Universidade Federal do Amazonas – UFAM

Instituto de Computação – IComp

Programa de Pós-Graduação em Informática – PPGI

Reconhecimento de Emoção por Expressão Facial Utilizando Redes Neurais de Convolução

Anderson Araújo da Cruz

Manaus – AM

Fevereiro, 2018

Anderson Araújo da Cruz

Reconhecimento de Emoção por Expressão Facial Utilizando Redes Neurais de Convolução

Texto de qualificação submetido ao Programação de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas para obtenção de pré-requisito para obtenção do título de mestrado *stricto sensu*.

Supervisor: Raimundo Barreto, D.Sc.

Manaus – AM

Fevereiro, 2018

PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

FOLHA DE APROVAÇÃO

"Reconhecimento de Emoção por Expressão Facial Utilizando Redes Neurais de Convolução"

Anderson Araújo da Cruz

Qualificação defendida e aprovada pela banca examinadora constituída pelos Professores:

Raimundo Barreto, D.Sc.– Presidente

Elaine Harada Teixeira de Oliveira, D.Sc.– Membro

Eulanda Miranda dos Santos, D.Sc.– Membro

Manaus – AM
Fevereiro, 2018

"Os loucos que acham que podem mudar o mundo,
são os que efetivamente o fazem"

Comercial Apple - "Pense Diferente", 1997

Resumo

Neste trabalho, uma abordagem é apresentada para o reconhecimento de emoção por meio da expressão facial utilizando redes neurais de convolução. Esta abordagem consiste em reconhecer emoções por meio de processamento digital de imagens e realizar diversas transformações na imagem com a finalidade de classificar as emoções. Esta proposta contempla problemas clássicos na classificação de imagens, como problemas de iluminação do ambiente e rotações do objeto principal, neste caso, a face. Uma revisão sistemática da literatura foi conduzida para mapeamento da área e a obtenção de uma visão geral. Foram descobertas as principais arquiteturas de redes neurais de convolução, e percebeu-se que ainda não há um direcionamento para qual caso é melhor optar por uma arquitetura ou outra e também a ausência de experimentos justos que possam verificar qual arquitetura alcança maior taxa de reconhecimento. Diversas aplicações foram localizadas na literatura, desde aplicações em interação humano-computador e humano-robô a cuidados com a saúde, entretanto, estas aplicações ainda não estão sendo usadas na prática, são apenas especulações, com potencial de fato serem utilizadas na próxima década. Apesar das redes neurais de convolução possuírem transformações de pré-processamento embutidas na própria arquitetura, o uso de pré-processamento na imagem antes da inserção na rede tem se mostrado eficaz e aumentado a taxa de reconhecimento. Isto evidencia a dificuldade de reconhecer emoções em expressão facial, pois geralmente a rede neural de convolução tem resolvido muitos problemas sem qualquer pré-processamento externo com acurácia elevada. O primeiro resultado parcial consistiu na utilização de uma API externa para reconhecer emoções, no qual monitorava estudantes em tempo real, enquanto estavam respondendo um simulado da prova do Exame Nacional do Ensino Médio (ENEM). Posteriormente, foram cruzadas as emoções detectadas com o desempenho no teste. Diante disso, esta prova de conceito norteou o desenvolvimento preliminar desta proposta gerando um reconhecedor de emoções baseado na arquitetura AlexNet, que foi treinada por mais de 28 mil imagens.

Palavras-chaves: Reconhecimento de Emoção, Expressão Facial, Redes Neurais de Convolução.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

RNC	Rede Neural de Convolução.
ILSVRC	ImageNet Large Scale Visual Recognition Challenge.

Sumário

1 Introdução

1.1 Contexto

Há décadas a comunidade científica tem se interessado no reconhecimento de emoções. As diversas maneiras de expressar as emoções humanas têm sido investigadas, tais como sinais fisiológicos, textos, envio de *emoticons*, dispositivo padrão de entrada de dados (e.g. teclado e mouse), voz e as expressões faciais. Esta última surgiu pelas anotações de Darwin (1965) e experiências de Ekman and Davidson (1994), que perceberam que todas as culturas emitem emoção pela expressão facial acreditando na existência de um grupo de emoções básicas (raiva, felicidade, tristeza, desprezo, medo e surpresa) que possuem a mesma expressão facial independente da cultura dos indivíduos. Apesar de tantos anos de pesquisa, a comunidade continua interessada neste assunto, pois reconhecer emoção tem sido desafiador, além de ser um campo promissor para a interação humano-computador e humano-robô. Ainda é considerado um problema em aberto, inclusive com a realização de concursos com premiação, como foi o caso do ICML'2013 e anualmente como o EmotiW dos anos de 2014 a 2017.

O progresso da área de aprendizado profundo ocasionou o surgimento de diversas técnicas poderosas de reconhecimento de padrões, gerando grande destaque para as redes neurais de convolução, que foram projetadas para processamento e classificação de imagem. As redes neurais de convolução têm sido bastante populares e utilizadas em diversos contextos dominando amplamente os trabalhos realizados pela comunidade ultimamente. Possibilitando, inclusive, o reconhecimento automático de emoção por meio da expressão facial, sendo que tal reconhecimento está próximo do que um humano reconheceria (Kim et al., 2016a). Estes resultados expressivos têm motivado pesquisadores a continuar aprimorando estas técnicas e a expressão facial tem se tornado uma abordagem eficaz para reconhecer emoções, pois não é uma abordagem intrusiva de coleta de dados quando comparada aos sensores fisiológicos. Tais sensores não são uma computação ubíqua que resulta no incomodo do usuário quando seus sinais fisiológicos são monitorados, além disso, a expressão facial possui a facilidade de ser obtida em uma captura de imagem devido a popularidade de dispositivos que possuem câmeras fotográficas (Cruz et al., 2017).

1.2 Motivação

O reconhecimento de emoção tem aplicação em muitas áreas. Destacamos alguns campos promissores. Na educação, por exemplo, segundo Jaques and Nunes (2013), estudantes durante o seu processo de aprendizagem emitem constantemente diversas emoções.

Além disso, sistemas educacionais como Ambientes Virtuais de Aprendizagem (AVA) e Sistemas de Tutores Inteligentes (STI) podem monitorar as emoções durante a interação com uma plataforma educacional em uma aula, por exemplo, para fornecer *feedback* personalizado para o estudante recomendando objetos de aprendizagem apropriados para aquele estado emocional e até mesmo, realizar ações que estimulem emoções positivas a fim de motivar os estudantes quando estes estiverem em um estado negativo. Outra área de aplicação para utilizar o reconhecimento de emoção é em realidade virtual. Segundo Riva et al. (2007), a realidade virtual pode estimular propositalmente emoções permitindo maior imersão do usuário à aplicação. Desta forma, o reconhecimento de emoção pode medir o quão efetivo é o método de estimular emoções ao usuário e, caso não seja satisfatório, o método de estímulo de emoção pode ser alterado. Para Li et al. (2015a), o reconhecimento de emoção pode auxiliar na construção de tecnologias assistivas para deficientes visuais que, quando possuem elevado grau de deficiência, apresentam dificuldades em reconhecer emoções na interação interpessoal. Em geral, é possível aplicar o reconhecimento de emoção na interação humano computador (Barsoum et al., 2016a; Chen et al., 2017a; Liu et al., 2016a; Wen et al., 2017a), e interação humano robô (Jung et al., 2015a; Shin et al., 2016a), criando a expectativa de que computadores do futuro possam reconhecer a emoção do usuário e realizar algum procedimento que ocasione maior aproximação entre homem e máquina.

1.3 Definição do Problema

Trata-se de um problema de classificação de imagem digital no qual há uma imagem ω formada por um conjunto de *pixels* (RGB) α pertencente a um conjunto de classes $\sigma = \{$neutralidade, raiva, felicidade, tristeza, desprezo, medo e surpresa$\}$, que são as emoções básicas definidas por (Ekman and Davidson, 1994), tal que haja uma função ϕ que saiba mapear ω por meio de α para σ.

Embora existam trabalhos que classifiquem emoções em imagens (Barsoum et al., 2016a; Kim et al., 2016a; Yu et al., 2016b), pouca atenção tem sido dada aos problemas clássicos em imagens como: (i) ausência de iluminação no ambiente; (ii) rotação do objeto principal, neste caso a face, e (iii) escala do objeto principal (face). Abordagens que tratam estes problemas em imagens são mais apropriadas para o uso em cenários reais no qual a exigência para classificação é maior devido as condições adversas do ambiente e pelas diferentes variações das características da face humana.

O problema considerado neste trabalho pode ser expresso na seguinte questão: *Como aprimorar os métodos de reconhecimento de emoções por meio da expressão facial a fim de permitir a classificação independente das características do ambiente e de indivíduos para o alcance de maior generalização?*

1.4 Objetivos

1.4.1 Objetivo Geral

Propor um método para reconhecer emoção humana por expressão facial para classificar emoções básicas em múltiplas faces de uma imagem e comparar a eficácia em cenários de uso real.

1.4.2 Objetivos Específicos

- Propor técnicas de eliminação de ruídos e detecção com recorte das diversas faces de uma imagem;
- Classificar cada face detectada separadamente estimando a probabilidade para cada emoção básica;
- Avaliar experimentalmente a solução proposta visando a comparação da eficácia.

1.5 Hipótese

As emoções básicas emitidas por expressão facial podem ser reconhecidas por uma rede neural de convolução, desde que esteja treinada e validada por instâncias representativas do problema (veja Seção 1.3). Este trabalho apoia-se na combinação entre uma rede neural de convolução profunda com a eliminação de ruídos da imagem por meio da utilização de técnicas de pré-processamento. A eliminação de ruídos antes da inserção da imagem na rede neural de convolução promove maior acurácia na classificação, aumento da generalização e do aprendizado para o reconhecimento de emoção em diferentes ambientes e variações da face humana.

1.6 Abordagem Proposta

Uma visão geral da solução proposta é apresentada na Figura 1. É uma solução que está direcionada para o reconhecimento de emoção por expressão facial no qual a abordagem recebe uma imagem qualquer que pode ter sido capturada por algum dispositivo de monitoramento que contém câmera fotográfica (e.g. smartphone, notebook, televisão e outros). Toda fotografia capturada pelo dispositivo de monitoramento é salva em um repositório de entrada de dados. Este repositório é consultado para obter uma imagem e enviá-la para a classificação. No processo de classificação é verificada a existência de uma face na imagem e, caso não exista, é encerrada a execução, pois é uma imagem que não contém uma face, logo, não existe uma expressão facial para classificar e novamente é consultado o repositório de entrada de dados para se obter uma nova imagem. Caso

exista uma face, uma função para recortá-la é chamada. Este procedimento é valoroso por dois aspectos: o primeiro por excluir o *background* da imagem, pois assim o classificador não necessita aprender a diferenciar o que é face e *background*, e o segundo é pela possibilidade de haver múltiplas faces na imagem realizando a classificação de cada face individualmente, reduzindo a complexidade do problema, pois é mais fácil classificar uma face por vez do que várias ao mesmo tempo. Posteriormente, a face recortada é enviada a um conjunto de filtros de pré-processamento que por sua vez operam sobre a imagem para eliminação de ruídos. Finalmente, a imagem pré-processada é enviada a uma rede neural de convolução para a classificação. Esta técnica pode ter uma característica em particular que é interessante: em vez de retornar à classe a que a expressão facial (ou uma instância) pertence, pode retornar às estimativas de probabilidade para cada classe da expressão facial (e.g. neutralidade: 0.95, felicidade: 0.025, medo: 0.025,...) possuindo assim uma propriedade em que a soma das probabilidades de todas as classes é igual a 1 e a classificação seria a maior probabilidade estimada (neutralidade com 95% de certeza), e as estimativas de probabilidade são salvas em um repositório de saída de dados. O processo anteriormente descrito deve ser repetido enquanto houver faces para classificar, isto é, quando uma imagem há múltiplas faces, e cada face é classificada uma por vez.

1.7 Organização do Trabalho

Este trabalho está dividido nos capítulos a seguir. O Capítulo 2 aborda os conceitos e definições necessários para o entendimento deste trabalho. O Capítulo 3 analisa os trabalhos relacionados. O Capítulo 4 apresenta a abordagem proposta. O Capítulo **??** discute os resultados parciais obtidos. O Capítulo **??** descreve o cronograma e detalha as atividades a serem realizadas enquanto o Capítulo **??** enfatiza as considerações finais, limitações do trabalhos e os trabalhos futuros.

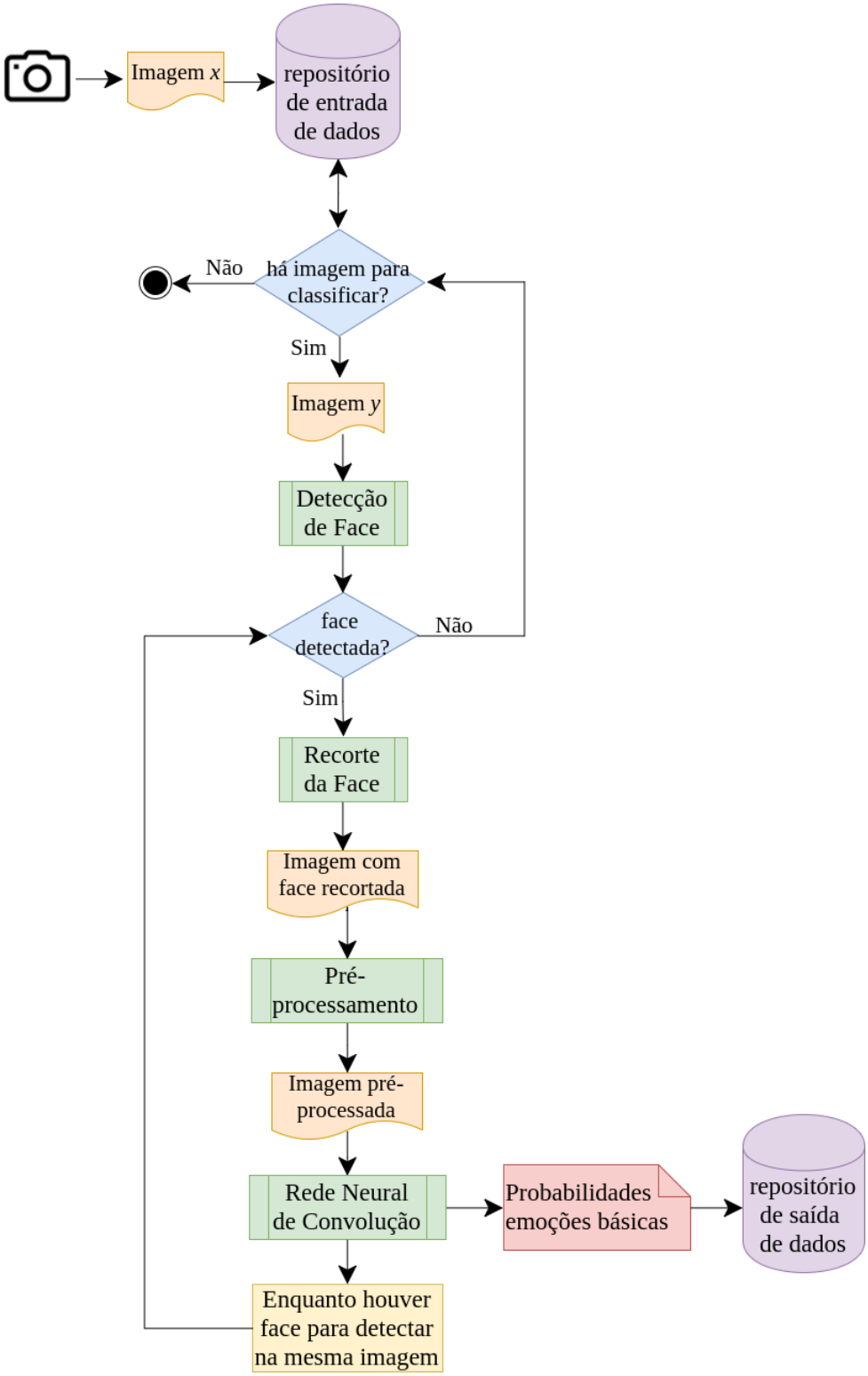

Figura 1 – Solução Proposta

2 Referencial Teórico

Neste capítulo são introduzidos os conceitos necessários para o entendimento deste trabalho e está organizado da seguinte forma. A Seção 2.1 define o que é o reconhecimento de emoção e os tipos de reconhecimento. A Seção 2.2 fundamenta a expressão facial emocional e mostra as expressões básicas existentes. A Seção 2.3 conceitua a área de aprendizagem de máquina. A Seção 2.4 apresenta o processo de classificação de imagem. A Seção 2.5 revisa as técnicas de pré-processamento que são utilizadas para eliminação de ruídos de uma imagem. As Seções 2.6 e 2.7 conceituam os aspectos básicos das redes neurais artificiais e das redes neurais de convolução, respectivamente. A Seção 2.8 descreve as arquiteturas de redes neurais de convolução utilizadas por este trabalho e para concluir a Seção 2.9 faz um resumo acerca deste capítulo.

2.1 Reconhecimento de Emoção

As emoções podem ser definidas como breves e intensas e são disparadas pela avaliação de um evento (Scherer, 2000). O reconhecimento de emoção tem sido explorado há algumas décadas e as emoções que tem sido frequentemente investigadas são as básicas como a raiva, alegria, tristeza, desgosto, medo e surpresa (Ekman and Davidson, 1994). Com objetivo de reconhecer emoção, os pesquisadores têm utilizado comumente as técnicas de reconhecimento de padrões, que por sua vez podem encontrar as características que são importantes para diferenciar as emoções, isto é, a busca pelo padrão relevante de uma entrada de dados, por exemplo uma imagem contendo uma expressão facial, de modo que a técnica consiga diferenciar a felicidade de neutralidade. Além disso, a comunidade tem gerado várias heurísticas para reconhecer emoções, entretanto com emprego somente em ambientes muito controlados.

É possível reconhecer as emoções por diversas formas (Nasoz et al., 2004): (i) sensores capturando os sinais fisiológicos; (ii) análise de expressões faciais; (iii) análise da variação da fala por microfone; (iv) movimento corporal por meio da captura de dados por dispositivos padrões de entrada (i.e. mouse e teclado) e (v) análise do texto ao escrever uma opinião.

Este trabalho está limitado ao uso de expressões faciais para o reconhecimento de emoção devido às justificativas a seguir: (i) a popularidade de dispositivos que possuem câmeras fotográficas (e.g. smartphone, tablet, smart TV e notebook) facilitam a captura da expressão facial do usuário; (ii) a evolução das técnicas de classificação de imagens que estão alcançando a taxa de reconhecimento a nível humano e (iii) por não causar qualquer intrusão ao usuário, pois fotografias da expressão facial podem ser capturadas dentro do

Figura 2 – Expressão facial emocional

cotidiano das pessoas, não causa incômodo, não necessita o uso de instrumentos especiais, além de ser imperceptível aos usuários.

2.2 Expressão Facial Emocional

Darwin (1965) verificou que fenômenos emocionais idênticos, principalmente relacionados as expressões faciais, podiam ser encontrados em diferentes culturas. Posteriormente, o trabalho de Ekman and Davidson (1994) apontou a existência de um conjunto de expressões faciais universais que representam as mesmas emoções em diferentes culturas e estão exemplificadas na Figura 2. Essas expressões faciais universais pertencem ao grupo das emoções básicas, portanto é possível reconhecer as seguintes emoções por expressão facial: raiva, alegria, tristeza, desgosto, medo e surpresa. Cada emoção que é emitida por um indivíduo possui a sua própria movimentação muscular facial, sendo assim, há a caracterização de vários padrões que são chamados de unidades de ação, como o movimento da sobrancelha, dos olhos fechando, ao levantar as bochechas e entre outros (Ekman and Friesen, 1977). As unidades de ação são os padrões relevantes para diferenciar cada tipo de emoção.

2.3 Aprendizagem de Máquina

Com intuito de criar sistemas que possuem a capacidade de reconhecer emoção de forma automática é muito comum o emprego de técnicas de aprendizagem de máquina,

que é um ramo da inteligência artificial em que máquinas aprendem a partir de uma experiência e são habilitadas para reconhecer padrões. Uma definição de aprendizagem de máquina foi dada por Alpaydin (2014): *"É a programação de computadores para otimizar um critério de desempenho usando dados de exemplo ou experiência passada"*. Na prática, isto pode ser entendido como a existência de um modelo definido com alguns parâmetros, no qual a ação de aprender consiste na execução de uma função de otimização, cujos parâmetros do modelo são otimizados, a partir dos dados de treinamento ou experiência passada. O modelo pode ser preditivo, para fazer previsões do futuro; descritivo, para obter conhecimento dos dados realizando a classificação; ou ambos.

Em aprendizagem de máquina, se as instâncias são conhecidas, isto é, cada instância possui o seu rótulo, então o aprendizado é supervisionado. Caso contrário, se as instâncias são desconhecidas, a aprendizagem é não supervisionada. Neste trabalho, o aprendizado utilizado é o supervisionado no qual os métodos possuem uma fase de treinamento e outra de teste. A primeira consiste na utilização de um conjunto de características com instâncias previamente rotuladas, também conhecido como base de treino, com objetivo de encontrar padrões nos exemplos e, assim, produzir um modelo para armazenar o aprendizado. Por fim, na fase de teste, o algoritmo deve classificar dados desconhecidos, por meio da base de teste, a partir dos padrões encontrados na fase anterior e mensurar o desempenho obtido. Além disso, o modelo gerado pode ser testado por outra base chamada de validação para averiguar com mais confiança que não há aprendizagem viciosa. Caso o desempenho esteja satisfatório o método está apto para a produção, senão deve voltar para fase de treinamento (Géron, 2017; Kotsiantis et al., 2007).

2.4 Processo de Classificação de Imagem

Nos problemas de classificação de imagem, quando a abordagem é por meio da aprendizagem de máquina, geralmente é seguido o processo: (i) a etapa inicial consiste em uma fase de pré-processamento em que são aplicadas várias técnicas com a intenção de eliminar o ruído da imagem, resultando em sua melhora considerável para as fases posteriores; (ii) a etapa de extração de característica foca em destacar ou retirar as principais formas da imagem que são importantes para a separação das classes e (iii) as características extraídas são enviadas para um classificador determinar qual a classe que a imagem pertence.

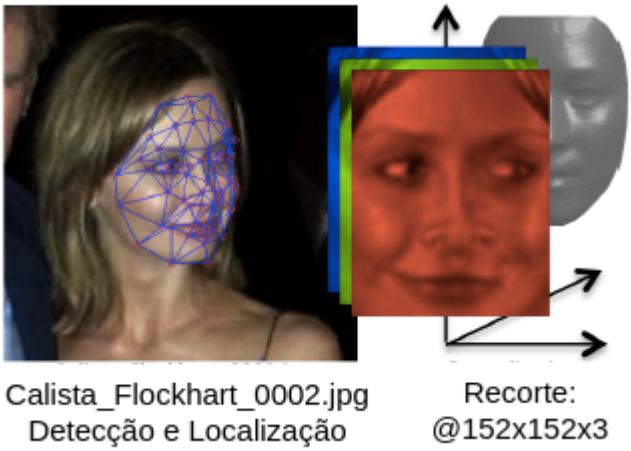

Figura 3 – Detecção Facial

2.5 Técnicas para pré-processamento de imagens em reconhecimento de emoção por expressão facial

2.5.1 Detecção Facial

Este procedimento consiste na utilização de técnicas que verificam a existência de uma face em uma imagem, seja em uma fotografia ou *frame* de vídeo. Geralmente é um problema difícil, pois dado uma imagem o método deve detectar em qual região há faces. No entanto, uma imagem pode ter diferentes objetos, *backgrounds* e ruídos. A união desses elementos pode induzir o método a identificar erroneamente uma face, causando confusão e a ocorrência de falsos positivos, isto é, objetos sendo incorretamente identificados como face.

O algoritmo Viola-Jones (Viola and Jones, 2001) é amplamente usado pela comunidade para detecção facial. Esta técnica possui vantagens como alta taxa de precisão, rapidez na execução e baixa taxa de falsos positivos. Este algoritmo é utilizado por Chen et al. (2017b), Shan et al. (2017), Shin et al. (2016b), Vo and Le (2016), Mayya et al. (2016), Ng et al. (2015) e Li et al. (2015b) para detectar a face em uma imagem e realizar o recorte com o intuito de excluir o *background* da imagem, reduzindo assim a complexidade do problema e diminuindo a carga de aprendizado da rede na qual, neste caso, o classificador não necessita mais aprender a diferenciar o que é *background* e face.

2.5.2 Histograma de Equalização (Normalização do Brilho)

A técnica de histograma de equalização é utilizada para normalizar o brilho da imagem (Jain, 1989). Obviamente há diferentes tipos de ambiente e consequentemente a iluminação pode variar bastante. Esta técnica permite que a intensidade das cores seja melhor distribuída. A atuação desta técnica equilibra o contraste da imagem nas regiões

em que há ausência.

Esta técnica foi utilizada por Shan et al. (2017), Kim et al. (2016b) e Shin et al. (2016b). Para todos os casos resultou no aumento da taxa de acurácia comparada a não aplicação desta técnica. Provavelmente o histograma de equalização funciona principalmente porque busca equalizar as cores da imagem retirando os ruídos de iluminação, isto é, realçando todos os pontos da imagem ocasionando maior facilidade para aprendizagem do problema devido a maximização da visibilidade nas regiões importantes que separam as classes, além disso, há a diminuição da carga de aprendizado do classificador, pois é menos necessário aprender a separar as classes nos casos de iluminação adversa.

2.6 Rede Neural Artificial

O ser humano inspirou-se nos pássaros para construir aeronaves e voar. A natureza também inspirou outras invenções da humanidade, por exemplo a dianteira de um trem-bala. Da mesma forma, as redes neurais artificiais tiveram a mesma inspiração, especificamente no cérebro, tendo como objetivo construir máquinas inteligentes (Géron, 2017; Goodfellow et al., 2016).

Um *perceptron*, que é uma simples arquitetura inspirada em um neurônio biológico, possui conexões de entrada e saída para conectar a outros neurônios. Cada conexão de entrada é associada a um peso, que recebe um sinal para o neurônio realizar uma computação baseada em uma função de ativação, gerando um sinal de saída que serve de entrada para outro neurônio (Géron, 2017).

A Figura 4 ilustra uma rede neural *perceptron* multicamadas. A rede neural *perceptron* possui uma estrutura em que consiste de uma camada de entrada, várias camadas intermediárias denominadas ocultas e uma camada de saída. Todas as camadas são compostas por neurônios *perceptron*. A camada de entrada recebe os dados oriundos de uma instância, por exemplo, os *pixels* de uma imagem, e encaminha os dados recebidos para a próxima camada. A camada oculta caracteriza-se por ser completamente conectada, isto é, cada neurônio conecta-se com todos da camada anterior e posterior. A camada de saída é responsável em fornecer o resultado da rede neural, por isso, nos casos de classificação, a quantidade de neurônio da camada de saída é a mesma das classes do problema. Além disso, cada neurônio da camada de saída está associada a uma classe e quando uma instância desconhecida for processada para classificação, o neurônio que terminar ativado da camada de saída representa a classificação desta instância (Géron, 2017; Goodfellow et al., 2016).

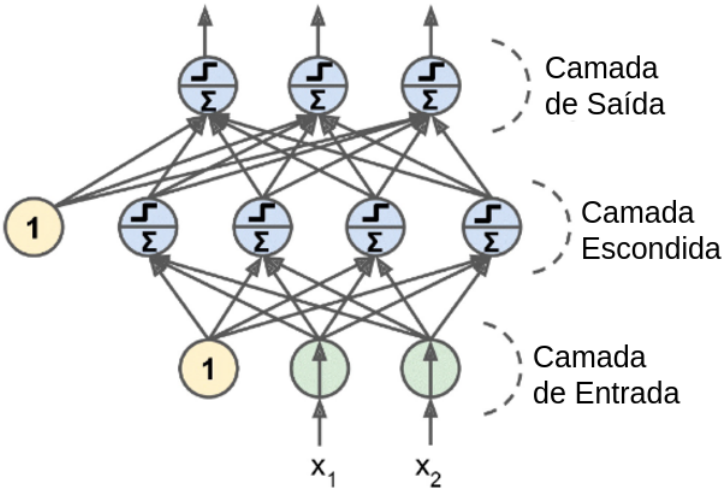

Figura 4 – Rede Neural Artificial

2.7 Rede Neural de Convolução

As redes neurais de convolução (RNC) surgiram dos estudos do córtex visual do cérebro e têm sido usadas em reconhecimento de imagem desde 1980 (Géron, 2017). Nos últimos anos, uma série de fatores contribuíram para a evolução das RNCs, principalmente relacionados ao aumento do poder de computação (hardware), ao surgimento da web, que proporcionou o aumento da quantidade de dados para treinamento e à evolução das técnicas de treinamento de uma rede neural. Este cenário favorável permitiu que as RNCs alcançassem nível super-humano em alguns problemas complexos de visão computacional. As RNCs têm sido utilizadas em larga escala tanto pela indústria como pelos pesquisadores, sobretudo em problemas como máquinas de busca, carros autônomos, sistemas de classificação automática de vídeo e imagens, entre outras tarefas.

Os trabalhos de Hubel (1959) e Hubel and Wiesel (1959) realizaram uma série de experimentos em gatos em 1958, e posteriormente, em macacos (Hubel and Wiesel, 1968), para encontrar intuições do funcionamento do córtex visual, que é a parte cerebral responsável em processar informação visual. Estes trabalhos levaram os autores a receberem o Prêmio Nobel em Fisiologia e Medicina em 1981. Seus trabalhos mostraram que muitos neurônios do córtex visual tem um pequeno campo de recepção local, ou seja, os neurônios reagem somente a um estímulo localizado na região limitada pelo campo visual. O campo de recepção local dos diferentes neurônios podem sobrepor um ao outro e a sua combinação gera o campo visual. Os autores mostraram que alguns neurônios somente reagem às imagens com padrões de linhas horizontais enquanto outros reagem às linhas com diferentes orientações. Notou-se que alguns neurônios têm um campo grande de recepção local, consequentemente, reagindo aos padrões mais complexos. Entretanto, os neurônios estão combinando um ao outro para gerar padrões menos complexos. Estas observações são evidências de que os neurônios são baseados na saída do vizinho.

O poderoso funcionamento do córtex visual está habilitado a detectar todos os padrões complexos em qualquer área do campo visual (Géron, 2017). Todos os estudos relacionados ao córtex visual foram gradualmente inseridos nas redes neurais artificiais para gerar a rede neural de convolução. A primeira RNC foi apresentada por LeCun et al. (1998), uma arquitetura denominada LeNet-5 que foi utilizada para reconhecer dígitos escritos no papel.

2.7.1 Camada de Convolução

A camada de convolução é o mais importante bloco de uma RNC e consiste em uma operação matemática que desliza uma função sobre a outra calculando a integral entre a multiplicação de duas funções. A Figura 5 ilustra o que cada camada de convolução recebe em seu campo visual dada uma imagem como entrada na RNC. Vale ressaltar que cada neurônio da primeira camada de convolução está conectado somente a alguns campos visuais de recepção da imagem, diferentemente da abordagem tradicional que é conectada a todos os *pixels*. Os neurônios da segunda camada de convolução estão conectados somente aos localizados no pequeno campo visual (retângulo) da primeira camada, novamente diferente da rede neural *perceptron* em que todos os neurônios são conectados com todos da camada anterior, e assim por diante (Géron, 2017). As vantagens da RNC sobre a abordagem tradicional são explicadas pela diferença entre ambas e são enfatizadas a seguir:

(i) A RNC tem característica esparsa como ilustrado na Figura 6, por isso, a RNC possui menor quantidade de parâmetros para serem treinados do que as redes neurais tradicionais, isto é, requer menos tempo de treinamento e recursos computacionais (Goodfellow et al., 2016);

(ii) A estrutura da RNC é comum no mundo real, por exemplo no córtex visual. Essa inspiração biológica é uma das razões para RNC funcionar tão bem no reconhecimento de imagens (Géron, 2017);

(iii) E por fim, o compartilhamento de parâmetros resulta no aprendizado de rotações dos objetos da imagem e os neurônios aprendem em conjunto ao invés de separados (Goodfellow et al., 2016).

2.7.2 Camada de *Pooling*

Esta camada tem como foco principal realizar subamostra, isto é, diminuir o tamanho de entrada da imagem entre as camadas de convolução com objetivo de reduzir a carga computacional e o número de parâmetros, ocasionando a diminuição do risco de

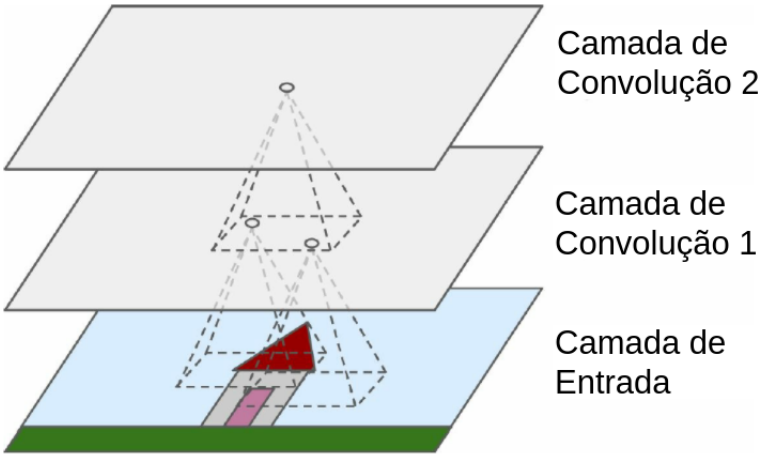

Figura 5 – Camada de Convolução com campos locais de recepção

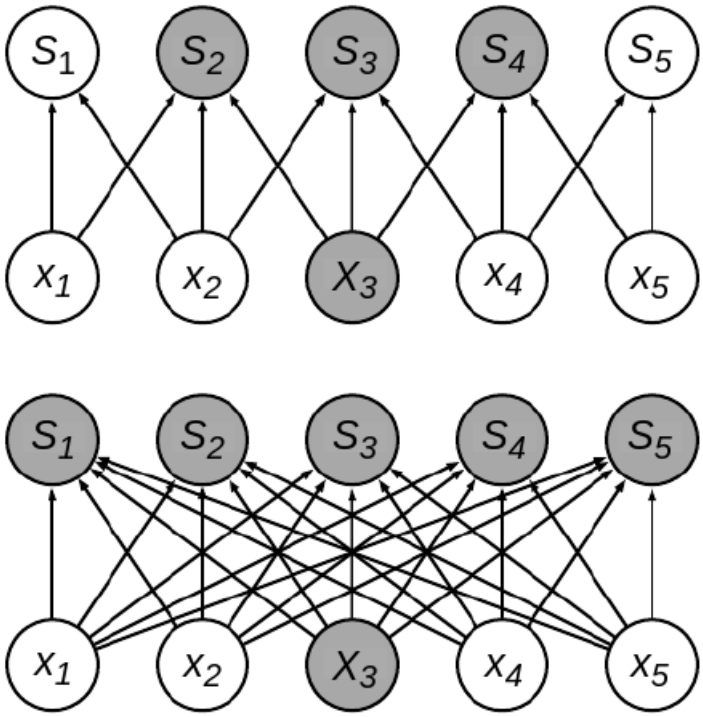

Figura 6 – Conectividade esparsa. É destacada a entrada x_3 e a saída em S que são afetadas por x_3. (Cima) Quando S recebe a convolução com um *kernel* de tamanho 3, somente três saídas são afetadas por x_3. (Baixo) Quando S é gerado por rede neural tradicional, todos são afetados por x_3.

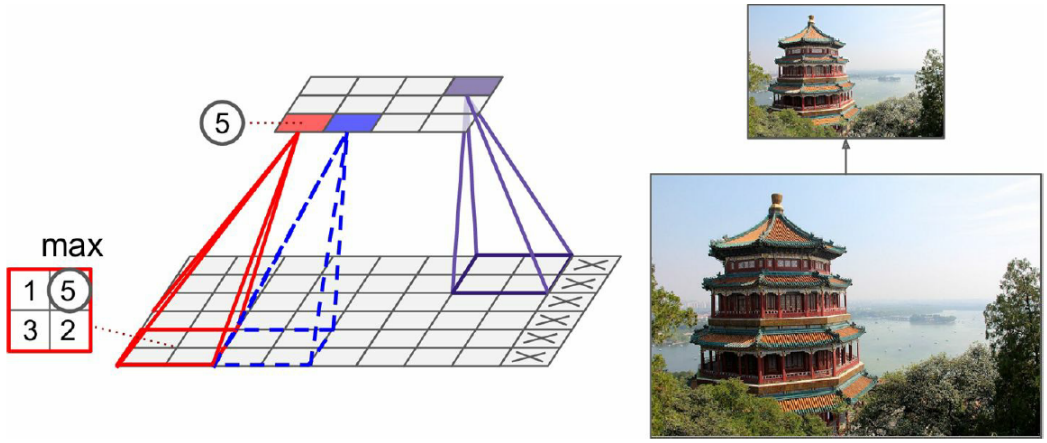

Figura 7 – Camada de *Max Pooling*

overfitting e o consumo de memória (Géron, 2017). Além disso, reduzir o tamanho de entrada da imagem faz a rede neural mais tolerável a variação do objeto principal, como a rotação da face.

Geralmente uma camada de *pooling* é implementada logo após uma camada de convolução, portanto recebe como entrada uma imagem processada anteriormente por uma camada de convolução com intuito de realizar subamostra da imagem e encaminhar o resultado para uma próxima camada de convolução (Goodfellow et al., 2016).

Existem duas principais funções de *pooling*. Por exemplo, o *max pooling* que considera o valor máximo de um campo de recepção e está exemplificado na Figura 7 por um *kernel* 2x2 que elimina 75% dos valores de entrada. Há também o *pooling* pela média de um campo de recepção e seu cálculo consiste na distância entre o *pixel* central e seus vizinhos (Goodfellow et al., 2016). A camada de *max pooling* é a mais comum operação de *pooling* utilizada em redes neurais de convolução.

2.7.3 Regressão *Softmax*

Geralmente uma arquitetura de RNC é composta em sua maior parte por camadas de convolução e *pooling*. Estas camadas processam a imagem com intuito de extrair os principais padrões para um classificador e determinar a classe dela. Por isso, ao final de uma arquitetura de RNC é necessário um classificador que tradicionalmente tem sido o *softmax*. Este classificador é um modelo generalizado de uma regressão logística, que por sua vez é normalmente usada para estimar probabilidades de uma instância pertencer a uma classe em particular, por exemplo, qual é a probabilidade de um email ser spam? (Géron, 2017). Se a estimativa de probabilidade for maior que 50%, então o modelo prevê que a instância pertence a classe positiva, caso contrário, pertence à classe negativa. Portanto, isto faz do regressor logístico um classificador binário. Um *softmax* é capaz de estimar probabilidades para múltiplas classes e não há a necessidade de treinar e combinar

Tabela 1 – Arquitetura AlexNet

Camada	**Tipo**	**Mapas**	**Tamanho**	**Kernel**	**Ativação**
Saída	Completamente Conectada	-	1000	-	Softmax
F9	Completamente Conectada	-	4096	-	ReLU
F8	Completamente Conectada	-	4096	-	ReLU
C7	Convolução	256	13 x 13	3 x 3	ReLU
C6	Convolução	384	13 x 13	3 x 3	ReLU
C5	Convolução	384	13 x 13	3 x 3	ReLU
S4	*Max Pooling*	256	13 x 13	3 x 3	-
C3	Convolução	256	27 x 27	5 x 5	ReLU
S2	*Max Pooling*	96	27 x 27	3 x 3	-
C1	Convolução	96	55 x 55	11 x 11	ReLU
Entrada	Entrada	3 (RGB)	224 x 224	-	-

múltiplos classificadores binários para tal tarefa. Um *softmax* está habilitado a estimar probabilidades para uma imagem processada em uma RNC e, para todos os efeitos, a imagem é uma expressão facial que pertence às emoções básicas: neutralidade, felicidade, surpresa, medo, raiva, tristeza ou desgosto.

2.8 Arquiteturas de Redes Neurais de Convolução

2.8.1 AlexNet

A arquitetura AlexNet foi desenvolvida por Alex Krizhevsky (por isso, o nome da mesma), Ilya Sutskever e Geoffrey Hinton. Destacando-se por ser grande e muito profunda, a AlexNet foi a primeira RNC a empilhar camadas de convolução diretamente em cima da outra, ao invés da tradicional conexão entre uma camada de convolução e a camada de *pooling*. Esta arquitetura pode ser consultada na Tabela 1.

A AlexNet usa uma normalização bastante eficiente entre as camadas C1 e C3 denominada *local response normalization*. Essa forma de normalização causa um efeito que contribui para inibição dos neurônios em ativar mais fortemente no mesmo local, ocasionando que outros mapas de características se tornem também especialistas em determinada região da imagem. Este comportamento também é observado nos neurônios biológicos (Géron, 2017).

O seu grande sucesso foi devido ao desafio de 2012 do ImageNet ILSVRC, que tem sido o principal concurso de classificação de imagens organizado pela comunidade científica, em que venceu por uma margem bastante grande: alcançou 17% no top-5 da taxa de erro, enquanto o segundo melhor alcançou somente 26%!

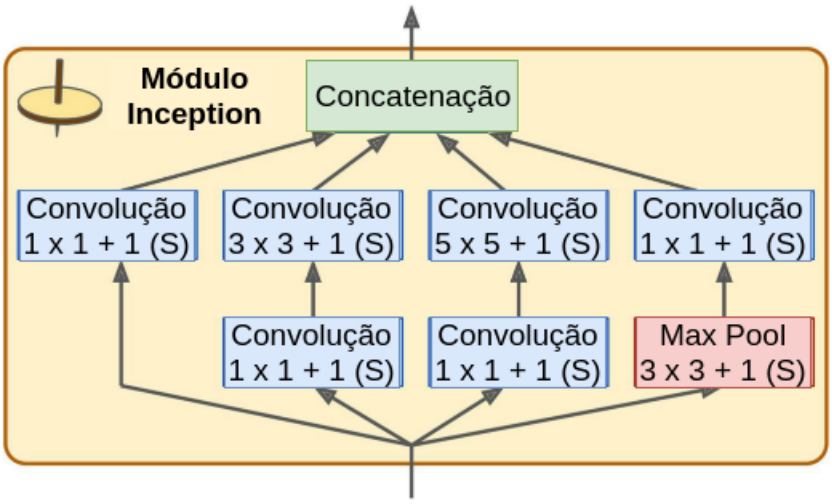

Figura 8 – Módulo *inception*

2.8.2 GoogLeNet

A arquitetura GoogLeNet (Szegedy et al., 2015) foi desenvolvida por Christian Szegedy do *Google Research*, e venceu o desafio do ILSVRC 2014 por alcançar a taxa de erro no top-5 abaixo de 7%. Este grande desempenho foi devido em grande parte pelo fato de que a rede foi muito mais profunda do que as anteriores. O aumento de profundidade está diretamente relacionado à criação de sub-redes chamadas de *inception* e está ilustrada na Figura 8. Essas sub-redes permitiram que a GoogLeNet utilizasse os parâmetros de forma mais eficiente comparada às arquiteturas anteriores e a dimensão desta eficiência pode ser elucidada pela diferença de parâmetros entre a GoogLeNet e a AlexNet, em que a primeira possui 10 vezes menos parâmetros do que a segunda (Géron, 2017).

Um módulo *inception*, que está ilustrado na Figura 8, possui uma notação ”3 x 3 + 2 (S)”. Isto significa que a camada usa um *kernel* 3 x 3, *stride* 2 e *SAME padding*. O sinal de entrada é primeiramente copiado e alimenta as camadas do módulo *inception* que estão divididas em dois conjuntos. O primeiro conjunto recebe o sinal para processar e encaminhar para o segundo. Vale ressaltar que todas as camadas de convolução utilizam a ReLU como função de ativação. É interessante observar que o segundo conjunto usa diferentes tamanhos de *kernel* (1 x 1, 3 x 3 e 5 x 5), permitindo que a rede possa capturar diferentes padrões de escala (Géron, 2017). No fim do módulo, há uma camada de concatenação, isto é, combina todas as saídas do segundo conjunto de camadas de convolução e encaminha um único sinal que é o resultado do módulo para uma próxima camada da rede. A rede GoogLeNet é composta por módulos *inception*, camadas de convolução, *max pooling*, camadas de *local normalization response*, camadas completamente conectadas e *softmax*.

Tabela 2 – Arquiteturas VGGNet

VGGNet configuração					
A	A-LRN	B	C	D	E
11 camadas	11 camadas	13 camadas	16 camadas	16 camadas	19 camadas
camada de entrada (224 x 224 imagem RGB)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
camada completamente conectada - 4096					
camada completamente conectada - 4096					
camada completamente conectada - 1000					
softmax					

2.8.3 VGGNet

Proposta por Simonyan and Zisserman (2014), a VGGNet foi a vice-campeã do desafio ILSVRC 2014, tendo alcançado 6.8% na taxa de erro top-5. A sua principal contribuição foi uma avaliação exaustiva de seis RNCs, que consistiu no aumento de profundidade enfatizando a utilização de filtros de convolução com tamanho muito pequeno (3 x 3), promovendo o aumento da profundidade da rede que passou de 16 para 19 camadas. Esta abordagem mostrou um aumento de eficiência significativo comparado às técnicas anteriores.

A VGGNet é composta principalmente por camadas de convolução, *max pooling*, completamente conectadas e *softmax*, e está ilustrada na Tabela 2. O fato de usar filtros de convolução com tamanho pequeno ocasionou no não aumento do número de parâmetros a serem ajustados a medida que a rede cresce, promovendo assim a eficiência.

2.8.4 *Ensemble*

Suponha uma questão complexa perguntada aleatoriamente para milhões de pessoas, então, todas as respostas são agregadas para obter o resultado. Em muitos casos, a resposta agregada é melhor do que a de um especialista. Isto é conhecido como sabedoria popular. Similarmente, se agregar as classificações de uma determinada instância oriunda de um grupo de RNCs e, frequentemente, tem mais acertos do que a classificação de uma única RNC. Um grupo de RNCs ou outros tipos de classificadores e regressores são chamados de *ensemble* (Géron, 2017). Geralmente, quem trabalha com *ensemble* no problema de reconhecimento de emoção tem calculada a média das probabilidades estimadas pelas diversas RNCs para decidir qual a classificação final da expressão facial.

2.9 Resumo

Neste capítulo foram apresentados os principais conceitos utilizados por esta proposta. Foi visto que uma RNC, que é uma técnica de aprendizagem de máquina, foi projetada para a classificação de imagem, justificando a sua escolha para a classificação de emoções em expressão facial. Esta técnica é muito poderosa e tem sido inspirada no cérebro dos mamíferos. A expressão facial tem se destacado como uma forma eficaz de coleta de dados para o reconhecimento de emoção, principalmente, por ser ubíqua, não intrusiva e pela popularidade de dispositivos que contém câmeras fotográficas. Entretanto, somente as emoções básicas são possíveis de reconhecer por meio da expressão facial. Isso significa que reconhecer tédio, frustração e confusão se torna bastante difícil por esse meio. Além disso, técnicas de pré-processamento, tais como histograma de equalização e detecção com recorte da face são úteis durante o processo de classificação de imagens, justamente por diminuir a carga de aprendizado da rede. Por fim, foram conceituadas algumas das principais arquiteturas de RNCs em que cada uma possui sua característica em particular que precisam ser experimentadas em diferentes cenários.

3 Trabalhos Correlatos

Os trabalhos relacionados são discutidos neste capítulo e está organizado da seguinte forma. Na Seção 3.2 analisa as principais extrações de características encontradas na literatura para o processamento da expressão facial. Na Seção 3.3 avalia os trabalhos correlatos categorizados pelas arquiteturas de RNC. Na Seção 3.4 é comentado as aplicações para o reconhecimento de emoção por expressão facial e enquanto a Seção 3.5 faz um resumo a respeito deste capítulo.

3.1 Preparação dos dados

Os problemas de classificação em geral, seja de imagem, vídeo, áudio ou qualquer tipo, tradicionalmente sofrem pela ausência de dados. Algoritmos de aprendizagem de máquina requerem quantidade de dados expressivos para apresentar soluções com desempenho satisfatório, especificamente as redes neurais profundas. Raramente há dados disponíveis e que sejam suficientes para treinar e validar uma rede neural de convolução, vale ressaltar que cada problema tem sua particularidade, isto é, quanto maior a complexidade mais dados são necessários.

Contudo, a comunidade de reconhecimento de emoção para amenizar esse problema utiliza a técnica de aumento de dados e multiplicação de imagens. Essa técnica consiste na geração de cópias de uma imagem original, que contém uma expressão facial, para gerar imagens duplicadas. Entretanto, tais imagens duplicadas são diferentes da imagem original, justamente por possuir alterações na posição da face com leves rotações da mesma, variação da intensidade das cores e redimensionamento com aplicação de *zoom*. As imagens aumentadas são usadas durante o treinamento contribuindo para a rede neural aprender a reconhecer emoção em diferentes rotações, intensidade de iluminação e escala.

Os trabalhos de Barsoum et al. (2016b); Huang and Lu (2016); Kim et al. (2016b); Shin et al. (2016b); Yu et al. (2016c) e Li et al. (2015b) utilizaram a técnica de aumento de dados. A técnica foi configurada para aumentar entre 5 a 10 vezes cada imagem original. Sendo assim, a base de dados original foi ampliada em até 10 vezes, gerando um ganho considerável dos dados. Tais trabalhos alcançaram boas taxas de reconhecimento em que o *fine tuning* dos modelos possuem generalização adequada e não apresentando *overfitting* e *underfitting*. O resultado expressivo foi viabilizado pela técnica de aumento de dados justamente pela rede ser treinada e validada com maiores quantidades de dados.

3.2 Extração de Característica

Uma etapa essencial durante o processo de classificação de imagem é a extração de característica. A extração de característica é sucintamente enfatizada na Seção 2.4 e tem como finalidade destacar ou retirar as formas mais relevantes da imagem que são cruciais para a separação das classes. A seguir, os principais tipos de extração de características empregados para o reconhecimento de emoção por expressão facial são analisados.

3.2.1 Extração Geométrica

A extração de características geométrica consiste na obtenção de pontos faciais ilustradas pela Figura 9. As características geométricas tem como finalidade capturar as deformações na face causadas pela ativação dos músculos a partir dos pontos faciais (Yu et al., 2016c). Esses pontos faciais podem ser mapeados pelos seguintes métodos: Yu et al. (2016a) e Yu et al. (2014). A extração geométrica é uma abordagem que realiza medições entre diversas partes da face tais como:

(i) Altura da sobrancelha esquerda/direita (distância vertical entre o ponto mais superior da sobrancelha e centro do olho);

(ii) Altura da pálpebra esquerda/direita (distância vertical entre o ponto mais superior do olho e parte inferior do olho);

(iii) Altura do nariz (distância vertical entre o ponto mais inferior do olho para o nariz e centro de ambos os olhos);

(iv) Largura do nariz (distância horizontal entre os pontos do nariz mais à esquerda e à direita);

(v) Altura do lábio superior (distância vertical entre o ponto mais superior e o centro da boca);

(vi) Altura do lábio inferior (distância vertical entre o ponto mais inferior e o centro da boca);

(vii) A distância do ponto da boca mais a esquerda para o centro da boca;

(viii) E por fim, a distância do ponto da boca mais a direita para o centro da boca.

A extração geométrica é amplamente empregada nas abordagens tradicionais de aprendizado de máquina, isto é, abordagens que não utilizam as redes neurais de convolução. Todavia, o trabalho de Yu et al. (2016c) realiza a extração geométrica concatenando com a RNC e obteve um pequeno ganho na taxa de precisão de 1%. Um diagrama da

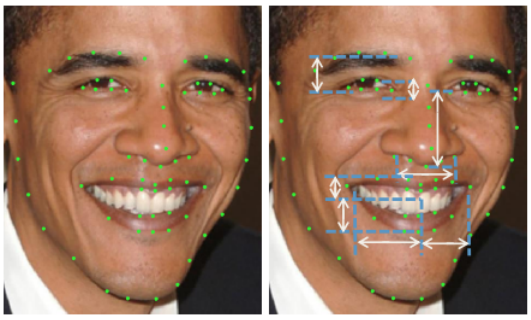

Figura 9 – Extração dos pontos faciais para características geométrica

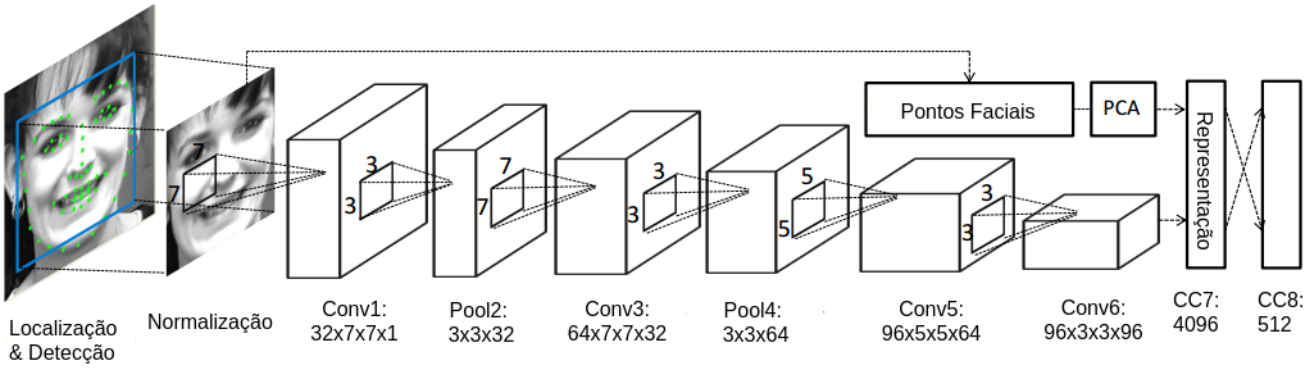

Figura 10 – Concatenação dos pontos faciais com uma rede neural de convolução

sua abordagem é ilustrada na Figura 10. No entanto, a combinação entre a extração geométrica e RNC, obviamente, aumenta o custo computacional devido a outros algoritmos serem executados como o mapeamento dos pontos faciais e as suas distâncias. Caso o foco do reconhecedor de emoções for aplicações em cenários reais, provavelmente, não é viável a concatenação devido o aumento na taxa de reconhecimento ser baixo, portanto não compensada pelo aumento do custo computacional, pois tais cenários requerem classificação instantânea e em tempo real.

3.2.2 Extração Aparente

A extração de característica aparente considera as sub-regiões faciais, principalmente próximas da boca e dos olhos, como características essenciais para a classificação, diferentemente, das características geométricas que foca na captura das deformações dos pontos faciais e possui como desvantagem não considerar as mudanças aparentes causadas por essas deformações capturadas (Yu et al., 2016c). A extração aparente foi amplamente estudada por Ekman and Davidson (1994), encontrando 96 unidades de ações (ou sub-regiões) na face correspondentes a movimentação de diversos músculos relacionada a uma determinada emoção. A RNC está habilitada naturalmente a realizar a extração de característica aparente, sendo assim, o processo de aprendizado está associado a descoberta de

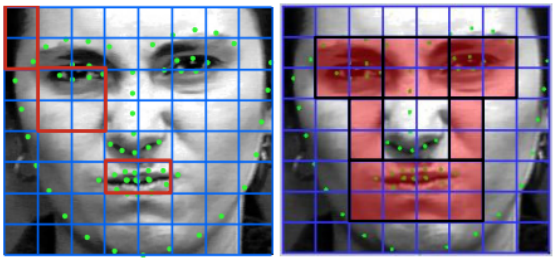

Figura 11 – Extração das sub-regiões faciais para características aparente

quais as sub-regiões da face mais relevantes para determinar qual a emoção está emitida em uma face. A Figura 11 ilustra as sub-regiões de uma face que são interessantes para a classificação depois de um processo de aprendizado.

3.3 Arquiteturas

3.3.1 AlexNet

A arquitetura AlexNet foi a RNC mais encontrada na literatura para reconhecimento de emoção, principalmente por ter sido a pioneira da família de métodos conhecido como aprendizado profundo que fez grande sucesso, inclusive esta rede venceu o desafio ILSVRC-2012. O trabalho de Kim et al. (2016b) utiliza AlexNet para reconhecer emoções, a sua abordagem é destacada por combinar a AlexNet com uma rede profunda *autoencoder* para alinhamento de face, esta foi uma grande contribuição do autor que apresentou uma solução funcional para o problema de rotação ou desalinhamento da face. O trabalho de Shan et al. (2017) propõe uma abordagem que consiste no emprego da técnica de equalização de histograma durante a fase de pré-processamento com intuito de resolver o problema da iluminação, desta forma, foi mostrado que a aplicação da técnica melhorava o aprendizado da rede. Nos trabalhos que utilizaram a AlexNet é notável que para alcançar maiores taxas de reconhecimento foi necessário o emprego de técnicas de pré-processamento.

3.3.2 VGG

Barsoum et al. (2016b) utilizou uma VGG de 13 camadas, isto é, uma arquitetura bastante profunda inclusive com camadas de *dropout* para alcançar maior generalização. Utilizou a técnica de aumento de dados, isto é, gerou novas imagens multiplicando por 10 a imagem original, as imagens geradas tinham variações da pose da face, pequenas rotações que contribuem para o aprendizado da rede. No trabalho de Ng et al. (2015), abordou a transferência de aprendizado para melhorar a performance de classificação, a VGG utilizada anteriormente tinha sido pré-treinada com a base do *ImageNet* (do tradicional

desafio ILSVRC), que é uma base que contém dezenas de objetos, e então, foi realizado um segundo auto-ajuste (treinamento) para a classificação das expressões faciais.

3.3.3 GoogLeNet

Guo et al. (2016) realizou um trabalho comparando a arquitetura *GoogLeNet* com a *AlexNet*. A primeira alcançou melhores resultados, principalmente por possuir camadas *inceptions* que são arquiteturas mais otimizadas para a classificação de imagens. Neste trabalho, também foi testado o classificador *kNN* na última camada, que alcançou melhor resultado que o tradicional *softmax*. Entretanto, vale ressaltar que, o *kNN* é uma técnica baseada em instâncias, isto é, não aprende e consequentemente não gera um modelo. Em outras palavras, caso haja $100k$ imagens na base de treino, o *kNN* para classificar qualquer instância aleatória da base de teste ou validação necessita realizar um conjunto de cálculos de distâncias entre as $100k$ imagens da base de treino para encontrar a classificação da instância, implicando em uma grande desvantagem pois este processo é repetido para cada instância que se deseja classificar.

3.3.4 Ensemble

Os trabalhos que utilizaram *Ensemble* combinaram CNN com outas técnicas ou com outras CNN. No trabalho realizado por Wen et al. (2017b) foram combinadas até 100 CNNs, no qual obteve melhor resultado que uma CNN sozinha. Entretanto, foi verificado que poucas CNNs, isto é, menos de 10, já alcançam o mesmo resultado que 100 CNNs juntas, ou seja, o aprendizado fica estável se continuar aumentando a quantidade de CNN a partir de um limiar. Isto implica que menos de 10 CNNs combinadas, é o suficiente para aprender o problema de reconhecimento de emoção por expressão facial. Liu et al. (2016b) implementou um *ensemble* de 3 CNNs com um único classificador *softmax* que recebia a extração de características das 3 CNNs. No trabalho de Shin et al. (2016b) foi treinada 20 redes diferentes com 5 entradas diferentes, também, foi testado a utilização do classificador Support Vector Machine (SVM) ao invés do tradicional *softmax* na última camada, e o Support Vector Machine alcançou resultados melhores.

3.4 Aplicações

Há diversas aplicações para o reconhecimento de emoção no mundo real, foi percebido que os pesquisadores de reconhecimento de emoção por expressão facial utilizando RNC, ultimamente concentraram seus esforços mais no desenvolvimento de reconhecedores de emoção do que a aplicação em cenários reais, mesmo assim, está aberto para trabalhos futuros inúmeras aplicações desses reconhecedores em diversas áreas, tendo destaque principalmente para:

Tabela 3 – Principais arquiteturas de Redes Neurais de Convolução para reconhecimento de expressões faciais. (*) Significa que a rede foi treinada (*fine-tuning*) por duas vezes.

Arquitetura	Trabalho	Base de Treino	Base de Validação	Acurácia
AlexNet	Chen et al. (2017b)	CK+	CK+	99.1%
		CK+	JAFFE	83.11%
		JAFFE	JAFFE	87.7%
	Shan et al. (2017)	JAFFE	JAFFE	76.7%
		CK+	CK+	80.3%
	Kim et al. (2016b)	FER	FER	73.73%
	Huang and Lu (2016)	FER	FER	76.9%
		CK+	CK+	97.3%
	Vo and Le (2016)	CK+	CK+	96.04%
	Yu et al. (2016c)	CK+	CK+	98.7%
		MMI	MMI	98.6%
	Ng et al. (2015)*	FER/EmotiW	FER/EmotiW	55.6%
	Jung et al. (2015b)	CK+/FER	CK+/FER	86.54%
	Li et al. (2015b)	CIFE	CIFE	81.5%
		CK+	CK+	83%
VGG	Barsoum et al. (2016b)	FER+	FER+	84.9%
	Ng et al. (2015)*	FER/EmotiW	FER/EmotiW	52.1%
GoogLeNet	Guo et al. (2016)	FER/SFEW2.0	FER/SFEW2.0	71.3%
Ensemble	Wen et al. (2017b)	FER	FER-Private	69.96%
		FER	CK+	76.05%
		FER	JAFFE	50.70%
		FER	EmotiW	34.09%
	Liu et al. (2016b)	FER	FER	65.03%
	Shin et al. (2016b)	FER/SFEW	FER-Test	66.67%
			SFEW	64.84%
			CK+	65.54%
			KDEF	50.66%
			JAFFE	49.17%

- Interação humano computador (Barsoum et al., 2016b; Chen et al., 2017b; Liu et al., 2016b; Wen et al., 2017b), onde pode ser possível projetar interfaces que se adaptam ao estado emocional do usuário;
- Psiquiatria e cuidados médicos (Chen et al., 2017b; Mayya et al., 2016; Wen et al., 2017b), no qual o reconhecedor de emoção deve monitorar constantemente o paciente ou usuário fornecendo dados emocionais que podem contribuir para diagnósticos;
- Deficiente visual (Li et al., 2015b), pois pessoas com alto grau de deficiência visual, tem dificuldades na interação entre pessoas para identificar qual a emoção que as pessoas em volta estão emitindo;
- Interação humano robô (Jung et al., 2015b; Shin et al., 2016b), fazendo com que robôs estejam habilitados a interagir com humanos podendo adaptar-se a emoção

dos humanos em volta, ou até mesmo emitir emoção se aproximando de um humanoide;

- Personagens virtuais e animação (Vo and Le, 2016; Yu et al., 2016c), habilitando avatares a copiar expressão humana que podem ser útil para gravações de filmes de animação, também pode ser usado em aplicações de animação como o popular aplicativo para *smartphone* o *Snapchat*, que identifica a expressão facial do usuário e retorna alguma animação sobrepondo a expressão anteriormente detectada do usuário.

3.5 Resumo

Neste capítulo foram apresentado os trabalhos relacionados, verificamos que o tema está em crescente investigação pela comunidade, visto que arquiteturas poderosas de RNC surgiram recentemente. Apesar de uma RNC ter embutido o pré-processamento em sua própria arquitetura, para o problema tratado por este trabalho, verificamos que pré-processamento adicionais (não originais da RNC) melhoraram consideravelmente a taxa de acurácia. É importante frisar que as principais arquiteturas empregadas tem sido as vencedoras ou com pouca variação do tradicional problema de aprendizado profundo: o desafio ILSVRC. Estas arquiteturas tem se saído bem no problema de reconhecimento de emoção alcançando resultados comparado a nível humano.

Um conjunto de aplicações foram identificadas para o reconhecimento de emoção por expressão facial, entretanto, investigando a literatura há um índice baixo de adesão em cenários de uso reais. Atualmente, os pesquisadores apenas tem falado que é possível usar em uma determinada área, mas de fato não o experimentaram na prática. Contudo, modelos de reconhecimento de emoção tem alcançado taxas de acurácia confiáveis para serem empregados na indústria e pesquisa.

4 Abordagem Proposta

Neste capítulo uma abordagem é proposta para reconhecer emoções por meio da expressão facial e está dividido da seguinte forma. Na Seção 4.1 descreve um módulo de detecção de face e recorte. Na Seção 4.2 aborda as operações de pré-processamento aplicadas na imagem. Na Seção 4.3 apresenta o classificador de expressões faciais e por fim um resumo do capítulo na Seção 4.4.

4.1 Detecção de Face e Recorte

Este procedimento consiste na detecção de todas as faces de uma imagem por meio do algoritmo Viola Jones (consultar Seção 2.5.1) gerando um conjunto de coordenadas para criar um retângulo indicando a localização da face. Vale ressaltar que esta atividade possui complexidade moderada, pois uma imagem contém vários objetos com diferentes geometrias, inclusive podendo assemelhar-se a uma face acarretando na geração de falsos positivos. Logo após a detecção de face é realizado o recorte utilizando o conjunto de coordenadas definido pela etapa anterior, tal atividade é valorosa para exclusão do *background*. Desta forma, é enviado somente a face recortada para a etapa de pré-processamento reduzindo a complexidade do problema, pois não há necessidade do classificador aprender a separar o *background* da face. Posteriormente, ao recorte da face, a imagem original que deve está com uma face recortada é mantida para nova averiguação de recorte de face. Caso exista outras faces na imagem, este processo é repetido até não existir mais faces para recortar. Obviamente caso seja enviada uma imagem para a etapa de detecção e recorte que não contém uma face (e.g. imagem de um avião) o processo é automaticamente encerrado, pois se não há uma face para detectar, logo não há uma expressão facial emocional para reconhecer.

4.2 Pré-Processamento

Uma face recortada é enviada pelo módulo de Detecção de Face e Recorte para a fase de Pré-processamento. Nesta etapa são aplicadas operações de pré-processamento que realçam características relevantes que diferenciam as expressões faciais com intuito de preparar a imagem para a classificação. Inicialmente, uma função de redimensionamento é chamada para transformar a imagem em uma escala de 60x60 *pixels*, como a imagem é colorida, isto é, possui 3 canais denominados RGB (do inglês: Red, Green e Blue) a imagem resultante possui 10.800 características que pode ser calculada por $Qtd\_Caracteristica = N\_Pixels\_X * N\_Pixels\_Y * N\_Canais$. Após o redimensionamento, é realizado a

normalização da imagem dividindo cada *pixel* por 255, isto é, o valor máximo que um *pixel* pode possuir resultando na normalização no intervalo de 0 a 1.

O problema visualizado por esta proposta consiste em classificar emoções em qualquer ambiente. É sabido que pela variação de ambiente que há diferença na intensidade da luz, ocorrendo a perda de características importantes da face que diferenciam as emoções. Vale destacar que esta proposta é baseada principalmente em redes neurais de convolução que originalmente possui vários filtros de pré-processamento de imagem. Entretanto, a literatura tem mostrado que filtros clássicos aplicados antes da inserção de uma imagem na rede tem sido eficazes na eliminação de ruídos, principalmente aqueles relacionados a iluminação e brilho. Assim, a imagem resultante ressaltará melhor os traços faciais, além da imagem transformada está com maior nitidez para a rede neural de convolução. Portanto, as técnicas de normalização de brilho e iluminação é parte da etapa de pré-processamento.

4.3 Rede Neural de Convolução

A rede neural de convolução é a parte central e mais importante desta abordagem em discussão. Por meio dela a imagem é processada enfatizando contornos, padrões, formas e características da imagem relevantes para a classificação. Além disso, funções são aplicadas para redução de dimensionalidade e normalização ocasionando que a rede não seja sensível a rotações, posições e escala da imagem. Tais aptidões são essenciais para um classificador de imagens que deve ser usado em cenários reais maximizando a generalização. Entretanto, um desempenho satisfatório da rede neural de convolução, assim como de qualquer algoritmo supervisionado de aprendizagem de máquina, está estritamente relacionado ao processo de treinamento e validação do modelo.

4.3.1 Treinamento

O treinamento da rede neural de convolução é parte fundamental para o reconhecedor de emoção alcançar generalização satisfatória e adquirir aprendizado suficiente para funcionar em variados ambientes. Para isso, o treinamento é apoiado pela técnica de aumento de dados com intuito de maximizar a generalização do aprendizado durante o treinamento. O aumento de dados consiste na multiplicação das imagens em tempo dinâmico modificando um pouco a imagem e seu contexto alterando as imagens de treinamento aplicando zoom, rotações, *blur*, *shear*, diferentes níveis de contraste e entre outros, resultando em maior aprendizado e generalização do modelo.

4.3.2 Extração de Caracteríticas e Classificação

A rede neural de convolução recebe a imagem pré-processada de uma face para classificá-la estimando a probabilidade para cada emoção: neutralidade, raiva, felicidade,

tristeza, desprezo, medo e surpresa, de acordo com as caraterísticas extraídas da imagem. A extração de característica é um procedimento responsável em identificar as zonas da imagem que são mais relevantes para a separação do problema, isto é, classificar uma expressão facial (*e.g* o sorriso humano é uma expressão facial indicadora para a emoção felicidade). A extração de característica está embutida na rede neural de convolução que consiste nas camadas operando sobre a imagem de entrada ressaltando todos os contornos. Algumas camadas de convolução são especialistas na extração dos padrões verticais, outras nos horizontais, até que um conjunto de características são extraídas para o *softmax* classificar estimando a probabilidade para cada emoção.

4.3.3 Computação embarcada e em nuvem

Esta proposta visa fornecer soluções para reconhecimento de emoção que contempla dois tipos de computação: em nuvem e embarcada. Tais computações estão em alta na academia, indústria e mercado. A primeira pelo crescimento da internet havendo bilhões de dispositivos conectados e a evolução da infraestrutura com aumento considerável de recursos computacionais e velocidade de conexão. A segunda pela explosão de dispositivos embarcados presentes em nosso cotiano. Além disso, os dispositivos embarcados de hoje tem uma autonomia energética periódica, hardware semelhante a desktops, sistemas operacionais e sensores embutidos, formando um dispositivo independente e poderoso. Há no mercado smartphones e smartwatches com processadores octa-core e dual-core respectivamente, com a memória RAM chegando a 8GB, e até mesmo com placas de vídeos embutidas para aceleração de computação.

A computação em nuvem hospedaria o melhor modelo gerado a partir das arquiteturas AlexNet, VGGNet, GoogLeNet e Residuais, considerando métricas de avaliação de desempenho como precisão, revocação e f1-score. Apesar de que o melhor modelo possa exigir elevada utilização de recursos computacionais por ser uma rede neural profunda, entende-se que um serviço em nuvem possuiria um hardware robusto capaz de suportar a demanda da rede neural de convolucão. Visto que estamos no boom da computação cognitiva, isto é, a capacidade de computadores pensarem como humanos. A ideia de ter um reconhecedor de emoções em um serviço em nuvem é para quaisquer aplicação, independente de qual linguagem de programação foi implementada ou em qual sistema operacional está sendo executada, enviar imagens via padrão REST com intuito de receber a classificação das imagens com as emoções detectadas.

A computação embarcada consiste na rede neural de convolução baseada na arquitetura MobileNet funcionar nativamente em um dispositivo embarcado. Essa arquitetura foi projetada para consumir menos recursos computacionais com o compromisso de perder o mínimo de precisão, sendo ideal para dispositivos embarcados que dispõe de menos recursos computacionais. Inclusive podendo funcionar nativamente no sistema operacional

Android que é amplamente usado por *smartphones*, *smartwatches* e *tablets*. Além disso, a arquitetura MobileNet pode ser também embutida em placas de desenvolvimento como Raspberry, Nvidia Jetson, Drones e outras.

Vale destacar que as maiores taxas de ocupação de recursos computacional de uma rede neural de convolução estão na fase de treinamento, isto é, durante a geração do modelo. E a fase de classificação exige menores taxas de ocupação de hardware, pois o principal procedimento que demanda recursos computacionais consiste em carregar o modelo na memória. Quando uma imagem é enviada para classificação, considerando que o modelo está carregado na memória, a rede neural opera sobre a imagem aplicando os pesos advindos do modelo. Diferentemente da fase de treinamento que exige muito processamento, pois é executado o algoritmo de otimização gradiente descendente para minimizar a função de perda realizando uma alta quantidade de cálculos vetoriais.

4.4 Resumo

,

Referências

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. (2016a). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283. ACM.

Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. (2016b). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283. ACM.

Basili, V. R., Caldiera, G., and Rombach, H. D. (1994). Experience factory. *Encyclopedia of software engineering*.

Biolchini, J., Mian, P. G., Natali, A. C. C., and Travassos, G. H. (2005). Systematic review in software engineering. *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES*, 679(05):45.

Chen, X., Yang, X., Wang, M., and Zou, J. (2017a). Convolution neural network for automatic facial expression recognition. In *Applied System Innovation (ICASI), 2017 International Conference on*, pages 814–817. IEEE.

Chen, X., Yang, X., Wang, M., and Zou, J. (2017b). Convolution neural network for automatic facial expression recognition. In *Applied System Innovation (ICASI), 2017 International Conference on*, pages 814–817. IEEE.

Cruz, A., Leitão, G., Colonna, J., Silva, E., Barreto, R., and Primo, T. (2017). Framework para coleta e inferência de estados emocionais de alunos baseado em reconhecimento de expressões faciais. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 997.

Darwin, C. (1965). *The expression of the emotions in man and animals*, volume 526. University of Chicago press.

Ekman, P. and Friesen, W. V. (1977). Facial action coding system.

Ekman, P. E. and Davidson, R. J. (1994). *The nature of emotion: Fundamental questions.* Oxford University Press.

Géron, A. (2017). Hands-on machine learning with scikit-learn and tensorflow: concepts, tools, and techniques to build intelligent systems.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Guo, Y., Tao, D., Yu, J., Xiong, H., Li, Y., and Tao, D. (2016). Deep neural networks with relativity learning for facial expression recognition. In *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*, pages 1–6. IEEE.

Huang, Y. and Lu, H. (2016). Deep learning driven hypergraph representation for image-based emotion recognition. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 243–247. ACM.

Hubel, D. H. (1959). Single unit activity in striate cortex of unrestrained cats. *The Journal of physiology*, 147(2):226–238.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243.

Jain, A. K. (1989). *Fundamentals of digital image processing*. Prentice-Hall, Inc.

Jaques, P. A. and Nunes, M. A. S. (2013). Ambientes inteligentes de aprendizagem que inferem, expressam e possuem emoções e personalidade. *Jornada de Atualização em Informática na Educação*, 1(1):30–81.

Jung, H., Lee, S., Park, S., Kim, B., Kim, J., Lee, I., and Ahn, C. (2015a). Development of deep learning-based facial expression recognition system. In *Frontiers of Computer Vision (FCV), 2015 21st Korea-Japan Joint Workshop on*, pages 1–4. IEEE.

Jung, H., Lee, S., Park, S., Kim, B., Kim, J., Lee, I., and Ahn, C. (2015b). Development of deep learning-based facial expression recognition system. In *Frontiers of Computer Vision (FCV), 2015 21st Korea-Japan Joint Workshop on*, pages 1–4. IEEE.

Kim, B.-K., Dong, S.-Y., Roh, J., Kim, G., and Lee, S.-Y. (2016a). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 48–57.

Kim, B.-K., Dong, S.-Y., Roh, J., Kim, G., and Lee, S.-Y. (2016b). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 48–57.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, W., Li, M., Su, Z., and Zhu, Z. (2015a). A deep-learning approach to facial expression recognition with candid images. In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, pages 279–282. IEEE.

Li, W., Li, M., Su, Z., and Zhu, Z. (2015b). A deep-learning approach to facial expression recognition with candid images. In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, pages 279–282. IEEE.

Liu, K., Zhang, M., and Pan, Z. (2016a). Facial expression recognition with cnn ensemble. In *Cyberworlds (CW), 2016 International Conference on*, pages 163–166. IEEE.

Liu, K., Zhang, M., and Pan, Z. (2016b). Facial expression recognition with cnn ensemble. In *Cyberworlds (CW), 2016 International Conference on*, pages 163–166. IEEE.

Mafra, S. N. and Travassos, G. H. (2006). Estudos primários e secundários apoiando a busca por evidência em engenharia de software. *Relatório Técnico, RT-ES*, 687(06).

Mayya, V., Pai, R. M., and Pai, M. M. (2016). Automatic facial expression recognition using dcnn. *Procedia Computer Science*, 93:453–461.

Nasoz, F., Alvarez, K., Lisetti, C. L., and Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, 6(1):4–14.

Ng, H.-W., Nguyen, V. D., Vonikakis, V., and Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449. ACM.

Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., Gaggioli, A., Botella, C., and Alcañiz, M. (2007). Affective interactions using virtual reality: the link between presence and emotions. *CyberPsychology & Behavior*, 10(1):45–56.

Scherer, K. R. (2000). Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162.

Shan, K., Guo, J., You, W., Lu, D., and Bie, R. (2017). Automatic facial expression recognition based on a deep convolutional-neural-network structure. In *Software Engineering Research, Management and Applications (SERA), 2017 IEEE 15th International Conference on*, pages 123–128. IEEE.

Shin, M., Kim, M., and Kwon, D.-S. (2016a). Baseline cnn structure analysis for facial expression recognition. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 724–729. IEEE.

Shin, M., Kim, M., and Kwon, D.-S. (2016b). Baseline cnn structure analysis for facial expression recognition. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 724–729. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

Vo, D. M. and Le, T. H. (2016). Deep generic features and svm for facial expression recognition. In *Information and Computer Science (NICS), 2016 3rd National Foundation for Science and Technology Development Conference on*, pages 80–84. IEEE.

Wen, G., Hou, Z., Li, H., Li, D., Jiang, L., and Xun, E. (2017a). Ensemble of deep neural networks with probability-based fusion for facial expression recognition. *Cognitive Computation*, pages 1–14.

Wen, G., Hou, Z., Li, H., Li, D., Jiang, L., and Xun, E. (2017b). Ensemble of deep neural networks with probability-based fusion for facial expression recognition. *Cognitive Computation*, pages 1–14.

Yu, X., Huang, J., Zhang, S., and Metaxas, D. N. (2016a). Face landmark fitting via optimized part mixtures and cascaded deformable model. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2212–2226.

Yu, X., Lin, Z., Brandt, J., and Metaxas, D. N. (2014). Consensus of regression for occlusion-robust facial feature localization. In *European Conference on Computer Vision*, pages 105–118. Springer.

Yu, X., Yang, J., Luo, L., Li, W., Brandt, J., and Metaxas, D. (2016b). Customized expression recognition for performance-driven cutout character animation. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE.

Yu, X., Yang, J., Luo, L., Li, W., Brandt, J., and Metaxas, D. (2016c). Customized expression recognition for performance-driven cutout character animation. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE.

Anexos

ANEXO A – Revisão Sistemática da Literatura

Neste anexo é descrito e discutido uma Revisão Sistemática da Literatura acerca do tema deste trabalho. Na Seção A.1 é descrito o protocolo seguido para a realização da revisão sistemática. Na Seção A.2 está o processo de condução da revisão sistemática e quantos artigos veio em cada filtro. Na Seção A.3 contém os resultados obtidos, assim como, as respostas para as questões de pesquisa e por fim na Seção A.4 o resumo e outras discussões sobre esta revisão sistemática da literatura.

A.1 Protocolo da Revisão Sistemática da Literatura

Este protocolo foi elaborado conforme especificado em: Biolchini et al. (2005), Mafra and Travassos (2006), e Kitchenham (2004):

A.1.1 Objetivo

O objetivo deste estudo será esquematizado a partir do paradigma GQM (goal, question, and metric) (Basili et al., 1994):

A.1.2 Questões de Pesquisa

Questão Principal: Como reconhecer emoções por meio da expressão facial utilizando redes neurais de convolução em uma imagem estática?

- **Q1**: Quais emoções têm sido reconhecidas por meio da expressão facial utilizando redes neurais de convolução?
- **Q2**: Quais tipos de pré-processamento tem sido realizado na imagem?
- **Q3**: Quais arquiteturas de redes de convolução têm sido mais utilizadas?

Tabela 4 – Objetivos da Revisão Sistemática

Analisar	Reconhecimento de emoções por meio da expressão facial em uma imagem estática.
Com o propóstio de	Identificar técnicas, métodos, abordagens, arquiteturas, base de dados e aplicações.
No que diz respeito a	Utilização de redes neurais de convolução.
Do ponto de vista do	Pesquisador.
No contexto	Acadêmico.

- **Q4**: Quais técnicas, métodos e abordagens têm sido utilizados para tratar problemas na imagem como iluminação, rotação, obstrução e escala?
- **Q5**: Quais bases de dados têm sido utilizadas?
- **Q6**: Quais aplicações podem utilizar o reconhecimento de emoção por expressão facial?

A.1.3 Biblioteca Digital

Scopus: http://www.scopus.com/ - Contempla as principais conferências da área (foi verificado por meio de busca manual)

A.1.4 Critérios de Inclusão e Exclusão dos Artigos

Critérios de Inclusão:

- **CI1:** Reconhecimento de emoção por expressão facial usando somente CNN com abordagem que funciona para classificação em imagem;
- **CI2:** Reconhecimento de emoção por expressão facial combinando CNN com várias arquiteturas de redes neurais com abordagem que funciona para classificação em imagem;
- **CI3:** Reconhecimento de emoção por expressão facial combinando CNN com outros métodos de aprendizado de máquina que funciona para classificação em imagem;
- **CI4:** Reconhecimento de emoção por expressão facial combinando CNN com técnicas de pré-processamento que não são originais da arquitetura CNN com abordagem que funciona para classificação em imagem.

Critérios de Exclusão:

- **CE1:** Trabalho somente apresenta teoria ou discussão relacionada ao reconhecimento de emoções;
- **CE2:** Não apresenta reconhecimento de emoção por expressão facial para classificação em imagens;
- **CE3:** Não utiliza redes neurais de convolução;
- **CE4:** Trabalho anterior ao ano de 2013;
- **CE5:** Reconhecimento por vídeo ou *streaming* de imagens;

- **CE6:** Trabalho utiliza durante a metodologia experimental uma base de dados não disponível para a comunidade científica;
- **CE7:** Publicação não disponível;
- **CE8:** Reconhecimento de emoção multimodal.

Observação: O *CE4* foi definido devido o surgimento das (atuais) redes neurais de convolução ter sido a partir de 2013.

A.1.5 Formulário de Extração de Informação

Inicialmente, no primeiro filtro serão analisados e considerados os seguintes itens:

- Título;
- Resumo;
- Palavras-chaves.

Posteriormente, no segundo filtro serão extraídas as seguintes informações:

- Autores do trabalho;
- Fonte: local que o trabalho foi publicado;
- Ano de Publicação;
- Emoções que foram reconhecidas;
- Aplicações para o reconhecimento de emoções por expressão facial;
- Arquiteturas de redes neurais de convolução utilizadas;
- Metodologia utilizada para o treinamento da rede neural de convolução;
- Base de dados utilizadas para treino e validação;
- Perspectivas futuras;
- Comentários.

A.1.6 *String* de Busca

As *strings* de busca foram definidas a partir das questões de pesquisa e do padrão PICO (*population*, *intervention*, *comparison*, *outcomes*) (KITCHENHAM e CHARTERS, 2007), conforme a estrutura abaixo:

- **População:** Reconhecimento de emoção por expressão facial;
- **Intervenção:** Por meio de redes neurais de convolução;
- **Comparação:** Não há;
- **Resultados:** Técnicas, métodos, arquiteturas, base de dados, aplicações e abordagens.

  ("emotion recognition"OR "emotion detection"OR "emotion identification"OR "emotion analysis"OR "emotion classification"OR "affect recognition"OR "affect detection"OR "affect analysis"OR "affect classification"OR "facial expression")

  AND

  ("convolutional neural network"OR "CNN"OR "long short term memory"OR "LSTM"OR "recurrent neural network"OR "RNN")

  AND

  ("technique"OR "method"OR "architecture"OR "database"OR "application"OR "approach")

Para a montagem da string de busca foi testado cada termo da população com todos os termos da intervenção mais resultado, desta forma, verificando e validando que todos os termos da população realmente contribuem para os artigos retornados. Este mesmo procedimento foi utilizado para verificar e validar os termos do resultado, no entanto foi verificado cada termo do resultado com todos os termos da intervenção e população.

A presença dos seguintes termos na intervenção: "long short term memory", "LSTM", "recurrent neural network"e "RNN", podem ser explicados devido à intuição do autor deste protocolo acreditar que a comunidade estava utilizando essas técnicas, que também são redes neurais profundas, combinadas com as redes neurais de convolução para classificação de expressões faciais em imagens.

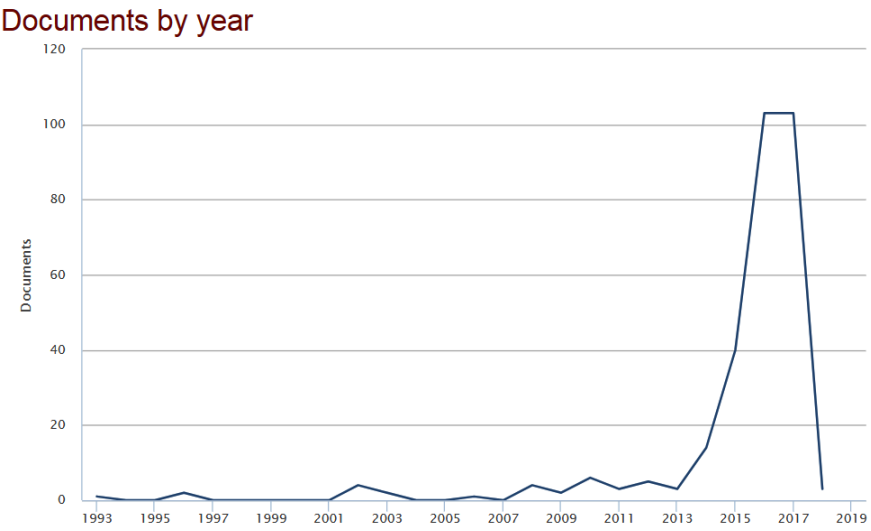

Figura 12 – Artigos por ano retornados pela *string* de busca

A.2 Condução da Revisão Sistemática da Literatura

A.2.1 Primeiro Filtro

No primeiro filtro foi lido somente o título, resumo e as palavras-chaves do artigo. A *string* de busca retornou 281 artigos para classificar no primeiro filtro. Foram aceitos 99 (35%) para o segundo filtro, 3 (1%) duplicados e 179 (64%) rejeitados.

Na Seção A.1.6, o autor do protocolo esclarece porque utilizou os seguintes termos na *string* de busca: "long short term memory", "LSTM", "recurrent neural network"e "RNN", depois do primeiro filtro, realmente comprovou-se que estas técnicas são combinadas com as redes neurais de convolução para classificação de emoção em expressão facial, porém, somente em vídeos ou streaming de imagens. Portanto, os artigos retornados por essas palavras receberam a classificação de rejeitado devido a esta revisão focar em trabalhos com classificação em imagens estática sem streaming.

A.2.2 Segundo Filtro

Para a realização do segundo filtro foi lido o artigo completo para a extração dos dados e, consequentemente a obtenção dos resultados. No segundo filtro tinham 99 artigos para classificar, onde 34 foram aceitos, 1 duplicados e 64 rejeitados.

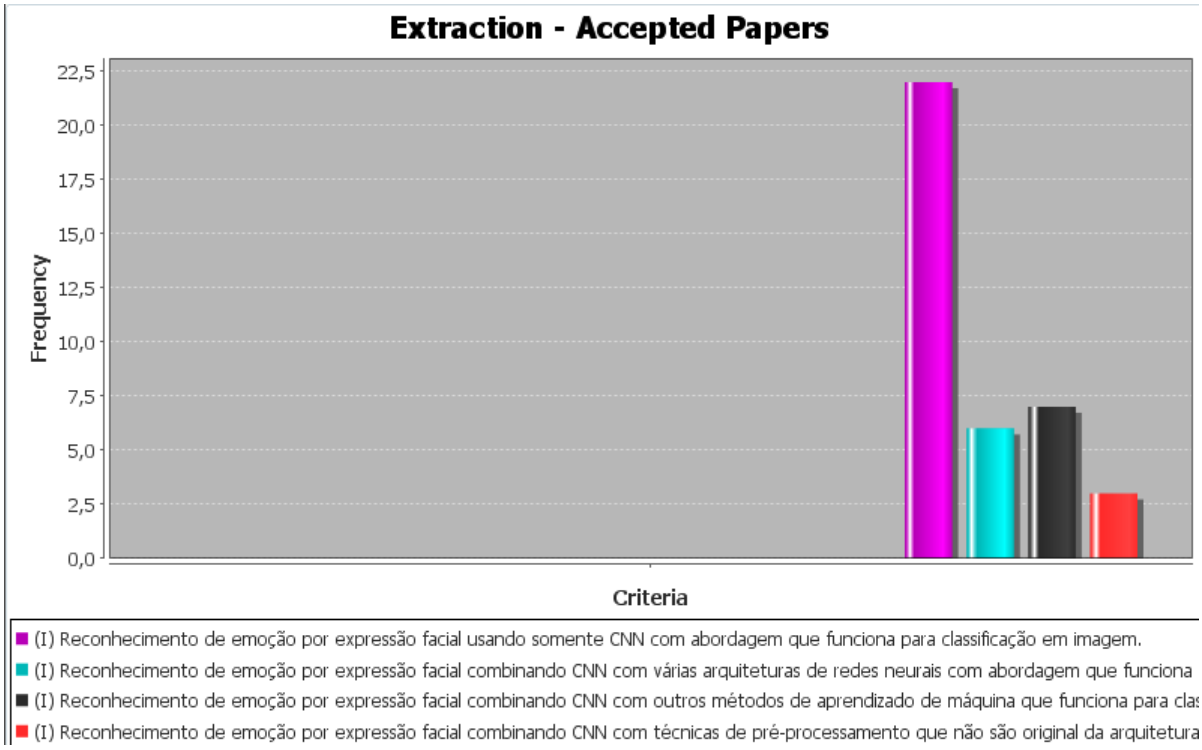

Figura 13 – Gráfico representando a frequência dos critérios de inclusão para os artigos aceitos do segundo filtro

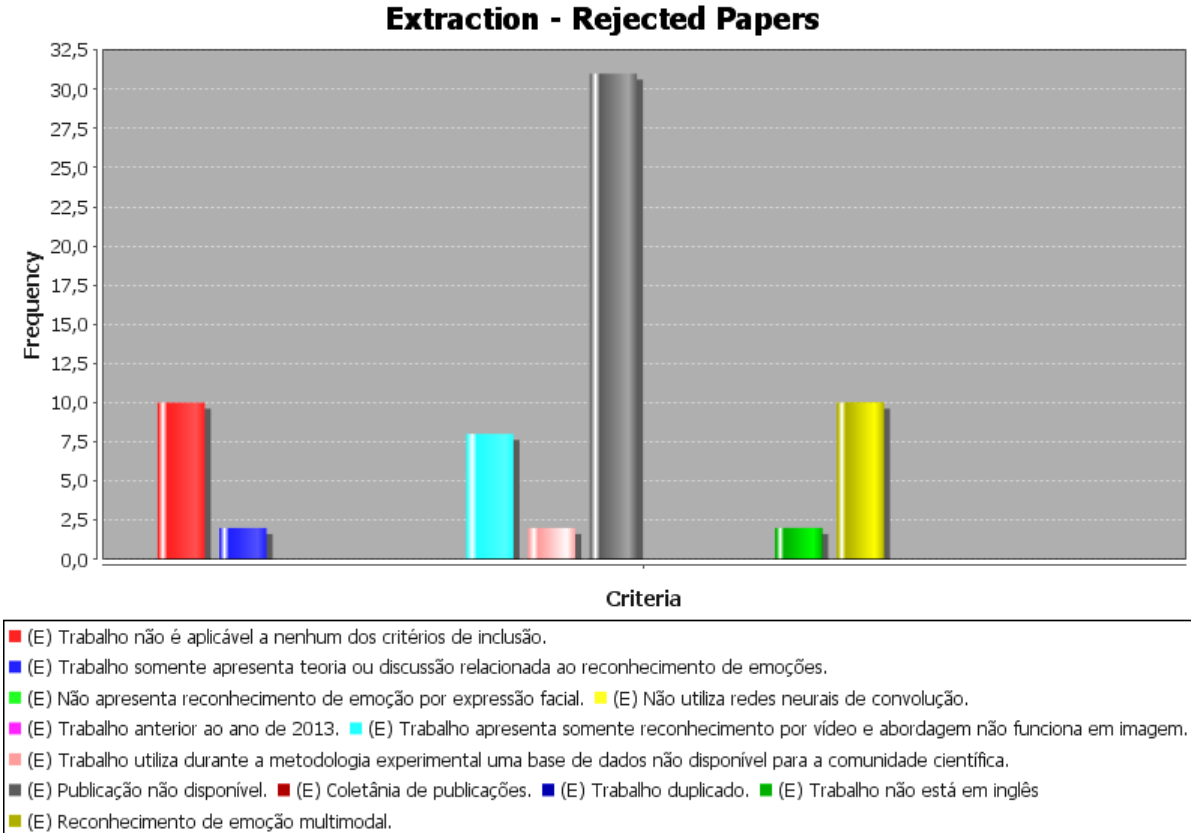

Figura 14 – Gráfico representando a frequência dos critérios de exclusão para os artigos rejeitados do segundo filtro

A.3 Resultados

A.3.1 Q1: Quais emoções têm sido reconhecidas por meio da expressão facial utilizando redes neurais de convolução?

As emoções reconhecidas obviamente são as emoções que são representadas e contidas em base de dados, logo são as emoções básicas: neutralidade, felicidade, surpresa, tristeza, raiva, desgosto e medo. Alguns trabalhos como Barsoum et al. (2016b) reconhece também nojo.

A.3.2 Q2: Quais tipos de pré-processamento tem sido realizado na imagem?

Abaixo segue uma lista com o pré-processamento e os trabalhos que fizeram utilização da técnica.

- **Detector de Face**: Consiste na detecção e recorte da face (Chen et al., 2017b; Li et al., 2015b; Mayya et al., 2016; Ng et al., 2015; Shan et al., 2017; Shin et al., 2016b; Vo and Le, 2016);
- **Normalização de Brilho (Equalização de histograma)**: Transformada para realce do contraste (Kim et al., 2016b; Shan et al., 2017; Shin et al., 2016b);
- **Normalização Min e Max**: Transformação linear baseada no valor mínimo e máximo da imagem (Kim et al., 2016b);
- **Pontos da Face (pontos geométricos)**: Extração de pontos da face e a distância entre os pontos (Yu et al., 2016c);
- **Escala de Cinza**: Transformação da imagem para escala de cinza (Mayya et al., 2016);
- **Diferença Gaussiana**: Detecta as bordas do objeto, neste caso, evidencia as bordas da face (Shin et al., 2016b);
- **Filtro de Difusão Isotópica**: (Shin et al., 2016b);
- **Normalização DCT**: Transformada discreta do cosseno utilizada em compressão de dados e eventualmente evidenciando informações relevantes da imagem (Shin et al., 2016b);
- **Alinhamento da Face**: Utilização de uma rede neural *autoencoder* para alinhar a face no centro (Kim et al., 2016b).

Arquitetura	Trabalhos que utilizaram a arquitetura
AlexNet	Chen et al. (2017b), Shan et al. (2017), Kim et al. (2016b), Huang and Lu (2016), Vo and Le (2016), Yu et al. (2016c), Ng et al. (2015), Jung et al. (2015b), Li et al. (2015b)
GoogLeNet	Guo et al. (2016)
VGG	Barsoum et al. (2016b), Ng et al. (2015)
Ensemble	Wen et al. (2017b), Liu et al. (2016b), Shin et al. (2016b)

Tabela 5 – Principais arquiteturas de redes neurais de convolução e os trabalhos que utilizaram

A.3.3 Q3: Quais arquiteturas de redes neurais de convolução têm sido mais utilizadas?

A Tabela 5 contém as arquiteturas encontradas e os trabalhos que a utilizaram, como podemos verificar a arquitetura AlexNet foi a mais utilizada.

A.3.4 Q4: Quais técnicas, métodos e abordagens têm sido utilizados para tratar problemas na imagem como iluminação, rotação, obstrução e escala?

A principal solução encontrada na literatura foi a ênfase na generalização adequada do aprendizado da rede neural de convolução, isto é, durante a fase de treinamento. Obviamente, as técnicas aplicadas no pré-processamento da imagem (ver Seção A.3.2), contribuem para resolver problemas de iluminação por meio da equalização de histograma e de rotação com o alinhamento de face. Entretanto, um achado bastante interessante foi a utilização da técnica de aumento de dados, que consiste durante a fase de treinamento da rede neural de convolução em multiplicar por 10 vezes uma instância (imagem), isto é, gerando 10 novas imagens com pequenos giros da faces, e variações da rotação da pose, escala e iluminação. Acrescendo em 10 vezes o tamanho da base de treinamento, com inserção de variações da imagem resultando em melhor aprendizado da rede.

A.3.5 Q5: Quais bases de dados têm sido utilizadas?

Esta seção tem enfoque nas bases de dados mapeadas para reconhecimento de emoção por expressão facial em uma imagem estática. Obviamente, é possível encontrar outras base de dados para reconhecimento de emoção que não seja por imagem estática, por exemplo, reconhecimento em vídeo, por sensores, em textos e outras.

As bases de dados para reconhecimento de emoção por expressão facial em uma imagem estática tem algo em comum, geralmente as amostras de expressões faciais são as mesmas emoções, as chamadas emoções básicas investigadas por Ekman and Davidson (1994) que são: neutralidade, felicidade, medo, desgosto, raiva, surpresa e tristeza, isto

Bases de Dados	Trabalhos que utilizaram a base para treinamento ou validação
CK+	Chen et al. (2017b), Shan et al. (2017), Wen et al. (2017b), Shin et al. (2016b), Huang and Lu (2016), Vo and Le (2016), Yu et al. (2016c), Mayya et al. (2016), Jung et al. (2015b), Li et al. (2015b)
JAFFE	Chen et al. (2017b), Shan et al. (2017), Wen et al. (2017b), Shin et al. (2016b), Mayya et al. (2016)
FER	Wen et al. (2017b), Kim et al. (2016b), Liu et al. (2016b), Shin et al. (2016b), Huang and Lu (2016), Guo et al. (2016), Ng et al. (2015), Jung et al. (2015b)
FER+	Barsoum et al. (2016b)
SFEW2.0	Shin et al. (2016b), Guo et al. (2016)
KDEF	Shin et al. (2016b)
MMI	Yu et al. (2016c)
CIFE	Li et al. (2015b)
EmotiW2015	Wen et al. (2017b), Ng et al. (2015)

Tabela 6 – Bases de Dados

significa que são essas as emoções que a comunidade tem reconhecido por expressão facial. As bases de dados mapeadas podem ser consultadas na Tabela 6.

A.3.6 Q6: Quais aplicações podem utilizar o reconhecimento de emoção por expressão facial?

Há diversas aplicações para o reconhecimento de emoção no mundo real, foi percebido que os pesquisadores de reconhecimento de emoção por expressão facial utilizando rede neural de convolução, ultimamente concentraram seus esforços mais no desenvolvimento de reconhecedores de emoção do que a aplicação em cenários reais, mesmo assim, está aberto para trabalhos futuros inúmeras aplicações desses reconhecedores em diversas áreas, tendo destaque principalmente para:

- **Interação humano computador**: no qual pode ser possível projetar interfaces que se adaptam ao estado emocional do usuário (Barsoum et al., 2016b; Chen et al., 2017b; Liu et al., 2016b; Wen et al., 2017b);
- **Psiquiatria e cuidados médicos**: no qual o reconhecedor de emoção deve monitorar constantemente o paciente ou usuário fornecendo dados emocionais que podem contribuir para diagnósticos (Chen et al., 2017b; Mayya et al., 2016; Wen et al., 2017b);
- **Deficiente visual**: pois pessoas com alto grau de deficiência visual, tem dificuldades na interação entre pessoas para identificar qual a emoção que as pessoas em volta estão emitindo (Li et al., 2015b),;
- **Interação humano robô**: fazendo com que robôs estejam habilitados a interagir com humanos podendo adaptar-se a emoção dos humanos em volta, ou até mesmo

emitir emoção se aproximando de um humanoide (Jung et al., 2015b; Shin et al., 2016b);

- **Personagens virtuais e animação**: habilitando avatares a copiar expressão humana que podem ser útil para gravações de filmes de animação, também pode ser usado em aplicações de animação como o popular aplicativo para *smartphone* o *Snapchat*, que identifica a expressão facial do usuário e retorna alguma animação sobrepondo a expressão anteriormente detectada do usuário (Vo and Le, 2016; Yu et al., 2016c).

A.3.7 Questão Principal: Como reconhecer emoções por meio da expressão facial utilizando redes neurais de convolução em uma imagem estática?

Diante das fichas de extração, foi percebido que a comunidade explorou diversas estratégias para processar imagens de expressão facial e reconhecer emoção. Algumas abordagens se destacam como: a técnica para aumentar os dados de treinamento e teste, utilizando a técnica flip fazendo até 10 pequenas rotações na imagem, para a CNN aprender a generalizar melhor sendo treinada e testada com uma base de dados maior. A técnica de normalização de brilho (equalização) no pré-processamento, no qual todos os trabalhos que utilizaram esta técnica aumentaram a acurácia do reconhecimento. Também merece destaque o trabalho de Kim et al. (2016a) em que na sua abordagem, a rede de convolução recebe duas expressões faciais de entrada: a saída de um autoencoder que alinha a face e a imagem original sem alinhamento da face. Esta abordagem melhorou bastante o reconhecimento.

Com relação à arquitetura da rede neural de convolução, quem utilizou um SVM como classificador ao invés de um tradicional softmax obteve maior acurácia. Teve trabalhos que utilizou uma rede com camadas inceptions, hipergrafo, ensembles e concatenação de redes, e todas essas abordagens superaram uma CNN simples. Neste caso, falta um trabalho que possa dizer experimentalmente qual dessas arquiteturas é a melhor.

Percebemos que existem várias bases de dados disponíveis para a comunidade. As bases de dados que foram mais exploradas foram a CK+, FER2013 e a JAFFE. A base CK+ é composta por expressões faciais capturadas em laboratório, por isso, tem altas taxas de reconhecimento, pois, sua dificuldade para o reconhecimento diminui. Já a base FER2013 foi capturada na "natureza", por isso sua taxa de reconhecimento é mais baixa sendo uma base bastante complexa para classificação.

Notoriamente os trabalhos utilizam o algoritmo Viola Jones para detecção de face pelo programa OpenCV e fazem o recorte da face excluindo o background. Desta forma, elimina o trabalho da rede em aprender a separar o que é background e o que é face, diminuindo a complexidade da classificação.

Portanto, para reconhecer emoção em uma imagem estática, é necessário o treinamento de uma rede de convolução com um classificador na última camada, com a maior quantidade de dados possível, realizando o recorte da face e utilizar técnicas de normalização na imagem, ocasionando um aumento da taxa de reconhecimento.

A.4 Resumo

Neste anexo apresentou uma revisão sistemática da literatura que investigou o estado-da-arte sobre o reconhecimento de emoção por expressão facial por meio de redes neurais de convolução. Verificamos que o tema está bem quente na comunidade, pois, antes de 2013 a String de busca retornou 33 artigos, e em 2013 (3 artigos), 2014 (14 artigos), 2015 (40 artigos), 2016 (103 artigos), 2017 (103 artigos) e 2018 (3 artigos), isso demonstra o crescimento exponencial da área.

Foram mapeadas as principais técnicas de pré-processamento, arquitetura de rede neural de convolução, base de dados, metodologias de treinamento e aplicações do reconhecimento de emoção. A impressão que fica é que a comunidade ainda não está utilizando esses classificadores no mundo real, e o amadurecimento rápido da área depois do surgimento do aprendizado profundo, nos levar acreditar que esses sistemas já estão prontos para ser posto em prática apoiando outras aplicações de interação humano computador, interação humano robô, educação, segurança, computação afetiva e etc.