

O *dropout* é uma técnica que aborda esses problemas. Isso impede o *overfitting* e fornece uma maneira de combinar de forma exponencial muitas arquiteturas de redes neurais diferentes de maneira eficiente. O termo “dropout” refere-se ao abandono de unidades (ocultas e visíveis) em uma rede neural. Ao descartar uma unidade, queremos dizer removê-la temporariamente da rede, acompanhada de todas as suas conexões de entrada e saída, como mostra a Figura 7

Figura 7: Exemplo de dropout em uma rede neural



Fonte: (SRIVASTAVA et al., 2014)

A aplicação de *dropout* a uma rede neural equivale a amostrar uma rede “thinned” dela. A rede *thinned* consiste em todas as unidades que sobreviveram ao *dropout*. Uma rede neural com *n* unidades pode ser vista como uma coleção de 2^n possíveis redes neurais diluídas. Para cada apresentação de cada caso de treinamento, uma nova rede diluída é treinada (SRIVASTAVA et al., 2014).

3 TRABALHOS RELACIONADOS

Este capítulo descreve alguns trabalhos relacionados ao principal tema da pesquisa. A Seção 2.1 detalha o sistema ErgoSV. A Seção 2.2 detalha o sistema RedFace. A Seção 2.3 detalha um sistema de reconhecimento de expressões faciais usando Redes Convolucionais.

3.1 ErgoSV

ErgoSV é um ambiente de avaliação da usabilidade de *software* apoiado por técnicas de processamento de imagens e reconhecimento de fala. O sistema oferece recursos para a aquisição de dados sobre a opinião por participante utilizando eventos (imagens faciais e palavras pronunciadas) de forma prática, rápida, pouco intrusiva durante a interação do usuário com o *software* (COLETI et al., 2013).

O ErgoSV apoia os testes de avaliação de usabilidade usando imagens registradas durante testes e processamento de fala que reconhece cinco palavras-chave pronunciadas pelos participantes: Excelente, Bom, Razoável, Ruim e Terrível por padrão, mas pode ser substituído por qualquer outro grupo de palavras como desejar o avaliador. Esses dados são usados para fornecer informações organizadas e relevantes para apoiar a análise de dados e a identificação de interfaces com possíveis problemas de usabilidade com o *software* (COLETI et al., 2013).

Porém, diferentemente do proposto pela pesquisa, a abordagem apresenta problemas devido o sistema não efetuar nenhum processamento aprofundado nas imagens obtidas durante o uso. Ao não permitir a associação das palavras-chave com a expressão do usuário de forma automatizada, o ErgoSV dificulta a identificação do foco do participante no momento de um evento, uma vez que o sistema não leva em conta que a interface do sistema analisado pode possuir vários recursos.

3.2 RedFace

RedFace é um sistema de reconhecimento de expressões faciais desenvolvido no padrão Cliente/Servidor de modo a encapsular o acesso ao AVA (Ambiente Virtual de Aprendizagem), dividido basicamente nas etapas: Captura das imagens; Rastreamento das características faciais; Classificação da expressão facial; Inferência da emoção (DINIZ et al., 2013).

O RedFace é capaz de inferir as emoções alegria, tristeza, raiva e desgosto em tempo real através de imagens sequenciais capturadas pela câmera do Kinect [Microsoft Research, 2011]. Para adquirir tais inferências, o sistema utiliza a técnica de Viola-Jones (Viola; Jones,

2004) para a detecção da face, e a técnica CANDIDE (Ahlberg, 2001) no processo de rastreamento dos pontos característicos da face. A classificação da emoção teve como base o sistema psicológico de codificação facial FACS (EKMAN et al., 2002) e regras que verificam as deformações geométricas da boca, olhos e das sobrancelhas.

A união dessas técnicas demonstrou ser bastante eficaz, o módulo de expressões faciais provou ser capaz de fornecer as características significativas e reduzir o tempo de classificação, com uma acurácia de 85,5%. Entretanto, diferentemente da proposta do presente trabalho, existe total dependência da tecnologia Kinect, impossibilitando a utilização apenas com uma câmera convencional, deixando os custos elevados e o sistema pouco atrativo, principalmente, após encerramento de sua produção.

3.3 Sistema de reconhecimento de expressões faciais usando Redes Convolucionais

No trabalho proposto por Lopes, Aguiar e Oliveira (2015) é abordado o uso de Redes Neurais Convolucionais (CNN) para o reconhecimento de expressões faciais. O sistema possui várias etapas para fazer a classificação das imagens em raiva, desdém, medo, felicidade, tristeza e surpresa. Primeiramente, é feito um processo que consiste em encontrar a localização dos olhos e, então, é feita uma normalização espacial para fazer o alinhamento dos olhos ao eixo horizontal.

Para que a CNN obtivesse melhores resultados foram realizados estudos sobre métodos de pré-processamento dos dados tais como normalizações nas imagens e a aplicação de *data argumentation* no conjunto de dados, onde são geradas novas imagens aplicando rotações nas imagens com o intuito de melhorar o tamanho do conjunto de dados e as imagens passam para a CNN. A união dessas técnicas com a arquitetura de CNN usada trouxeram ótimos resultados, possibilitando a precisão de 97,81% para a classificação.

4 PROCEDIMENTOS METODOLÓGICOS

O presente trabalho tem o objetivo de desenvolver uma abordagem automatizada aos testes de usabilidade através do uso de técnicas de processamento de imagens e redes neurais como: reconhecimento facial, e reconhecimento de expressões faciais, com o intuito de diminuir o tempo de análise dos dados obtidos a partir de gravações. Dessa maneira, será analisada de que formas as expressões feitas pelo usuário, durante uso de uma interface, serão classificadas e associar essas expressões a interação do usuário com sistema em dado momento.

A Seção 5.1 mostra informações sobre a base de dados utilizada. A Seção 5.2 o desenvolvimento do módulo de captura de vídeo. A Seção 5.3 aborda como se deu a classificação das expressões.

4.1 Base de dados

A base de dados utilizada para o trabalho de reconhecimento de expressões faciais foi o banco de imagens JAFFE que proporciona 213 imagens de 6 expressões básicas sendo elas: alegria, tristeza, surpresa, raiva, desgosto e medo, como pode ser visto na Figura 8, contando também com expressão neutra, posadas por 10 modelos femininas japonesas (LYONS, et al., 1998).

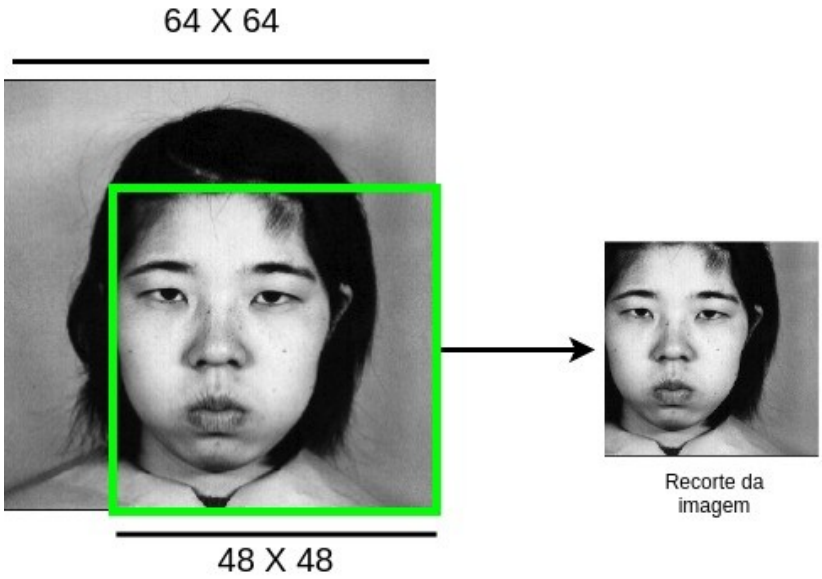
Figura 8: Exemplo de expressões do JAFFE database, da esquerda para a direita as expressões de desgosto, surpresa, felicidade, tristeza e medo.

Fonte: (LYONS, et al., 1998)

Dentre as expressões que a base proporciona, foram excluídas as imagens das emoções tristeza e medo, devido à dificuldade da associação de tais expressões com os problemas no uso de sistemas computacionais que podem ser encontrados, restando 150 imagens com as demais expressões.

As imagens inicialmente foram redimensionadas para 64x64 *pixels* e, para que a classificação não fosse prejudicada pela baixa quantidade de imagens, foram selecionados cortes de dimensão 48 x 48 *pixels* das imagens originais como pode ser visto na Figura 9, totalizando 16 cortes para cada imagem, algo nos moldes do que foi feito por DACHAPALLY (2017).

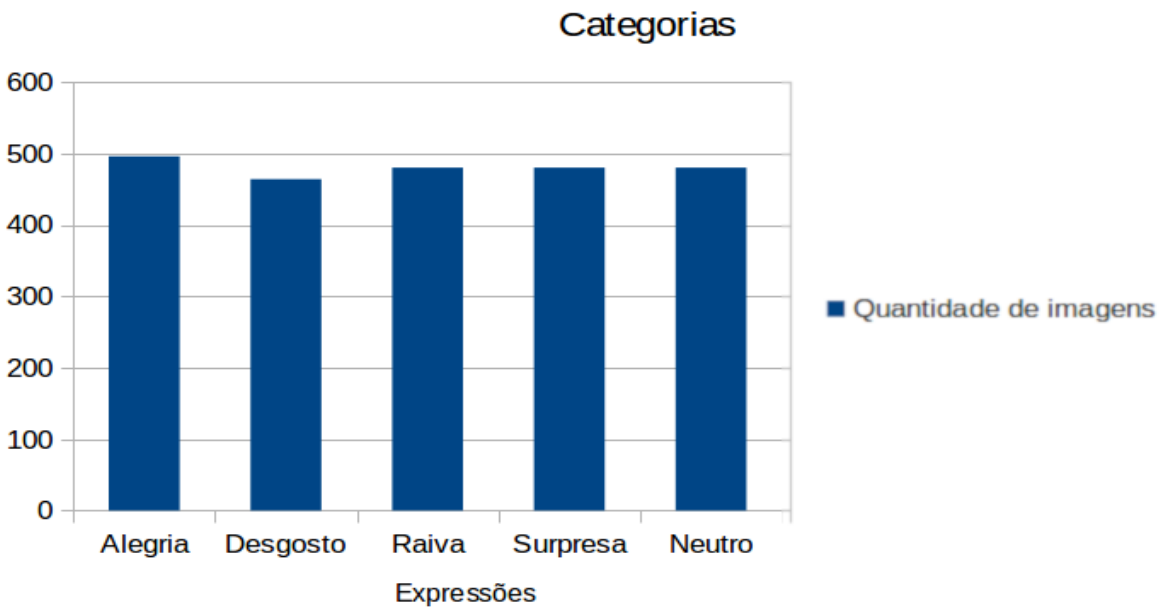
Figura 9: Demonstração de um corte de 48x48 da imagem original



Fonte: Próprio autor

Essa técnica foi utilizada para aumentar os dados para classificação, com objetivo de melhorar o desempenho da rede, fazendo o número de imagens da base aumentar para 2400 imagens distribuídas como visto na Figura 10.

Figura 10: Distribuição do conjunto de dados



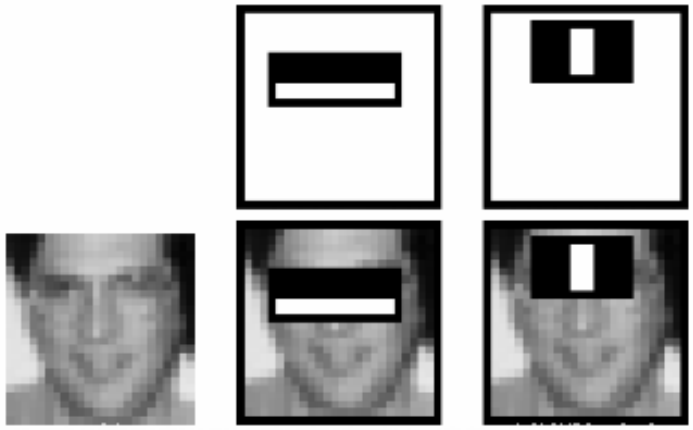
Fonte: Próprio autor

4.2 Módulo de captura de vídeo

Nessa etapa foi desenvolvido um módulo de captura de vídeo para que fosse possível registrar as expressões dos usuários e com isso poder fazer a extração de imagens da face utilizando *OpenCV.* Uma biblioteca *open source,* desenvolvida inicialmente em C/C++ amplamente usada para reconhecimento de objetos, detecção de faces e demais atividades relacionadas a imagens.

Uma das vantagens de usar a biblioteca *OpenCV* é que ela que dispõe de um classificador em cascata *Haar Features-based* conforme na Figura 11. Cada *Feature* é um valor único obtido subtraindo a soma dos *pixels* abaixo do retângulo branco pelo soma dos pixels abaixo do retângulo preto.

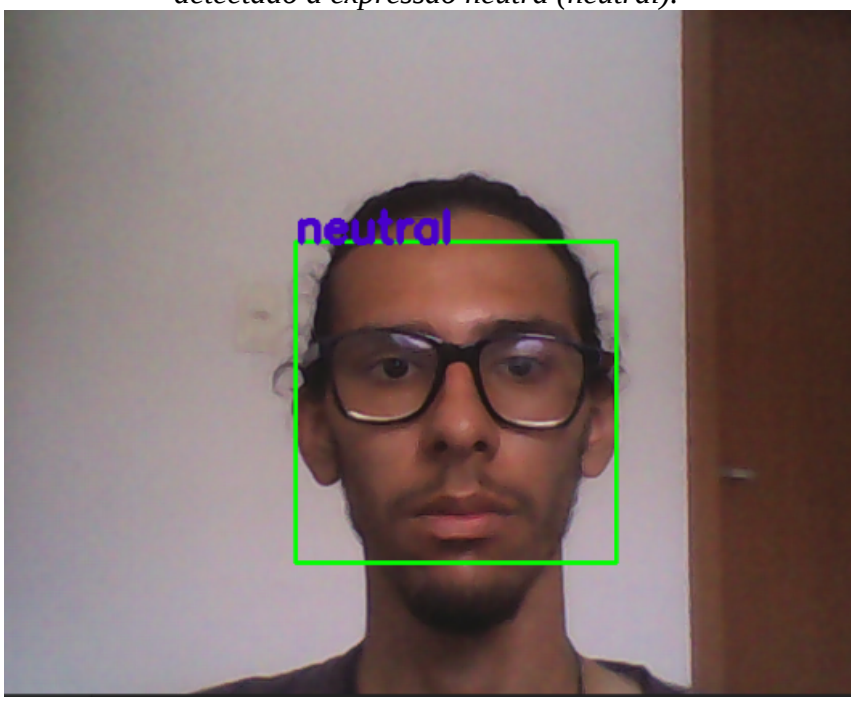
Figura 11: Haar Features-based aplicadas a uma face



Fonte (VIOLA; JONES, 2014)

Desenvolvido na linguagem de programação Python, que também possui suporte a biblioteca *openCV, o* módulo captura frame a frame imagens do dispositivo principal de gravação de vídeo e a tela do computador em que está sendo executado e utiliza o classificador em cascata para detectar a face, gerando corte da imagem e enviando como entrada para uma rede neural capaz de classificar a expressão. Após a classificação, o módulo recebe como retorno a emoção identificada na imagem e a indica seu nome visualmente como pode ser visto na Figura 12.

Figura 12: Módulo de captura de vídeo em tempo real onde foi detectado a expressão neutra (neutral).



Fonte: Próprio autor

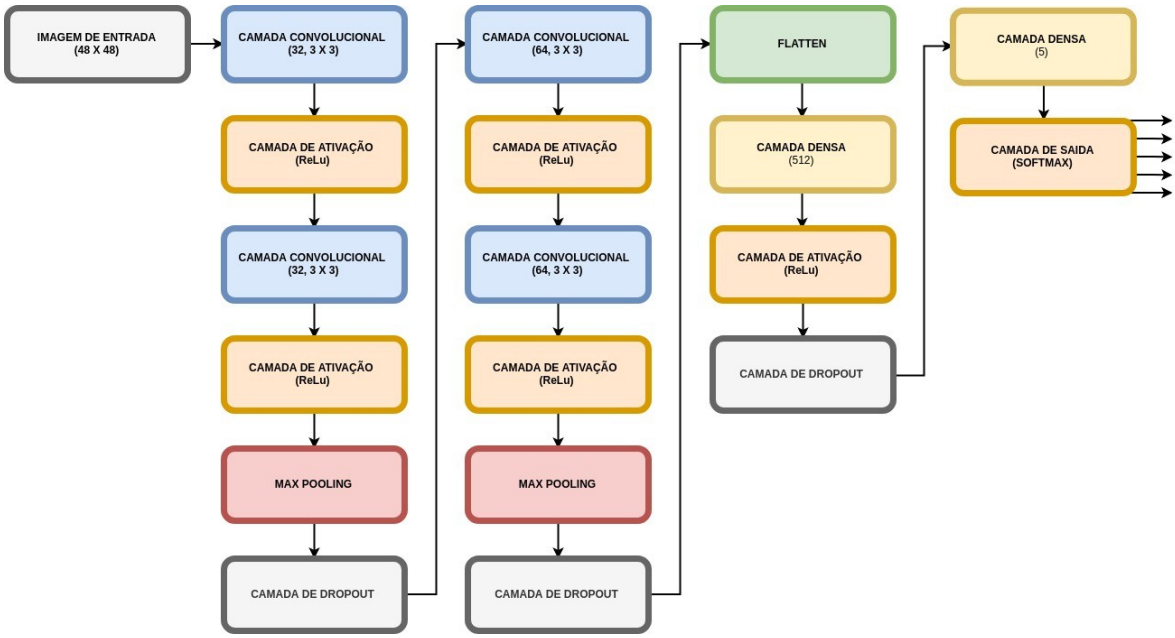
4.3 Classificação das expressões

Para a etapa de classificação das imagens do conjunto de dados final descrito na Seção 5.2 foi utilizado uma Rede Neural Convolucional. Dado sua grande relevância atual no processamento de informações visuais, principalmente aplicadas a imagens, visto que, permitem filtrá-las considerando sua estrutura bidimensional a CNN mostrou-se a melhor opção para essa etapa.

Foram realizados experimentos com algumas configurações de arquitetura de CNNs com baixo número de épocas, que define quantas vezes o treinamento será repetido, no intuito de encontrar a que obtivesse melhores resultados, dado o pequeno número de amostras para treinamento em cada categoria. A CNN que se mostrou com maior potencial foi uma arquitetura semelhante a LeNET (LECUN et al., 1990), porém, foi aplicada a rede camadas de *dropout*, responsáveis pela desativação dos neurônios de menos relevância durante a etapa de aprendizagem da rede, precedendo as camadas de *polling* para auxiliar na diminuição da ameaça de *overfitting.*

A arquitetura de CNN utilizada consiste em 18 camadas compostas inicialmente com camadas intercaladas de convolução e ativação, seguidas de uma camada de *maxpooling* e uma de *dropout. O* padrão se repete com mais intercalações de convolução e ativação (ReLu), depois de *maxpooling* e de *dropout.* Logo após, vem uma camada densa, uma de ativação, uma terceira camada de *dropout*, finalizando com mais uma camada densa e uma de ativação como mostrado na Figura 13.

Figura 13: Representação da arquitetura CNN utilizada



Fonte: Próprio autor

5 RESULTADOS

Este capítulo apresenta os resultados da implementação CNN propostas. Inicialmente Na Seção 6.1 mostra a configuração do ambiente computacional. Na Seção 6.2 é demonstrado como foram feitos os experimentos e seus resultados. A Seção 6.3 mostra os trabalhos futuros.

5.1 Ambiente de testes

A Implementação da CNN abordada na Seção 5.3, foi utilizada a Keras 2.0.8, uma biblioteca para redes neurais que funciona em conjunto com o TensorFlow implementado na linguagem Python. Os experimentos realizados foram executados utilizando uma máquina virtual dedicada do Google Colab com suporte ao Tensorflow em GPU.

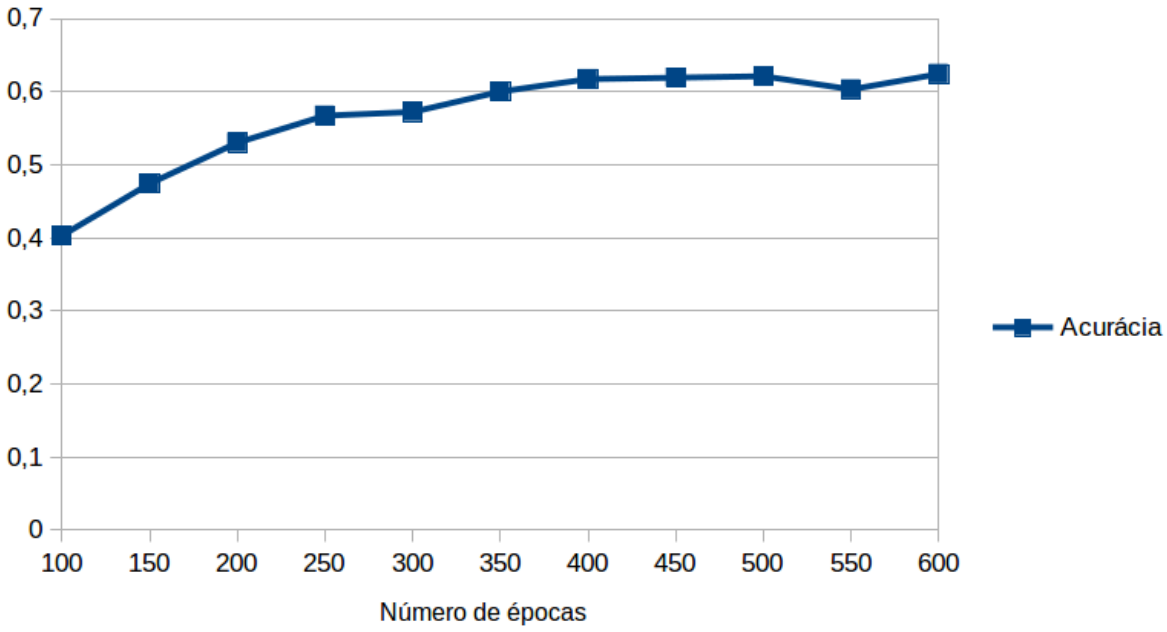
5.2 Experimentos e resultados

O conjunto de dados descrito na Seção 5.2 foi dividido em dois subconjuntos: conjunto de treino e conjunto de validação, na proporção 80% para o conjunto de treino e 20% para o conjunto de validação. Totalizando 1920 amostras para treino, distribuídas em: 393 amostras para alegria, 371 amostras para desgosto, 384 amostras para raiva, surpresa e neutro. E para validação 480 amostras distribuídos em: 100 amostras para alegria, 93 amostras para desgosto, 96 amostras para raiva, surpresa e neutro.

No decorrer dos experimentos, vários parâmetros, como número de épocas, número de neurônios em cada camada foram arbitrariamente modificados. Inicialmente com número de 100 épocas foi analisado a quantidade de neurônios em cada camada densa, aplicando o valor de 512 e de 1024 na primeira camada densa e 5 (número de classes) na segunda, a rede com 512 neurônios na primeira camada densa se mostrou melhor em relação a com 1024, com ambas atingindo uma acurácia inicial de aproximadamente 40%, porém com a de 512 conseguindo atingir, mas rapidamente o resultado nas 100 épocas.

A rede foi executada 50 épocas de cada vez. Com isso, a maior precisão de validação foi alcançada após 500 iterações como visto na Figura 14. E nas épocas subsequentes a acurácia da rede não melhorou consideravelmente, sendo executado ainda até 600 iterações, porém, começou a sofrer com o *overfitting* nesse ponto.

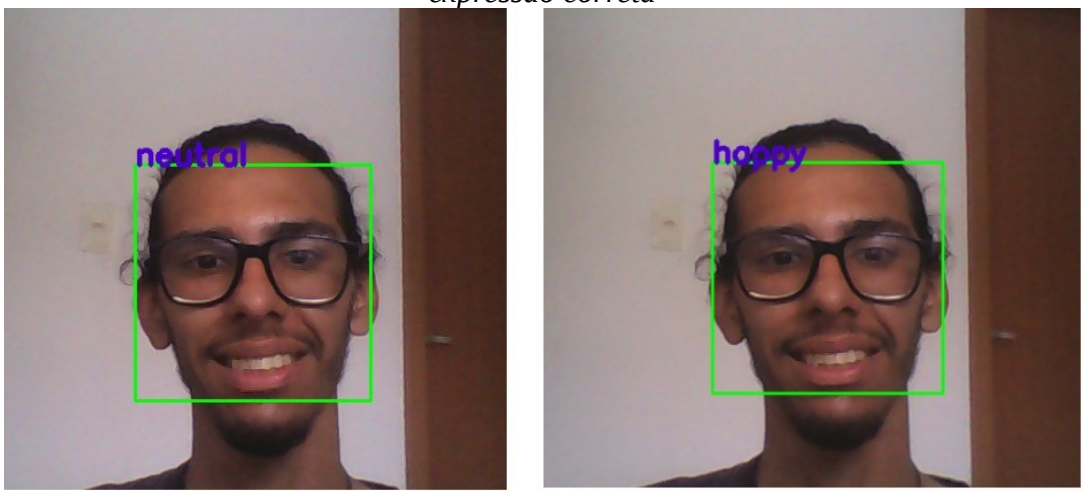
Figura 14: Acurácia do modelo



Fonte: Próprio autor

A CNN utilizando o conjunto de dados aumentado da base JAFFE, com a arquitetura final auxílio das 3 camadas gradativas de *dropout*, conseguindo atingir uma acurácia de 62%, calculado pela divisão do número de classificações corretas, pelo número total de amostras para validação, considerando que nenhum pré-processamento foi feito nas imagens e o baixo número de imagens utilizadas para treino mediante a complexidade de se classificar expressões faciais.

Figura 15: Exemplo de classificação de expressão errada e classificação de expressão correta



Fonte: Próprio autor

A Figura 15, mostra a classificação feita pela CNN, porém, como pode ser visto o classificador a partir de duas imagens com a expressão feliz classificou uma de forma erada, indicando a expressão neutro.

6 CONCLUSÃO

Pode-se ver que, mesmo com condições desfavoráveis, isto é, base de dados com número baixo de imagens, e a ausência de pré-processamento nas imagens, a CNN conseguiu-se mostrar uma excelente escolha para a classificação das expressões.

O módulo de captura e classificação permite conhecer o impacto emocional que um sistema computacional pode causar ao usuário, e com essa informação tomar medidas para melhorar a interação do usuário. Assim concluímos que é possível utilizar CNN para efetuar classificação de expressões aplicadas a testes de usabilidade, podendo proporcionar até mesmo resposta em tempo real para quem efetua o teste.

6.1 Trabalhos Futuros

Esse trabalho abre espaço para diversos trabalhos futuros, podemos destacar que a criação de uma base de dados de expressões faciais em ambiente controlado, com grande quantidade de imagens de pessoas de diferentes sexos, idades e etnias, poderia ter aumentado consideravelmente a acurácia da CNN.

Para aumentar a acurácia da rede é possível também fazer uso do módulo em testes de usabilidade reais com objetivo de gravar e utilizar as imagens obtidas para alimentar a base com mais imagem, porém, para isso será necessário validar manualmente cada predição da rede antes de fazer o retreinamento com as novas imagens, abrindo espaço também a pesquisas com o intuito de encontrar formas automatizadas de fazer essa validação.

Em próximos trabalhos também pode ser aprimorado o modulo com a geração de relatórios que venham a auxiliar mais na etapa de análise dos dados gerados, podendo, por tanto, ser aplicado a diversas maneiras como: teste de usabilidade, computação afetiva e demais contextos que possam usufruir da identificação da emoção do usuário, como intuito de tornar o uso dos sistemas computacionais mais prazeroso.

REFERÊNCIAS

AGARWAL, M.; AGRAWAL, H.; JAIN, N.; KUMAR, M. Face Recognition Using Principle Component Analysis, Eigenface and Neural Network. In: **international conference on signal acquisition and processing**.10., 2010, Bangalore. IEEE, 2010.

AHLBERG, J. C. “CANDIDE-3 – an updated parameterized face”. Report No. **LiTH-ISYR2326**, Dept. of Electrical Engineering, Linköping University, Sweden, 2001

COLETI T. A.; MORANDINI M.; DE LOURDES DOS SANTOS Nunes F. Analyzing Face and Speech Recognition to Create Automatic Information for Usability Evaluation. In: Kurosu M. (eds) **Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments. HCI 2013**. Lecture Notes in Computer Science, vol 8004. Springer, Berlin, Heidelberg, 2013

CHIU, K.; RASKAR, R.; Computer vision on tap, Computer Vision and Pattern Recognition Workshops. **CVPR Workshops 2009. IEEE Computer Society Conference on**, vol., no., p.31-38, 2009.

CYBIS, W.; BETION, A. H.; FAUST, R. **Ergonomia e Usabilidade:** Conhecimentos, métodos e técnicas**.** [S.1.]: Novatec Editora, 2010.

DACHAPALLY R. P. Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units, 2017.

DEMUTH, H. B.; BEALE, M. H.; De JESS O.; HAGAN, M. T.; Neural network design. **Martin Hagan**, 2014.

DINIZ, F. A. Um sistema de reconhecimento facial aplicado a um ambiente virtual de aprendizagem composto por agentes pedagógicos. **Proceedings of International Conferenceon Engineeringand Computer Education.** Vol. 8, 2013.

DIX A.; FINLAY J.; ABOWD G.; BEALE R. **Human-Computer Interaction. Prentice-Hall**, Inc., Upper Saddle River, NJ, USA, 1997.

DUMAS J. S.; REDISH J. C. **A practical guide to usability testing**. Intl. Specialized Book Service Inc., Portland, 2000.

EKMAN, P. Facial Expressions. **In: DALGLEISH, T.; POWER, T. (Eds.). The Handbook of Cognition and Emotion. Sussex.** Reino Unido: John Wiley & Sons Ltd. p.301-320, 1999.

EKMAN, P.; FRIESEN, W.V.; HAGER, J.C. **Facial Action Coding System: Investigator's guide**. Research Nexus division of Network Information Research Corporation, Salt Lake City, Estados Unidos, 2002.

FUKUSHIMA, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. **Biological cybernetics**, [s. l.], n. 36, p.193-202, 1980.

GONZALEZ, R. C.; WOODS R. **Digital Image Processing**. 3rd ed. Upper Saddle River, N.J.: Prentice Hall, 2008.

HOLLINGSED, T.; NOVICK, D. G. Usability inspection methods after 15 years of research and practice. **SIGDOC'07**, 2007.

ISO 9241-11 **ISO 9241-11.** Geneve: International Organization of Standardization, 2018.

KAVUKCUOGLU K.; RANZATO M.; FERGUS R.; LECUN Y. Learning invariant features through topographic filter maps," in CVPR'09.

LAWRENCE S.; GILES C. L.; CHUNG A. T.; BACK A. D. Face recognition: a convolutional neural-network approach, ***IEEE Transactions on Neural Networks***, p. 98-113, 1997.

LECUN, Y.: BOTTOU, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. ***Proceedings of the IEEE***, 1998.

LECUN, Y.; DENKER, J. S.; HEDERSON, D.; HOWARD, R. E,; HUBBARD, W.; JACKEL, L. D. Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems, 1990.

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep Learning. Nature,** p. 436-444, 2015.

LYONS, M. J.; AKEMASTU S.; KAMACHI M.; GYOBA J. Coding Facial Expressions with Gabor Wavelets, **3rd IEEE International Conference on Automatic Face and Gesture Recognition**, p. 200-205, 1998.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, p. 115-133, 1943.

MORAN, T. The Command Language Grammars: a representation for the user interface of interactive computer systems. **International Journal of Man-Machine Studies**, Academic Press, p. 3-50, 1981.

NIELSEN, J. **Usability Engineering, Academic Press**,1993.

NORMA, D.; MILLER, J.;HENDERSON, A. What you see, some of what's in the future, and how we go about doing it: HI at **Apple Computer. In Conference Companion on Human Factors in Computing Systems (CHI '95)**, New York, NY, USA, 1995.a

PARKER J. R. Algorithms for image processing and computer vision. **John Wiley & Sons**, 2010.

PREECE, J.; ROGERS, Y.; SHARP H. **Design de Interação, além da interação homem-computador.** Bookman, 2005.

PREECE, J. et al. **Human-Computer Interaction**, Addison-Wesley, 1994.

ROSENBLATT, F. **The perceptron: A probabilistic model for information storage and organization in the brain. Psychological review**, p. 386, 1958.

RUBIN, D. B. For objective causal inference, design trumps analysis. **Ann. Appl. Stat.** 2, no. 3, p. 808-840, 2008.

SAMER, C. H.; RISHI K.; ROWER, Image Recognition Using Convolutional Neural Networks. **Canadence Whitepaper,** p. 1-12, 2015.

STAN Z.; ANIL K. **Handbook of Face Recognition** (2nd ed.). Springer Publishing Company, Incorporated., 2011.

VIOLA, P.; JONES, M. Robust real-time object detection. Technical report, **University of Cambridge**, 2001.