Universidade Federal do Amazonas – UFAM

Instituto de Computação – IComp

Programa de Pós-Graduação em Informática – PPGI

Uma Abordagem para Reconhecimento de Emoção por Expressão Facial baseada em Redes Neurais de Convolução

Anderson Araújo da Cruz

Manaus – AM

Agosto, 2019

Anderson Araújo da Cruz

Uma Abordagem para Reconhecimento de Emoção por Expressão Facial baseada em Redes Neurais de Convolução

Dissertação submetida ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas para obtenção de título de mestrado *stricto sensu*.

Orientador: Raimundo da Silva Barreto, D.Sc.

Manaus – AM

Agosto, 2019

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Cruz, Anderson Araújo da
C957u Uma abordagem para reconhecimento de emoção por expressão facial baseada em redes neurais de convolução / Anderson Araújo da Cruz. 2019
120 f.: il. color; 31 cm.

Orientador: Raimundo da Silva Barreto
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Reconhecimento de Emoção. 2. Expressão Facial. 3. Redes Neurais de Convolução. 4. Computação Afetiva. 5. Detecção de Afeto. I. Barreto, Raimundo da Silva II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Uma Abordagem para Reconhecimento de Emoção por Expressão Facial baseada em Redes Neurais de Convolução"

ANDERSON ARAÚJO DA CRUZ

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Raimundo da Silva Barreto - PRESIDENTE

Profa. Elaine Harada Teixeira de Oliveira - MEMBRO INTERNO

Prof. Daniel Lins da Silva - MEMBRO EXTERNO

Manaus, 02 de Agosto de 2019

Agradecimentos

Inicialmente, agradeço ao dom da vida, às dádivas e benevolências recebidas. Muitos dizem ser sinal da presença de Deus neste mundo e em nossas vidas. Sendo assim, a Ele toda glória, honra e louvor.

Agradeço ao meu orientador e querido amigo, professor Raimundo Barreto pelo tempo dedicado, conselhos, conversas, companheirismo e confiança. O senhor me acompanhou desde o início da minha jornada nesta área, contribuindo significativamente para minha formação profissional e humana. Espero ter correspondido às inúmeras oportunidades concedidas. Tamo juntos!

Agradeço à minha esposa, Giselle Almeida, Gi, pelo amor verdadeiro, companheirismo, paciência, carinho e zelo. Iniciamos esta jornada juntos, em que há muito tempo atrás, você me ensinou meu primeiro "Hello World!". Obrigado por tudo meu amor! Amo você!

Agradeço à minha família, em especial ao Elisamar, Maria de Nazaré e Fátima, por todo amor, carinho, dedicação e cuidado. Nesta família amada, que aprendi a ser forte, perseverante, focado e outros valores essenciais para execução deste trabalho. Obrigado minha querida família por todos os momentos!

Agradeço à professora Elaine e ao professor Daniel pela participação da banca de defesa e por todo acompanhamento desta pesquisa com dicas, conselhos e apontamentos. Obrigado!

Agradeço à CAPES pelo suporte financeiro para execução deste trabalho. Parte dos resultados apresentados neste trabalho foram obtidos por meio do projeto de pesquisa “Sistemas para Avaliação de Comportamento e Recomendação Inteligente em Ambientes Educacionais e de Saúde Remota”, financiado pela Samsung Eletrônica da Amazônia Ltda., no âmbito da Lei no. 8.387 (art. 2o)/91.

Por fim, agradeço aos meus amigos e amigas que me apoiaram durante esta jornada em que trocamos inúmeras ideias para execução deste trabalho. Um agradecimento especial ao Juan Colonna, Gabriel Leitão, Edwin Juan, Márcio Alencar, Edson Silva, Romário Lira, Rafael Gurgel, Carlos Júnior, Alexandre Costa, Tiago Custódio e Juliana Postal.

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

Comercial Apple - "Pense Diferente", 1997

Resumo

Desenvolver a percepção emocional dos computadores é uma tendência tecnológica. O reconhecimento de emoção compõe sistemas cognitivos com aplicabilidade em diversas áreas. A expressão facial é uma maneira efetiva para reconhecer emoções, sobretudo por ser menos intrusiva na coleta de dados, quando comparada aos outros métodos, e pela facilidade de obter imagens da face diante da popularização das câmeras. Por meio das expressões faciais é possível classificar o grupo das emoções básicas (alegria, medo, surpresa, tristeza, desgosto e raiva) e neutralidade. Atualmente, as redes neurais de convolução (CNN) tem sido o estado da arte para classificação de imagens. Diante desse contexto, esta dissertação apresenta uma abordagem para reconhecer emoções por expressão facial utilizando CNN denominada como *Single Shot Facial Expression Recognition* (SSFER) e o seu uso em um estudo de caso. Inicialmente, um estudo experimental foi realizado para avaliar quatro detectores de faces em bases de expressões faciais e na VOC-2007. O método MMOD-CNN foi o melhor alcançando 91.89% de acurácia. Posteriormente, um outro estudo experimental foi conduzido a fim de comparar cinco arquiteturas de CNNs alternando quatro classificadores na última camada com intuito de classificar expressões faciais. As CNNs foram: VGGNet, InceptionResNetV2, InceptionV3, MobileNetV2 e ResidualNet, e os classificadores: Softmax, SVM, Random Forest e KNN. A ideia é que a CNN funcione como um extrator de características enviando um vetor unidimensional para o classificador definir a emoção. A melhor combinação foi a VGGNet com SVM alcançando 78.95% de acurácia. Desta forma, a abordagem proposta (SSFER) venceu com uma diferença de 9.74% de acurácia a API da *Microsoft Cognitive Services* em um comparação avaliando bases de expressões faciais. De um modo geral, as emoções alegria e surpresa foram as que tiveram maiores taxas de precisão. Em contrapartida, as emoções medo e raiva alcançaram as menores taxas de precisão. Um estudo de caso foi executado em um cenário real voltado para educação digital. Participaram vinte e sete estudantes do ensino médio com objetivo de responder um simulado do ENEM em uma plataforma digital. Durante a prova as expressões faciais dos estudantes foram coletadas, assim como, todas as interações com a plataforma. Após o simulado, as expressões faciais foram processadas para correlacionar com as interações de cliques e desempenho no teste. Análises de dados sugerem que a neutralidade pode estar relacionada ao estado de concentração e que estudantes passam a maior parte do tempo no estado de neutralidade. O estado de surpresa pode ser confundido aos bocejos possibilitando o reconhecimento de sonolência. E os estudantes que alcançaram as melhores notas no exame foram os que tiveram menor taxa de detecção de surpresa. Por fim, a abordagem proposta demostrou ser positiva para ser utilizada em aplicações gerais e, em particular, na educação digital.

Palavras-chaves: Reconhecimento de Emoção, Expressão Facial, Redes Neurais de Convolução, Computação Afetiva, Detecção de Afeto.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

RNC	Rede Neural de Convolução
CNN	Convolutional Neural Network
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
UA	Unidade de Ação
IA	Inteligência Artificial
AM	Aprendizagem de Máquina
VC	Visão Computacional
SE	Sistemas Embarcados
CN	Computação em Nuvem
OCR	Reconhecimento Ótico de Caracteres
ReLU	Rectified Linear Unit
ResNet	Residual Network
MobileNet	Mobile Network
HoG	Histogram of Gradients
MTCNN	Multi-Task cascade Convolutional Neural Network
MMOD	Maximun-Margin Object Detector
SVM	Support Vector Machine
kNN	k-Nearest Neighbor
RaFD	Radboud Faces Database
KDEF	Karolinska Directed Emotional Face
JAFFE	Japanese Female Facial Expression
CK	Cohn-Kanade
CIFE	Candid Images Facial Expression

FER	Facial Expression Recognition
SSFER	Single Shot Facial Expression Recognition
TA	Taxa de Aprendizagem
SaaS	Software as a Service
MCS	Microsoft Cognitives Services

Sumário

1 Introdução

1.1 Contexto

Há décadas a comunidade científica está interessada no reconhecimento de emoções. As diversas maneiras de expressar as emoções humanas têm sido investigadas, e as seguintes fontes de dados têm sido exploradas, tais como sinais fisiológicos, textos, envio de *emoticons*, padrão de uso em dispositivos de entrada de dados (teclado e mouse), voz e as expressões faciais. Esta última surgiu pelas anotações de Darwin (1965) e experiências de Ekman e Davidson (1994). As investigações científicas anteriormente citadas concluíram que todas as culturas expressam emoção pela face, de tal forma que há um grupo de emoções básicas (raiva, alegria, tristeza, desgosto, medo e surpresa) que possuem a mesma expressão facial independente da cultura dos indivíduos. Posteriormente, Ekman (1999) definiu o conjunto de estados negativos composto por medo, raiva, desgosto e tristeza, deliberou a alegria como fazendo parte dos estados positivos e, por fim, a surpresa como estado meio termo. Apesar de tantos anos de pesquisa, a área de reconhecimento de emoção continua entusiasmando pesquisadores principalmente por ser uma solução com aplicabilidade em vários campos, destacando-se a interação humano-computador e humano-robô em sistemas cognitivos. Embora haja várias maneiras de reconhecer emoção, a expressão facial tem recebido mais atenção evidenciada pelo maior número de publicações científicas comparada às outras maneiras de reconhecimento, e também, pela realização de concursos como foi a competição ICML'2013 [1] e anualmente como o EmotiW [2] dos anos de 2013 à 2018.

A expressão facial é uma maneira eficaz para reconhecer emoções, sobretudo por não ser uma abordagem intrusiva de coleta de dados quando comparada aos sensores fisiológicos. A popularização das câmeras fotográficas, seja em dispositivos pessoais usados em momentos esporádicos ou no segmento de vigilância, em que constantemente há um monitoramento do ambiente, facilita a captura de expressão facial caracterizando-se como um método de coleta acessível e que as pessoas sabem manusear (Cruz et al., 2017). Deve-se ressaltar que a presença desses equipamentos (câmeras fotográficas) no cotidiano da população tem sido cada vez mais invisível, portanto gerando uma redução da sensação de invasão de privacidade e alteração de comportamento por estar sendo vigiado (Cruz et al., 2017). Enquanto que, para obter medições dos sinais fisiológicos humanos, é necessário vestir ou colocar algum aparelho no corpo, neste caso, tal maneira de reconhecimento não é uma computação ubíqua gerando algum nível de incômodo ao usuário quando seus

[1] ICML'2013: https://bit.ly/1LgpbFL
[2] EmotiW: https://sites.google.com/view/emotiw2018

sinais são monitorados.

O progresso da área de aprendizado de máquina profundo ocasionou o surgimento de diversas técnicas poderosas de reconhecimento de padrões destacando-se as redes neurais de convolução. Esta técnica foi projetada especialmente para aplicações de visão computacional atuando no processamento, extração de características e classificação. Ultimamente, as redes neurais de convolução têm sido amplamente utilizadas em diversos contextos dominando os trabalhos realizados pela comunidade de visão computacional em problemas de classificação de imagens. Em reconhecimento de emoção por meio da expressão facial, as redes neurais de convolução são tão efetivas que as métricas de avaliação dos modelos estão próximas do que um humano reconheceria (Kim et al., 2016).

1.2 Motivação

O reconhecimento de emoção tem aplicação em muitas áreas. Destacamos alguns campos promissores. Na educação, por exemplo, segundo Jaques e Nunes (2013), estudantes durante o seu processo de aprendizagem transmitem constantemente diversas emoções. Portanto, 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. O intuito seria de fornecer *feedback* personalizado para o estudante através da recomendação de 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 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 tem sido o método de estímulo de emoções ao usuário e, caso não seja satisfatório, o método pode ser alterado. Para Li et al. (2015), 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, além das aplicações já mencionadas, é possível aplicar o reconhecimento de emoção na interação humano-computador (Barsoum et al., 2016; Chen et al., 2017; Liu et al., 2016; Wen et al., 2017), e interação humano-robô (Jung et al., 2015; Shin et al., 2016), criando a expectativa de que computadores do futuro possam reconhecer a emoção do usuário e adaptar-se para incentivar emoções positivas.

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 σ

= {neutralidade, raiva, alegria, tristeza, desgosto, medo e surpresa}, que são as emoções básicas definidas por (Ekman e Davidson, 1994) mais neutralidade, tal que haja uma função ϕ que saiba mapear ω por meio de α para σ.

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 com maior alcance de generalização e qual é o comportamento desta abordagem em um cenário real de uso?*

1.4 Objetivos

1.4.1 Objetivo Geral

Propor uma abordagem para reconhecer emoção humana por expressão facial para classificar emoções básicas em múltiplas faces de uma imagem e mensurar a eficácia em cenário real de uso.

1.4.2 Objetivos Específicos

- Propor um detector de face com eficácia para ambientes internos e externos;
- Validar arquiteturas de redes neurais de convolução visando a eficiência em ambientes internos e externos;
- Comparar classificadores para compor o nível de decisão observando a acurácia;
- Avaliar a abordagem proposta em um estudo de caso real mensurando a eficácia.

1.5 Hipótese

As emoções básicas e a neutralidade são transmitidas por expressões faciais que podem ser capturadas por câmeras fotográficas representadas em imagens. As redes neurais de convolução destacam-se como a melhor técnica para processamento de imagem. Porém, a rede neural de convolução é uma técnica que possui muitas configurações. Diante desse contexto, este trabalho foca em encontrar o ajuste ideal da rede neural de convolução para maximizar os acertos ao classificar as imagens. Este ajuste consiste em introduzir imagens de expressões faciais capturadas em ambientes internos e externos. Além de aplicar recursos de treinamento que destaca o melhor modelo gerado durante o ajuste de pesos.

1.6 Abordagem Proposta

A abordagem proposta consiste nos seguintes componentes: coleta de dados, detecção de face, extração de características e classificação. A coleta de dados monitora o indivíduo capturando imagens continuamente. Desta forma, gera-se um *buffer* de imagens a serem processadas. O detector de face, que é uma rede neural de convolução, recebe uma imagem e computa a quantidade e a localização (coordenadas) das faces existentes. Depois, é executado um processo de recorte da face caracterizando-se como uma etapa de pré-processamento. A face recortada é enviada para um verificador de posicionamento da face. Esta etapa analisa se o posicionamento da face está totalmente dentro da imagem, se positivo, a face é enviada para rede neural de convolução extrair as características, caso não esteja, o processo é encerrado. A extração de características resulta em um vetor unidimensional com tamanho bastante inferior da imagem original com redução de até 98.75% do tamanho total da imagem para o classificador analisar. Após averiguação do classificador, tem-se a decisão que é mapeada para um vetor de probabilidades distribuída para cada emoção, onde a probabilidade mais alta é a emoção detectada nesta imagem. Portanto, o método proposto retorna as coordenadas das faces localizadas com a emoção correspondente para a aplicação consumidora. Desta forma, é possível saber de qual face se trata e qual foi a emoção detectada. Este processo é ilustrado na Figura 1. Para concluir, é importante salientar que em uma imagem com múltiplas faces o processo consiste em classificar uma por vez, pois é mais simples reconhecer a emoção de uma única face do que todas ao mesmo tempo.

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 5 discute os resultados de um estudo de caso, enquanto o Capítulo 6 enfatiza as considerações finais, limitações do trabalhos e os trabalhos futuros.

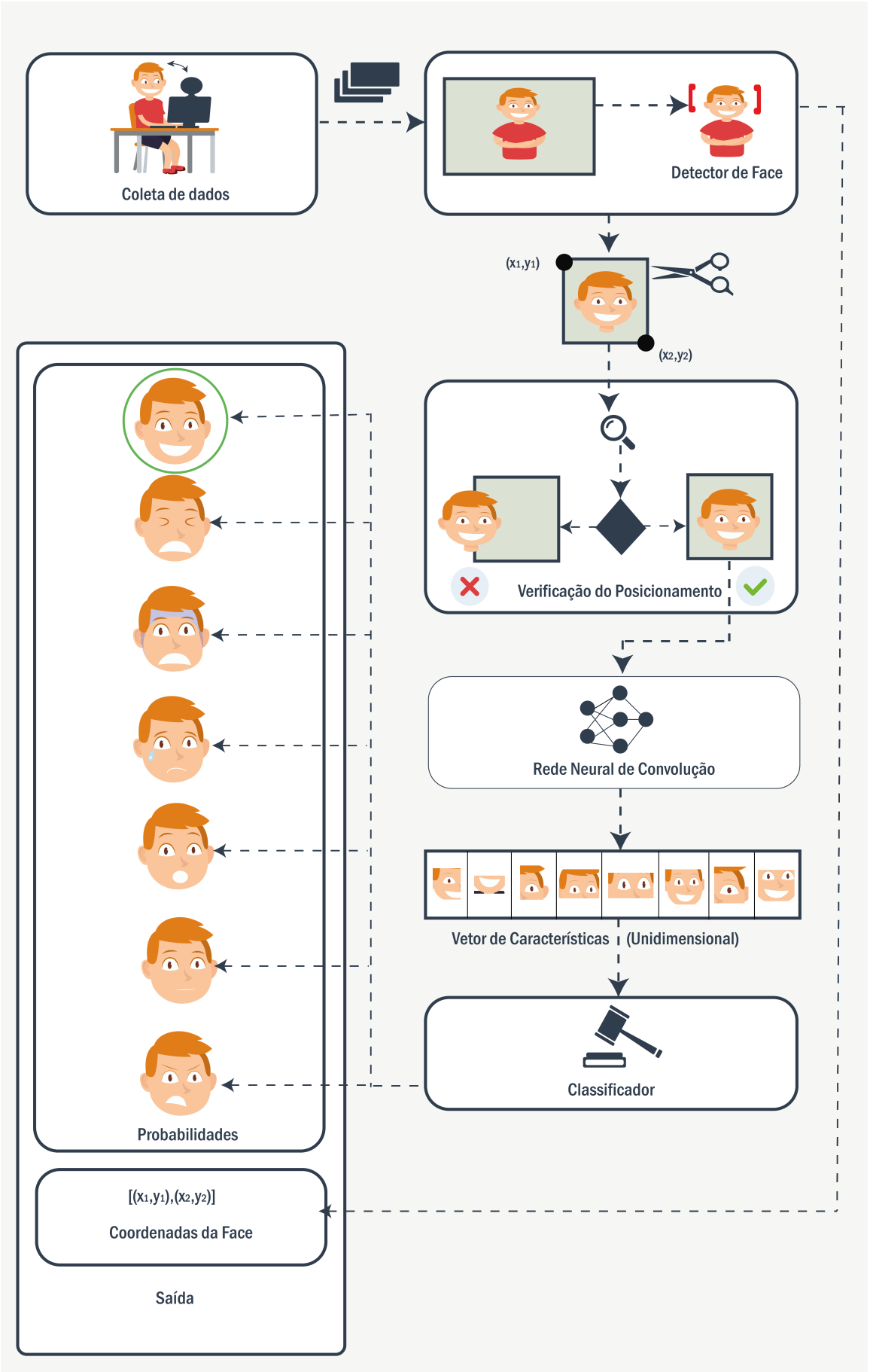

Figura 1 – Abordagem Proposta

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 abordagens para identificação da emoção. A Seção 2.2 fundamenta a expressão facial emocional e mostra as expressões faciais que podem ser relacionadas com o conjunto das emoções básicas. 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 os métodos para detecção de face. 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 neste trabalho. A Seção 2.10 retrata as principais métricas para avaliar classificadores e para concluir a Seção 2.11 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 sendo que as emoções que tem sido frequentemente investigadas são as básicas como a raiva, alegria, tristeza, desgosto, medo e surpresa (Ekman e Davidson, 1994). Com objetivo de reconhecer o conjunto das emoções básicas, 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, em uma imagem que retrata uma expressão facial, a técnica deve diferenciar a alegria de neutralidade e a característica de um sorriso é um padrão preponderante para esta atividade. Além disso, para reconhecer estados não básicos ou secundários a comunidade tem gerado várias heurísticas, entretanto, devido a complexidade tais heurísticas têm como desvantagem o emprego somente em ambientes com as variáveis controladas e a necessidade de uma dupla verificação com avaliação de um observador humano (D'Mello et al., 2018).

É possível reconhecer as emoções de 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) em relação aos outros métodos de coleta de dados, é o método que causa menor sensação de intrusão ao usuário, quando comparado com outros métodos de coleta de dados, pois fotografias da expressão facial podem ser capturadas dentro do cotidiano das pessoas, não necessita o uso de instrumentos especiais, podendo ser imperceptível aos usuários quando coletadas por meio de câmeras de vigilâncias.

2.2 Expressão Facial Emocional

Darwin (1965) verificou que fenômenos emocionais idênticos expressado em faces humanas podiam ser encontrados em diferentes culturas. Posteriormente, o trabalho de Ekman e 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. O conjunto das emoções básicas composta por: raiva, alegria, tristeza, desgosto, medo e surpresa, pertence ao grupo das expressões faciais universais. Portanto, as emoções básicas podem ser reconhecidas através da identificação das expressões faciais universais. Adicionalmente, uma outra expressão facial pode ser reconhecida e relacionada a um estado emocional que é a própria neutralidade. Particularmente, cada emoção básica transmitida por um indivíduo possui uma movimentação muscular facial por meio da sobrancelha, dos olhos, das bochechas, dentre outros, sendo que cada movimento é definido como uma unidade de ação (UA) (Ekman e Friesen, 1977). Sendo assim, é possível definir quais são as UAs que caracterizam uma determinada expressão facial emocional (Ekman e Friesen, 1977). Consequentemente é através das UAs que é possível caracterizar uma expressão facial para ser mapeada para uma emoção básica, desta forma, permitindo a identificação e a diferenciação de cada tipo de emoção.

2.3 Aprendizagem de Máquina

Uma maneira efetiva para identificar as UAs é por meio da aprendizagem de máquina (AM). Com intuito de criar sistemas que possuem a capacidade de reconhecer emoção de forma automática é muito comum o emprego de técnicas de AM, que é um ramo da inteligência artificial (IA), em que máquinas aprendem a partir de uma experiência e são habilitadas para reconhecer padrões. Uma definição de AM 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

Figura 2 – Expressão facial emocional

Fonte: (Shojaeilangari et al., 2014)

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, há três abordagens básicas de aprendizagem: supervisionada, não supervisionada e por reforço. Quando as instâncias são conhecidas, isto é, cada instância possui o seu rótulo informando a sua descrição, então o aprendizado é supervisionado. Em contrapartida, quando as instâncias não tem rótulo, há apenas o conjunto de dados mas não se sabe exatamente a qual classe as amostras pertencem, então a aprendizagem é não supervisionada. Enquanto que a aprendizagem por reforço está fora do escopo desta dissertação, pois este aprendizado é usualmente aplicado nas áreas de jogos e robótica. A ideia básica do aprendizado por reforço é que não há nenhuma base dados ou poucos dados e o aprendizado é baseado na interação de um agente com o ambiente, em que quando o agente realiza uma ação correta ganha-se uma recompensa e enquanto que o oposto, quando a escolha é errada, uma punição. A repetição deste ciclo de realizar ações e receber recompensas ou punições gera aprendizado. Neste trabalho, o aprendizado empregado é o supervisionado. Portanto, o problema de reconhecimento de emoção consiste na identificação de sete classes (as emoções básicas mais a neutralidade), parte-se da premissa que o conjunto de dados, mas especificamente para cada imagem, usados para geração dos modelos possui uma descrição (rótulo) informando qual é a classe da amostra (Géron, 2017; Kotsiantis et al., 2007).

O conjunto de métodos supervisionado possui duas fases básicas: uma de treinamento e outra de teste. A primeira consiste na utilização de um conjunto de dados, também conhecido como base de treino, com objetivo de encontrar padrões nesta amostragem e, assim, produzir um modelo para armazenar o aprendizado. Por fim, na fase de teste, o algoritmo deve classificar (prever) dados desconhecidos utilizando a base de teste, a partir dos padrões encontrados (aprendizado) na fase anterior e mensurar o desempenho obtido comparando os rótulos verdadeiros com os rótulos previstos. 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

Na área de Visão Computacional (VC) as abordagens empregadas para a classificação de imagem geralmente segue um processo comum. Quando o processo de classificação é por meio da aprendizagem de máquina há as seguintes etapas: (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 características 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 é a classe da imagem.

2.5 Detecção Facial

A detecção facial é um pré-processamento amplamente utilizado em problemas de visão computacional relacionada à face como objeto alvo. A localização da face consiste no emprego 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 o método deve encontrar o conjunto de pontos que representa as posições das faces em uma imagem. No entanto, uma imagem pode ter diferentes objetos, *backgrounds* e ruídos. A união desses elementos complicadores pode induzir o método a identificar erroneamente uma face, causando confusão e a geração de falsos positivos, isto é, objetos sendo incorretamente identificados como face.

Após a detecção de face, geralmente é realizado o recorte utilizando o conjunto de coordenadas definidas pela etapa anterior. O recorte da face é uma operação com valor significativo, ocasionando a exclusão do *background* da imagem. Desta forma, somente a face recortada é enviada para as análises posteriores, reduzindo a complexidade do

problema, pois não há necessidade, das etapas seguintes ao pré-processamento, a separar o *background* da face. Posteriormente ao recorte da face, a imagem original, que deve estar com uma face recortada, é mantida para nova averiguação de recorte de face. Caso exista outras faces detectadas 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 existe uma face (e.g. imagem de um avião) o processo de reconhecimento é automaticamente encerrado, pois se não há uma face para detectar, logo não há uma expressão facial emocional para reconhecer. Portanto, para as análises posteriores parte-se da premissa que a imagem a ser analisada é de uma face de um único indivíduo.

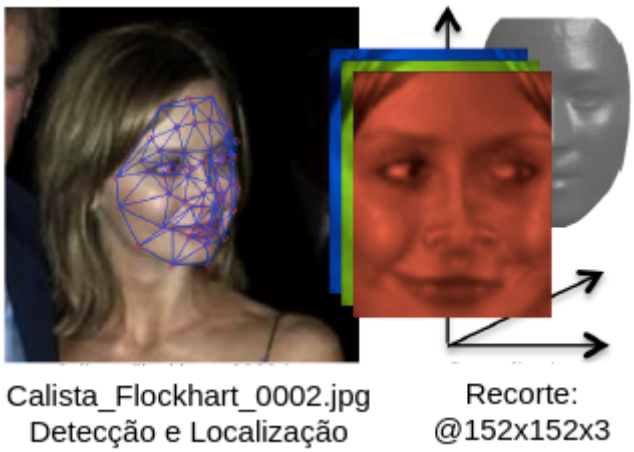

Figura 3 – Detecção Facial

Fonte: (Taigman et al., 2014)

O algoritmo Viola-Jones (Viola e Jones, 2001) é amplamente usado pela comunidade para detecção facial. Este método é composto por vários classificadores "fracos", sendo que cada classificador é responsável em analisar uma sub-região da imagem para encontrar uma característica específica. Desta forma, após análise é feito uma junção dos resultados e caso haja uma alta taxa de detecção das características é determinado como positivo a presença do objeto (face) na imagem. Este algoritmo é utilizado por Chen et al. (2017), Shan et al. (2017), Shin et al. (2016), Vo e Le (2016), Mayya et al. (2016), Ng et al. (2015) e Li et al. (2015).

Outro método bastante popular é o *Histogram of Gradients SVM* (HoG-SVM) (Dalal e Triggs, 2005). Este método extrai as características baseada nos gradientes da imagem. Os gradientes extraídos fornece uma intuição da forma dos objetos presentes na imagem. Após a captura dos gradientes é enviado para um SVM classificar se há a presença do objeto na imagem.

Embora o Viola-Jones e o HoG-SVM seja amplamente utilizado e aceito pela comunidade, recentemente outros métodos vem surgindo e ganhando popularidade, principal-

mente os baseados em redes profundas como o *Multi-Task Convolutional Neural Network* (MTCNN) (Zhang et al., 2016) e o *Maximun-Margin Object Detector Convolutional Neural Network* (MMOD-CNN) (King, 2015). Os métodos baseados em redes profundas são o estado-da-arte em acurácia nos *benchmarks* tanto para detecção quanto alinhamento de face. Além disso, os métodos baseados em redes profundas, quando executado em GPUs, tem alto desempenho no quesito de velocidade, consolidando-se como um método adequado para processamento de faces em tempo real e de imagens com alta resolução.

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 foram criadas por meio de inspiração biológica, especificamente, através do funcionamento do cérebro dos mamíferos. O foco das redes neurais é construir máquinas inteligentes (Géron, 2017; Goodfellow et al., 2016).

Um *perceptron* é o mais simples componente de uma rede neural também designado de neurônio. Este componente é baseado em um neurônio biológico e, por isso, possui conexões de entrada e saída para conectar aos outros neurônios enviando e recebendo sinais. Cada conexão de entrada é associada a um peso, e também, possui a abertura para receber um sinal de entrada de um outro neurônio. Tanto o peso como o sinal de entrada são parâmetros para realização de uma computação através de uma função de ativação. Esta computação intermediada pela função de ativação gera uma saída, que é justamente o sinal propagado para a entrada de outros neurônios de uma camada posterior (Géron, 2017).

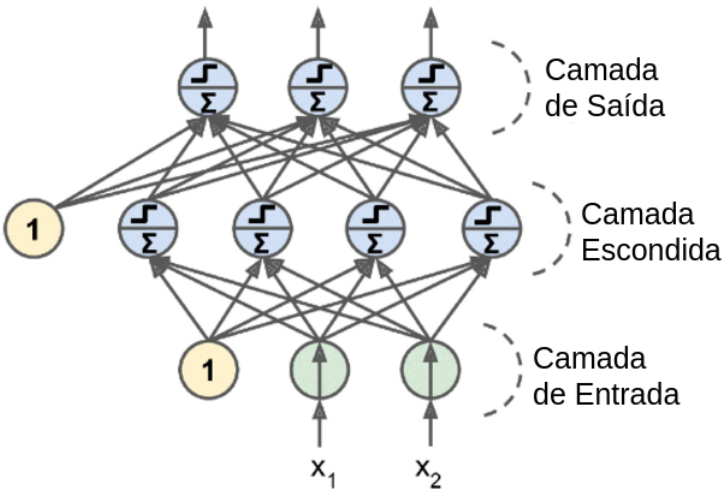

Figura 4 – Rede Neural Artificial

Fonte: (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 escondidas (ou 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 escondida 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. Isto significa que se o problema for reconhecer as emoções básicas mais a neutralidade, a camada de saída terá sete neurônios, pois o conjunto das emoções básicas mais neutralidade totaliza sete classes. Além disso, cada neurônio da camada de saída está associada a uma classe e quando uma instância desconhecida for processada para classificação, o neurônio da camada de saída que terminar ativado representa a classificação desta instância (Géron, 2017; Goodfellow et al., 2016).

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 (algorítmicas) de treinamento de uma rede neural. Este cenário favorável permitiu que as RNCs aproximassem ao nível humano na questão de reconhecimento 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, dentre outras tarefas.

Os trabalhos de Hubel (1959) e Hubel e Wiesel (1959) realizaram uma série de experimentos em gatos em 1958, e posteriormente, em macacos (Hubel e Wiesel, 1968). Estes experimentos tinham o objetivo de encontrar intuições do funcionamento do córtex visual dos mamíferos, 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 têm um campo de recepção local, ou seja, uma sub-área do campo visual em que os neurônios reagem somente a um estímulo ocorrido nessa sub região. O campo de recepção local dos diferentes neurônios pode sobrepor um ao outro e a sua combinação gera o campo visual. Os autores mostraram que alguns neurônios reagem somente às imagens com padrões de linhas

horizontais enquanto outros reagem às linhas com diferentes orientações. Descobriu-se que quanto maior o campo de recepção local maior a capacidade desse neurônio a reagir aos padrões mais complexos. Através dos experimentos, notou-se também que, de fato alguns neurônios têm um campo grande de recepção local, consequentemente, reagindo aos padrões mais complexos. Entretanto, o córtex visual busca combinar os neurônios com intuito de gerar campos visuais menores, e assim, cada neurônio processar padrões menos complexos. Estas observações são evidências de que os neurônios são baseados na saída do vizinho e através dessas conexões vão se adaptando ao longo do tempo trabalhando de forma distribuída combinando as suas saídas para processar todo o campo visual.

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 criar uma nova rede neural denominada rede neural de convolução (RNC) que é focada no processamento de imagens. A primeira RNC foi apresentada por LeCun et al. (1998), uma arquitetura denominada LeNet-5 que foi utilizada para reconhecer dígitos escritos no papel. Este problema é clássico na área de visão computacional que também é conhecido como reconhecimento ótico de caracteres (do inglês: *Optical Character Recognition* - OCR).

2.7.1 Camada de Convolução

A camada de convolução é o mais importante bloco de uma RNC. A convolução é uma operação matemática que consiste no deslizamento de uma função sobre a outra calculando a integral da soma do produto de ambas a partir da sobreposição gerada pelo deslizamento. A Figura 5 ilustra o que cada camada de convolução recebe em seu campo de recepção local (campo visual) a partir de uma imagem de entrada. Vale ressaltar que cada neurônio da camada de convolução 1 está conectado somente a alguns *pixels* da imagem, portanto tal neurônio processa somente uma sub-região da imagem. Diferentemente da abordagem tradicional em que um neurônio de uma camada escondida estaria conectado com todos os dados (*pixels* da imagem) de entrada. Os neurônios da camada de convolução 2 estão conectados somente a um pequeno campo visual da camada de convolução 1, novamente diferente da rede neural *perceptron* (abordagem tradicional) em que todos os neurônios são conectados com todos da camada anterior e posterior (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);

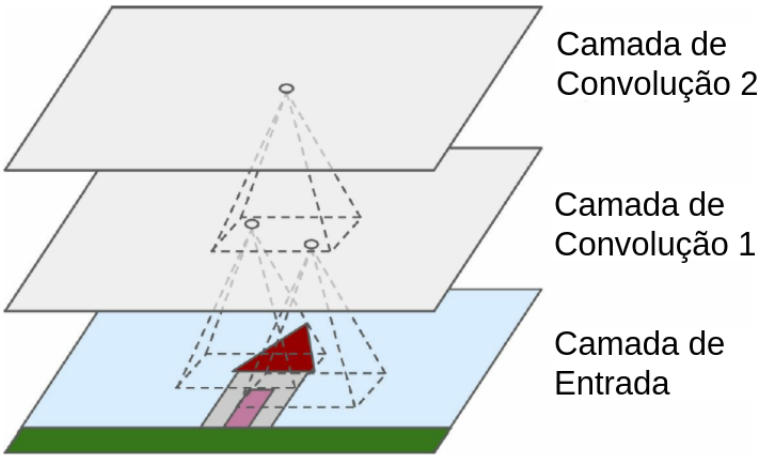

Figura 5 – Camada de Convolução com campos locais de recepção

Fonte: (Géron, 2017)

(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 (pesos) 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*

A camada de *pooling* é uma operação essencial na CNN. Esta camada tem como foco principal realizar subamostragem, isto é, diminuir o tamanho da imagem durante o processamento da RNC com objetivo de reduzir a carga computacional e o número de parâmetros, ocasionando a diminuição do risco de *overfitting* e do consumo de memória (Géron, 2017).

A operação de *pooling* faz alusão ao funcionamento do córtex visual em que cria os campos de recepção local consistindo de uma sub região do campo visual pertencente à um determinado neurônio. Além disso, a vantagem em reduzir o tamanho da imagem faz a rede neural mais tolerável à variação do objeto. Desta forma, o modelo aprende que a posição dos objetos, mas especificamente as informações dos *pixels* que constituem o objeto, sofrem variações de suas posições, diferentemente da visão computacional tradicional que as heurísticas somente funcionam quando os objetos estão em posições fixas. O fato de reduzir o tamanho da imagem também elimina os menores ruídos.

Geralmente uma camada de *pooling* é implementada logo após uma camada de convolução. Portanto a *pooling* recebe como entrada os *pixels* de uma imagem proces-

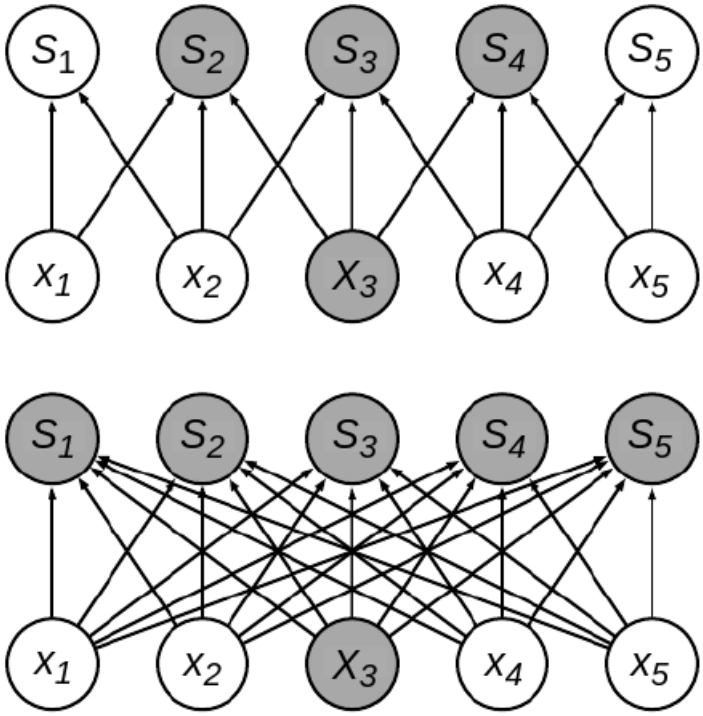

Figura 6 – Conectividade esparsa. É destacada a entrada x_3 e a saída em S que são afetadas por x_3. (Cima) Quando S recebe a convolução com um *kernel* de tamanho 3, somente três saídas são afetadas por x_3. (Baixo) Quando S é gerado por rede neural tradicional, todos são afetados por x_3.

Fonte: (Goodfellow et al., 2016)

sados por uma camada de convolução, a fim de aplicar operações de subamostragem e encaminhar a saída para uma próxima camada de convolução (Goodfellow et al., 2016). Em uma arquitetura de RNC, há várias camadas de convolução e *pooling*, portanto a junção dessas camadas é a base da RNC.

Há 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 nas RNCs.

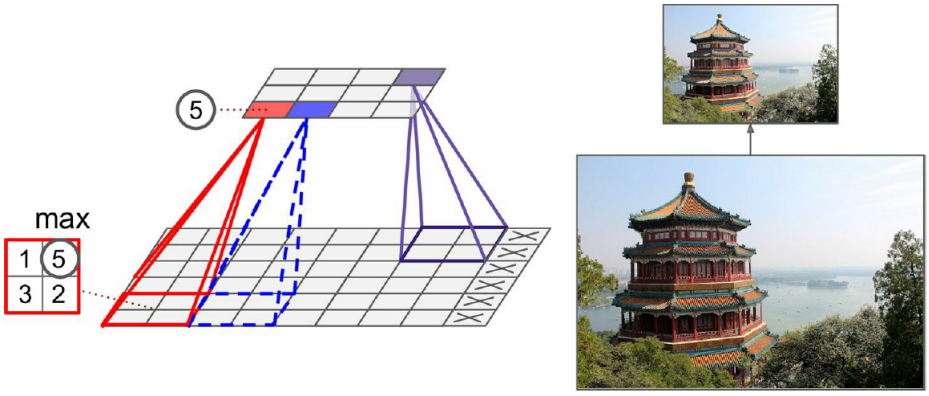

Figura 7 – Camada de *Max Pooling*

Fonte: (Géron, 2017)

2.7.3 *Softmax*

As arquiteturas de RNCs são compostas em sua maior parte por camadas de convolução e *pooling*. Estas camadas processam a imagem com intuito de extrair os principais padrões para servir de entrada à um classificador a fim de determinar a classe. Por isso, ao final de uma arquitetura de RNC é necessário um classificador, que usualmente nas principais arquiteturas tem sido o *Softmax* (ver Seção 2.8). 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. O *Softmax* é um modelo generalizado da regressão logística, no entanto com um diferencial tem a capacidade de estimar probabilidades para múltiplas classes. Um *Softmax* está habilitado a estimar probabilidades recebendo um vetor de características extraídos de uma imagem e, para todos os efeitos, a imagem é uma expressão facial que pertence ao conjunto das emoções básicas mais a neutralidade.

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, ao invés da tradicional conexão entre uma camada de convolução e a camada de *pooling*. Esta arquitetura pode ser consultada na Tabela 1.

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	-	-

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

Esta arquitetura causou impacto no concurso do ano de 2012 do ImageNet ILSVRC. Este concurso é o principal na área de classificação de imagens em que são testados os mais variados objetos. A AlexNet venceu por uma margem considerável: alcançou 17% no top-5 da taxa de erro, enquanto o segundo melhor alcançou somente 26%!

2.8.2 GoogLeNet

A arquitetura GoogLeNet (Szegedy et al., 2015) foi desenvolvida por Christian Szegedy do *Google Research*. Também foi vencedora do concurso do ILSVRC justamente no ano de 2014. A GoogLeNet alcançou a incrível taxa de erro no top-5 abaixo de 7%. Este grande desempenho foi devido, em grande parte, pelo fato da rede ser mais profunda que as anteriores. O aumento de profundidade está diretamente relacionado à criação de sub redes chamadas de *inception*. Um módulo simples de *inception* está ilustrado na Figura 8. Essas sub redes permitiram que a GoogLeNet utilizasse os parâmetros de forma mais eficiente quando comparada às arquiteturas anteriores. 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 para serem otimizados durante o treinamento do que a segunda (Géron, 2017). A GoogLeNet é a base das variações de arquiteturas de CNNs denominadas de Inception.

Um módulo *inception*, que está ilustrado na Figura 8, possui a seguinte notação:

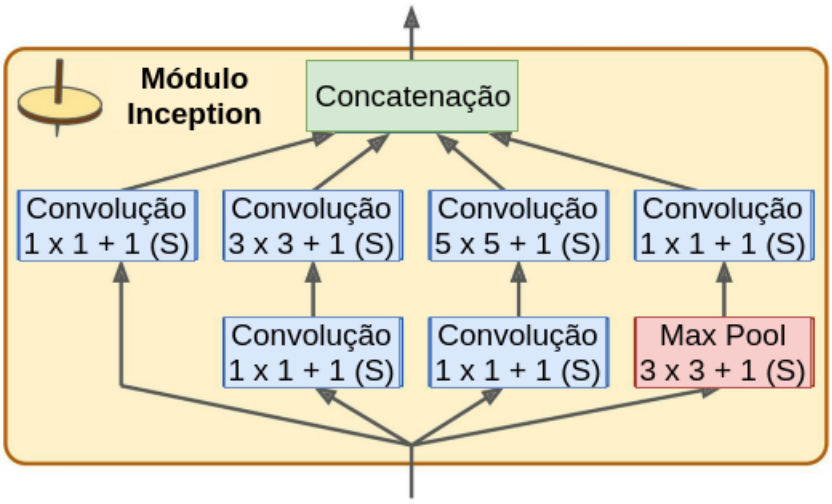

Figura 8 – Módulo *inception*

Fonte: (Géron, 2017)

"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, em que o segundo está empilhado no primeiro. Vale ressaltar que todas as camadas de convolução utilizam a Rectified Linear Unit (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*.

2.8.3 VGGNet

Proposta por Simonyan e 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 resultou em que não haja um aumento significativo de

parâmetros a serem ajustados a medida que a rede cresce, desta forma, gerando eficiência.

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					

2.8.4 Residual Network

A Residual Network (ou ResNet) foi desenvolvida por He et al. (2016). Esta técnica foi a vencedora do concurso ImageNet ILSVRC 2015. A ResNet avaliou a base de teste do ImageNet que tinha 1000 classes e alcançou incríveis 3.6% em taxa de erro no top-5. A avaliação top-5 consiste na verificação das 5 classes com maiores probabilidades após uma classificação, e caso esta imagem de fato pertença ao grupo destas 5 classes é considerado uma classificação correta. A ResNet proposta para o concurso tinha 152 camadas caracterizando uma rede muito profunda.

O segredo desta técnica consiste na inovação dos blocos residuais. A novidade desta arquitetura é possibilitar que um bloco residual faça o mapeamento de outros blocos distantes através das conexões de saltos, isto é, os sinais propagados pelos neurônios

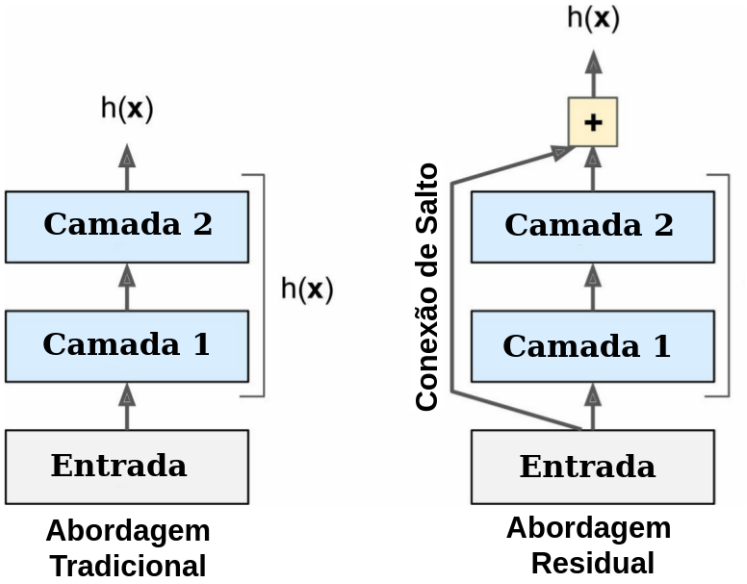

Figura 9 – Bloco residual

Fonte: (Géron, 2017)

tanto na direção para frente como para trás podem ser enviados diretamente de uma camada para qualquer outra independente da posição como ilustrado na Figura 9. Isto é diferente da abordagem tradicional que somente permite propagação de sinal para a camada anterior e posterior. A conexão de salto trouxe também impactos positivos na aprendizagem, principalmente atuando na função objetiva, pois quando a mesma está próxima de convergir ocorre um aumento considerável da velocidade de treino comparada às redes tradicionais que não possuem as conexões de salto.

2.8.5 Mobile Network

A Mobile Network (MobileNet) foi desenvolvida pela equipe do Google (Howard et al., 2017). Esta rede tem como característica principal ser apropriada para aplicações de visão computacional em *smartphones* e embarcadas. Por isso, tem como pilar a eficiência para consumir menos recursos computacionais. A MobileNet para obter eficiência consiste de um conjunto de hiper parâmetros com valores baixos nas camadas da rede. Obviamente isso faz com que a MobileNet esteja atrás na métrica de acurácia em vários problemas comparada à arquiteturas mais robustas. A MobileNet utilizada por este trabalho é a versão 2 (Sandler et al., 2018). Esta MobileNet configura-se como mais poderosa do que a versão inicial, pois inclui em sua arquitetura camadas residuais ilustrada pela Figura 9.

2.9 Classificadores

2.9.1 *Ensemble*

Suponha uma questão complexa perguntada aleatoriamente para milhões de pessoas. Posteriormente todas as respostas coletadas são agregadas para obtenção do resultado final. Em muitos casos, a resposta agregada é melhor do que a de um especialista, isto é conhecido como sabedoria popular. Similarmente, a agregação das classificações de uma determinada instância a partir de um grupo de RNCs frequentemente tem mais acertos do que a classificação de uma única RNC. Um grupo de RNCs ou de outros classificadores quando combinados para decidir a classificação de uma instância são chamados de *ensemble* (Géron, 2017). Geralmente, quem trabalha com *ensemble* no problema de reconhecimento de emoção tem utilizado da média das probabilidades estimadas pelo *Softmax* das RNCs para decidir qual a classificação final da expressão facial.

2.9.2 K-Nearest Neighbors

O K-Nearest Neighbors (kNN) é um simples algoritmo de classificação baseado em cálculos de distâncias (Witten et al., 2016). Este método pertence à família apoiada nas instâncias ou preguiçosa. Desta forma, não é gerado um modelo e o seu aprendizado consiste em armazenar a base de dados em uma estrutura de hierarquia. Um parâmetro fundamental do kNN é o valor de K. Suponha que temos k igual a 5, então, aplicando uma instância para o kNN classificar, o método irar procurar quais são as 5 instâncias da base armazenada que mais se aproximam da entrada baseando-se na distância. Após localizar as 5 instâncias que mais se aproximam, é determinado para a instância de entrada a classe majoritária deste conjunto.

2.9.3 Support Vector Machine

O Support Vector Machine (SVM) é uma técnica poderosa e versátil de aprendizado de máquina (Géron, 2017). É um dos modelos mais populares de aprendizado de máquina, especialmente, para aplicações de visão computacional. O SVM é um problema de otimização em que o foco é dispor as instância em um espaço, de modo que seja possível traçar um plano ótimo de separação entre as classes (Boser et al., 1992). Entretanto, esse plano de separação tem algumas propriedades como ter a maior margem possível entre as classes como mostra a Figura 10. Nesta figura, é ilustrado um problema de classificação binária em que há a classe íris-versicolor e íris-setosa. O objetivo é encontrar o nível de decisão ótimo que separa as duas classes. O ponto de partida do SVM é a figura do lado esquerdo em que é traçado uma reta que ainda não é a ótima. Após algumas iterações, os vetores de suportes são identificados, isto é, as instâncias dos extremos de cada classe,

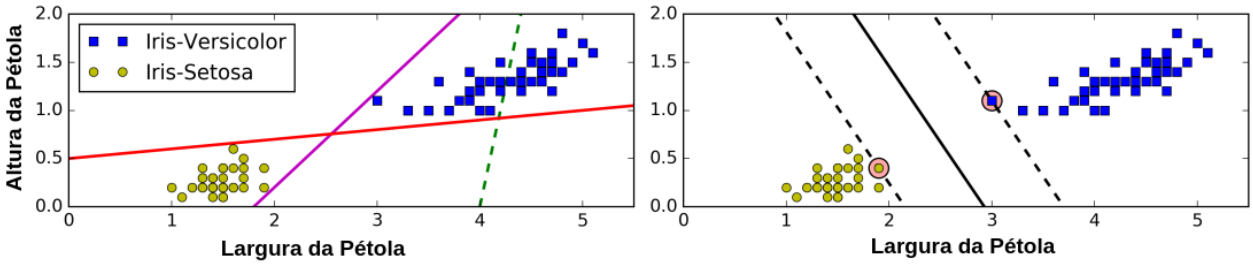

Figura 10 – SVM - hiperplanos de separação

Fonte: (Géron, 2017)

sendo possível calcular a margem máxima de separação representada pela reta na figura da direita.

2.9.4 Random Forest

O Random Forest é uma espécie de *ensemble* de árvores de decisões. Por isso, primeiramente deve-se compreender o que é uma árvore de decisão. Uma árvore de decisão é formada por regras ou declarações de condições que tem como propriedade estrutural a raiz que está no topo da árvore, nós intermediários e as folhas nas extremidades. A classificação vai da raiz até uma folha, em que a folha representa a classe identificada. A concepção das regras é baseada em um processo estatísticos geralmente conduzida por uma métrica chamada entropia. A Figura 11 ilustra um exemplo de árvore de decisão para classificar de acordo com os atributos de tempo, umidade e vento se deve jogar uma partida de tênis. O Random Forest é um *ensemble* de um conjunto de árvores de decisão em que cada árvore é diferente uma da outra com diferentes profundidades. E a classificação final é composta pela junção dos resultados individuais de cada árvore (Breiman, 2001).

2.10 Métricas de Avaliação de Desempenho para Classificadores

Um dos objetivos da avaliação de desempenho para classificadores é a identificação de *underfitting* e *overfitting*. *Underfitting* ocorre quando o classificador não treinou o suficiente, desta forma, o aprendizado ainda é baixo e necessita de aprimoramento, pois não aprendeu o bastante e não funcionará na prática. Já *overfitting* é quando treinou muito, caracterizando uma espécie de vício ou uma tendência a escolher as classes majoritárias ignorando as minoritárias. As principais métricas utilizadas para avaliação de desempenho de classificadores são: acurácia, precisão, revocação e f1-score. Por meio dessas métricas é possível comparar os classificadores determinando os pontos fortes e fracos, investigar quais são as classes que o modelo tem maiores desempenho, verificar a existência de *over-*

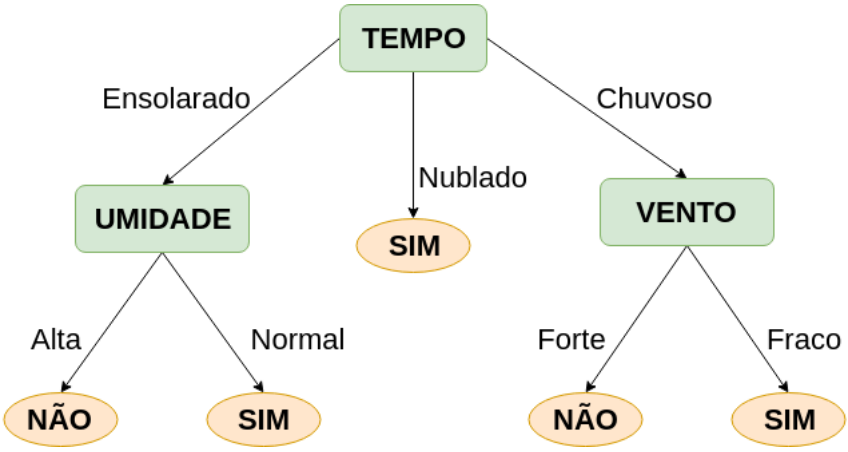

Figura 11 – Árvore de Decisão - Jogar uma partida de tênis

Fonte: (Witten et al., 2016)

fitting em uma classe específica, e também, se houve aprendizado em classes com menores amostras. É relevante tais verificações, pois quando a base de treino é desbalanceada, isto é, quando há classes com muita amostragem e outras com baixa amostragem, pode gerar um modelo que consiste de chutar as classes com maiores amostras, pois se está acertando as classes majoritárias a taxa de acerto está sendo maximizada, no entanto, as classes com menores amostras podem estar com baixa acurácia, isto claramente caracteriza um modelo com *overfitting*.

A acurácia é a proporção dos casos corretamente classificados e é calculada a partir da Equação 2.1, em que TP é a taxa de verdadeiros positivos, TN é a taxa de verdadeiros negativos, FP é a taxa de falsos positivos e FN é a taxa de falsos negativos. A precisão é a proporção das instâncias que foram classificadas como positiva sobre todas as instâncias que de fato são positivas, sendo assim, é a informação do quão bem o modelo trabalhou e a fórmula pode ser consultada em 2.2. A revocação (ver Equação 2.3) nos diz de todas as instâncias que foram classificadas como positiva, o quanto efetivamente é positiva ou o quão frequente o modelo classifica como positiva.

Classificadores com uma alta taxa de revocação têm uma taxa baixa de instâncias positivas classificadas incorretamente. Vale destacar que é fácil construir um classificador que alcança altas taxas de precisão ou revocação, difícil é construir com ambas taxas em nível alto. Imagine um conjunto binário de classe positiva e negativa, suponha que o classificador prevê todas as instâncias são da classe positiva, o modelo teria uma taxa perfeita de revocação e precisão para a classe positiva, entretanto uma baixa precisão no geral. Por isso, criar um classificador que maximiza tanto precisão como revocação é um desafio.

É frequentemente conveniente combinar precisão e revocação em uma única métrica chamada de f1-score, principalmente quando é necessário comparar de forma sim-

plória classificadores distintos. A f1-score é a média harmônica de precisão e revocação (Equação 2.4). Enquanto a média tradicional trata todos os valores igualmente, isto é, sem pesos, a média harmônica favorece os valores mais baixos. Portanto, caso tenha alta revocação e baixa precisão, o resultado de f1-score será mais próximo de precisão, pois esta métrica adiciona pesos nos valores mais baixos, que neste caso é o valor de precisão.

$$acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$precisão = \frac{TP}{TP + FP} \tag{2.2}$$

$$revocação = \frac{TP}{TP + FN} \tag{2.3}$$

$$f1 - score = \frac{2}{\frac{1}{precisão} + \frac{1}{revocação}} = 2 * \frac{precisão * revocação}{precisão + revocação} \tag{2.4}$$

2.11 Resumo

Neste capítulo foram apresentados os principais conceitos utilizados por este trabalho. 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 causar menor sensação de intrusão e pela popularidade de dispositivos que contém câmeras fotográficas caracterizando-se como uma tecnologia ubíqua. Entretanto, somente as emoções básicas mais a neutralidade 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 detecção e 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 e os classificadores em que cada um possui sua característica em particular que precisam ser experimentadas em diferentes cenários e avaliadas por meio das métricas para definir a melhor combinação entre arquitetura e classificador.

3 Trabalhos Correlatos

Os trabalhos relacionados são discutidos neste capítulo e está organizado da seguinte forma. A Seção 3.1 discute as abordagens para preparação dos dados. A Seção 3.2 analisa as principais extrações de características encontradas na literatura para o processamento da expressão facial. A Seção 3.3 avalia os trabalhos correlatos categorizados por arquiteturas de RNC. A Seção 3.4 discute as áreas de aplicações para o reconhecimento de emoção por expressão facial e finalmente a Seção 3.5 faz um resumo a respeito deste capítulo.

3.1 Preparação dos dados

Em aprendizagem de máquina, há um problema clássico presente em qualquer forma de classificação, seja de imagem, vídeo, áudio ou outro, que é a ausência de dados. Os algoritmos de aprendizagem de máquina requerem quantidade de dados expressivas para apresentar soluções com desempenho satisfatório, especificamente as redes neurais profundas. Raramente há dados disponíveis suficientes para treinar e validar uma rede neural de convolução. Vale ressaltar, entretanto, que cada problema tem sua particularidade, isto é, quanto maior a complexidade mais dados são necessários.

Para amenizar esse problema, a comunidade de reconhecimento de emoção utiliza a técnica de aumento de dados gerando uma multiplicação de imagens. Essa técnica consiste na geração de cópias de uma imagem original contendo uma expressão facial para, a partir dessa imagem, gerar novas imagens. Entretanto, estas novas imagens são diferentes da imagem original, justamente por possuir alterações na posição da face como leves rotações, distorções, variação do brilho, variação 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, iluminação e escala.

Os trabalhos de Barsoum et al. (2016); Huang e Lu (2016); Kim et al. (2016); Shin et al. (2016); Yu et al. (2016) e Li et al. (2015) 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 de tal forma que o ajuste dos modelos alcançaram generalização adequada 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 uma 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 característica geométrica consiste na obtenção de pontos faciais ilustradas pela Figura 12. As características geométricas tem como finalidade capturar as deformações na face causadas pela ativação dos músculos a partir dos pontos faciais (Yu et al., 2016). Esses pontos faciais podem ser mapeados pelos métodos propostos por Yu et al. (2015) 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 visão computacional por meio de heurísticas, isto é, sem o uso das redes neurais de convolução. Todavia, o trabalho de Yu et al. (2016) realiza a extração geométrica concatenando

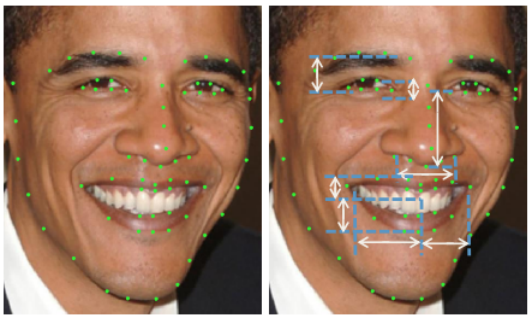

Figura 12 – Extração dos pontos faciais para características geométrica

Fonte: (Yu et al., 2016)

à extração da RNC e obteve um pequeno ganho de 1% na taxa de precisão. Esta abordagem é ilustrada na Figura 13. No entanto, a combinação entre a extração geométrica e RNC, obviamente, aumenta o custo computacional devido à necessidade da execução de outros algoritmos como o mapeamento dos pontos faciais e cálculos das suas distâncias. Caso o foco do reconhecedor de emoções seja por aplicações em cenários reais, configura-se uma situação em que o uso da extração geométrica é pouco viável devido ao aumento na taxa de reconhecimento ser baixo, portanto não compensada pelo aumento do custo computacional, pois tais cenários requerem classificação em tempo real.

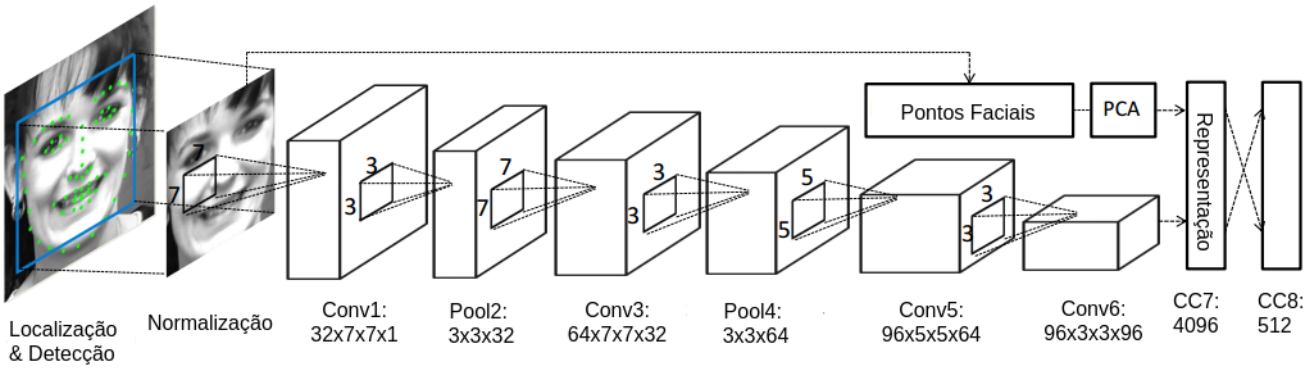

Figura 13 – Concatenação dos pontos faciais com uma rede neural de convolução

Fonte: (Yu et al., 2016)

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. Ao contrário das características geométricas que foca na captura de informações a partir dos cálculos de distância considerando os pontos faciais, desta forma, ignorando as sub-regiões próximas dos pontos e a deformação no rosto causada pela movimentação muscular

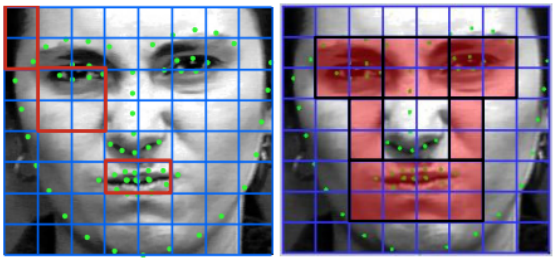

Figura 14 – Extração das sub-regiões faciais para características aparente

Fonte: (Yu et al., 2016)

facial (Yu et al., 2016). A extração aparente foi amplamente estudada por Ekman e Davidson (1994), encontrando 96 unidades de ações (ou sub-regiões) na face correspondentes a movimentação de diversos músculos associada a uma determinada emoção. A RNC possui naturalmente embutida a extração de característica aparente, sendo assim, o processo de aprendizado está associado à descoberta de quais são as sub-regiões faciais e os padrões manifestados relevantes que permite determinar qual a emoção está sendo transmitida. A Figura 14 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 AlexNet foi a arquitetura mais encontrada na literatura para reconhecimento de emoção por expressão facial. A justificativa plausível pela AlexNet ter sido tão experimentada é por ser a pioneira da família de métodos conhecido como aprendizado profundo, desta forma, seu pioneirismo alcançou grande sucesso, inclusive vencendo o desafio ILSVRC-2012. O trabalho de Kim et al. (2016) utiliza AlexNet para reconhecer emoções, entretanto, a sua abordagem é destacada por combinar a AlexNet com outra rede profunda denominada *autoencoder*. Portanto, nesta abordagem, a rede *autoencoder* é responsável pelo alinhamento de face, desta forma, provendo uma grande contribuição do autor que apresentou uma solução funcional para o problema de rotação e alinhamento da face, e também, gerou um aumento na taxa de acurácia. 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 e abordagens híbridas de extração de características.

3.3.2 VGG

Barsoum et al. (2016) utilizou uma VGG de 13 camadas caracterizando uma arquitetura bastante profunda inclusive com camadas de *dropout* para alcançar maior generalização. Houve também o emprego da técnica de aumento de dados durante o treinamento, consistindo na geração de novas imagens aumentando a base em até 10 vezes através de operações matemáticas aplicadas em cada imagem original da amostra, desta forma, as imagens aumentadas não são uma cópia fiel da original possuindo variações referentes às poses das faces, rotações, variação da iluminação e outros atributos que contribuem para o aprendizado da rede por consistir de uma amostra rica de variabilidade.

O trabalho de Ng et al. (2015) propôs como novidade o emprego da transferência de aprendizado durante o treino para melhorar a performance de classificação. Neste caso, a estratégia foi inicialmente pré-treinar a VGG a partir da base do *ImageNet* (Deng et al., 2009) base tradicional do desafio ILSVRC, que é uma base que contém dezenas de objetos, e posteriormente, realizar um segundo treinamento a partir da base de expressões faciais para enfim reconhecer emoções.

3.3.3 GoogLeNet

Guo et al. (2016) propôs a *GoogLeNet* para reconhecer emoção por expressão facial. Houve um estudo comparativo entre modelos treinados a partir das arquiteturas *GoogLeNet* e *AlexNet*. A primeira alcançou melhores resultados, principalmente por possuir camadas *inceptions* que são módulos otimizados para a classificação de imagens, inclusive em fase de treinamento e produção consomem menos recursos computacionais do que *AlexNet*. Neste trabalho, também foi testado o classificador *kNN* na última camada, que alcançou melhor resultado que o tradicional *Softmax*. Desta forma, a *GoogLeNet* funcionou como extrator de características para servir de entrada o *kNN*.

Entretanto, vale ressaltar que, o *kNN* é uma técnica da família de métodos baseada em instâncias, isto é, não gera um modelo que encapsula o aprendizado obtido do treino. Portanto, inicialmente foi treinado a *GoogLeNet* com *Softmax* na última camada, posteriormente, retirou-se o *Softmax* da rede, desta forma, a última camada passou a ser a camada que antecedia o *Softmax*, isto é, uma camada de neurônios (também chamada de completamente conectadas). A saída dessa camada retorna as características extraídas da imagem pelo processamento na rede. Logo, para treinar o *kNN* foi necessário processar toda a base de treino salvando as características das imagens em uma outra base. Em Guo et al. (2016), o *kNN* para classificar uma instância recebe como entrada as características da imagem extraídas pela *GoogLeNet*, realiza um conjunto de cálculos de distâncias consultando a base de treino que contém somente as características das imagens e as instâncias que mais se aproximaram é a predição final. Obviamente, este processo de classificação há desvantagem pois para cada instância que se deseja classificar é necessário

analisar a base de características e quanto maior esta base maior será o tempo necessário para calcular o conjunto de distâncias para classificar.

3.3.4 *Ensemble*

Em Wen et al. (2017) a abordagem consistiu em combinar até 100 CNNs. A saída das classificações foram agrupadas em um nível de decisão baseada no produto das probabilidades para predizer a classe. O *ensemble* de CNNs obteve melhor resultado do que somente uma CNN. Entretanto, no experimento verificou-se que um *ensemble* com 10 CNNs atinge o melhor resultado. A medida que vai aumentando o tamanho do *ensemble* a acurácia permanece estável até às 60 CNNs. E a partir das 60 CNNs há um decréscimo da taxa de acurácia até às 100 CNNs. Vale destacar que o *ensemble* é uma abordagem bastante custosa, pois há necessidade de treinar vários modelos e há também a complexidade em definir como funcionará o nível de decisão.

Liu et al. (2016) implementou um *ensemble* de 3 CNNs com um único classificador no nível de decisão, o *Softmax*. Desta forma, o classificador recebia a extração das características das 3 CNNs. O interessante desse trabalho foi diminuir a complexidade no nível de decisão possuindo apenas o *Softmax*. Vale destacar que o *Softmax* é um classificador para problemas simples, nesta abordagem, o mesmo recebe um grande conjunto de informações dificultando seu aprendizado pelo alto grau de dimensionalidade dos dados.

Em Shin et al. (2016) o *ensemble* tinha 20 CNNs. Entretanto, as CNNs eram configuradas de tal forma que havia variação no tamanho da imagem nas camadas de entradas, isto é, uma CNN aceitava imagem com resolução de 60x60, outra por 80x80, assim por diante. Este trabalho experimentou o classificador *Support Vector Machine* (SVM) ao invés do tradicional *Softmax* na última camada. O SVM alcançou melhor acurácia embora necessite de um consumo maior de processamento e memória para classificar comparado ao *Softmax*.

A Tabela 3 apresenta um comparativo entre as arquiteturas com ênfase na acurácia. Vale ressaltar que, não é possível afirmar que o trabalho de Chen et al. (2017), uma *AlexNet*, por alcançar 99.1% de acurácia é melhor que a VGG de Barsoum et al. (2016) que alcança somente 84.9%, pois a base utilizada por Chen et al. (2017) é a CK oriunda de laboratório e o Barsoum et al. (2016) usa a base FER+ oriunda da natureza. Obviamente, é mais fácil classificar uma base oriunda de laboratório do que da natureza. Provavelmente, um modelo treinado por bases laboratoriais teria desempenho inferiores em cenário real comparado aos modelos treinados por bases da natureza. Portanto, com a finalidade de comparar de forma justa, devemos analisar a Tabela 3 somente os trabalhos que têm as mesmas bases nas colunas de treino e validação respectivamente.

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. (2017)	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. (2016)	FER	FER	73.73%
	Huang e Lu (2016)	FER	FER	76.9%
		CK+	CK+	97.3%
	Vo e Le (2016)	CK+	CK+	96.04%
	Yu et al. (2016)	CK+	CK+	98.7%
		MMI	MMI	98.6%
	Ng et al. (2015)*	FER/EmotiW	FER/EmotiW	55.6%
	Jung et al. (2015)	CK+/FER	CK+/FER	86.54%
	Li et al. (2015)	CIFE	CIFE	81.5%
		CK+	CK+	83%
VGG	Barsoum et al. (2016)	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. (2017)	FER	FER-Private	69.96%
		FER	CK+	76.05%
		FER	JAFFE	50.70%
		FER	EmotiW	34.09%
	Liu et al. (2016)	FER	FER	65.03%
	Shin et al. (2016)	FER/SFEW	FER-Test	66.67%
			SFEW	64.84%
			CK+	65.54%
			KDEF	50.66%
			JAFFE	49.17%

3.4 Aplicações

Há diversas aplicações para o reconhecimento de emoção no mundo real. Mapeando os trabalhos relacionados foi percebido que os pesquisadores de reconhecimento de emoção através das expressões faciais usando RNC, ultimamente concentraram seus esforços mais no desenvolvimento de algoritmos do que na aplicação em cenários reais. Desta forma, há um cenário favorável para pesquisas em casos reais tendo destaque para:

- Interação humano-computador (Barsoum et al., 2016; Chen et al., 2017; Liu et al., 2016; Wen et al., 2017), onde pode ser possível projetar interfaces que se adaptam ao estado emocional do usuário;

- Psiquiatria e cuidados médicos (Chen et al., 2017; Mayya et al., 2016; Wen et al., 2017), 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., 2015), pois pessoas com alto grau de deficiência visual, têm dificuldades para identificar qual a emoção que as pessoas ao seu redor estão transmitindo;
- Interação humano-robô (Jung et al., 2015; Shin et al., 2016), para que os robôs estejam dotados com a habilidade de interagir com humanos podendo adaptar-se às emoções dos humanos em volta, ou até mesmo transmitir emoção em uma face robótica para melhorar a interação com humanos;
- Personagens virtuais e animação (Vo e Le, 2016; Yu et al., 2016), dotando avatares com a habilidade de copiar a expressão facial emocional útil para gravação de filmes de animação. Além disso, pode ser utilizado em aplicações de animação como as rede sociais do *Snapchat*, *Facebook Messenger* e outros, que identificam a expressão facial do usuário e retorna alguma animação sobrepondo a expressão anteriormente detectada.

3.5 Resumo

Neste capítulo foram apresentados os trabalhos relacionados. Verificamos que o tema está em alta pela comunidade, diante do surgimento das arquiteturas de RNC que tem elevado o estado da arte de forma surpreendente em vários problemas de VC. Apesar de uma RNC ter embutido o pré-processamento em sua própria arquitetura, para o problema tratado por este trabalho, os trabalhos mostraram que pré-processamentos adicionais (não originais da RNC) melhoram a taxa de acurácia entre 2% a 8% dependendo do trabalho. É importante frisar que as principais arquiteturas empregadas são as vencedoras ou com pequenas modificações do tradicional problema de VC: o desafio ILSVRC. Estas arquiteturas funcionam para o problema de reconhecimento de emoção alcançando resultados jamais encontrados com técnicas anteriores, inclusive aproximando do nível humano.

Um conjunto relevante 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 prospectaram a possibilidade de usar em uma determinada área, mas de fato há pouca experiência na prática. Contudo, os modelos de reconhecimento de emoção têm alcançado taxas de acurácia confiáveis e o amadurecimento adequado para serem empregados na indústria e no mercado em geral.

4 Abordagem Proposta

A abordagem proposta por este trabalho é descrita neste capítulo e está dividida da seguinte forma. A Seção 4.1 apresenta a coleta de dados das imagens dos usuários. A Seção 4.2 detalha a detecção de face que é uma operação de pré-processamento aplicada à imagem. A Seção 4.3 define a fase de extração de características que é baseada em rede neural de convolução. A Seção 4.4 especifica o classificador que é responsável em definir a emoção da(s) face(s) detectada(s). A Seção 4.5 explana a integração entre os componentes e por fim há um resumo do capítulo na Seção 4.6.

4.1 Coleta de Dados

A coleta de dados é baseada em qualquer dispositivo que possui câmera fotográfica. Há dois cenários possíveis para a coleta de dados: (i) quando há um monitoramento periódico do usuário coletando um *streaming* de imagens de acordo com um intervalo de tempo, por exemplo, a cada 2 segundos e, (ii) quando deseja-se analisar uma foto isolada, sem um monitoramento constante, como a captura de um momento social especial corriqueiramente anexada em postagens nas redes sociais.

É importante destacar a coleta de dados nas aplicações mencionadas no Capítulo 3. Uma aplicação que é relevante para utilização deste trabalho são em ambientes educacionais inteligentes. Geralmente, as soluções de ambientes educacionais inteligentes possui um *tablet* ou *notebook* em sua arquitetura. Estes dispositivos executam um *software* educacional que permite a interação do estudante para participar das atividades. Por meio das plataformas educacionais abre-se a possibilidade da coleta de dados do estudante para gerar analíticos de aprendizado e, desta forma, o professor ou a plataforma pode oferecer um aprendizado personalizado. Então, para a viabilidade deste trabalho neste contexto, deseja-se que nas plataformas haja também o monitoramento por meio da câmera frontal (*tablet*) ou *webcam* (*notebook*) do estudante durante o desenvolvimento das atividades. Portanto, este cenário configura-se um ambiente favorecedor, pois a câmera frontal ou *webcam* sempre estarão capturando as reações da face com ângulo frontal.

Em um outro exemplo, destacamos uma aplicação para deficientes visuais em que a coleta de dados consiste de uma câmera apropriada para dispositivos vestíveis (do inglês: *wearables*). Esta câmera deve estar anexada à roupa do usuário na região peitoral, para que diante de uma conversa entre um grupo de indivíduos esta possa capturar imagens das faces das pessoas. Entretanto, este cenário, ao contrário dos ambientes educacionais, possui maiores desafios do ponto de vista da coleta de dados, pois o usuário que está com a câmera pode movimentar-se capturando imagens tremidas, desfocadas e borradas,

além de faces com ângulos variados. Tais problemas, entretanto, podem ser tratados por equipamentos de qualidade e tais discussões estão fora do escopo deste trabalho. Portanto, enquanto o monitoramento está ocorrendo, as imagens são enviadas para um repositório de entrada de dados.

4.2 Detecção de Face

A detecção de face e o posterior recorte é uma etapa de processamento aplicado à imagem. Visto que a detecção de face é um problema bastante explorado pela literatura. Quatro métodos populares da comunidade *Open Source* foram selecionados para definir de forma experimental qual funciona melhor. Os métodos selecionados foram: Viola-Jones (Viola e Jones, 2001)[1], *Histogram of Gradients* (HoG) com *Support Vector Machine* (SVM) (Dalal e Triggs, 2005)[2], *Multi-Task cascade Convolutional Neural Network* (MTCNN) (Zhang et al., 2016)[3] e uma CNN com *Maximum-Margin Object Detector* (MMOD) (King, 2015)[4].

Primeiramente, realizamos um experimento em que avaliamos os métodos de detecção facial para classificar as bases de dados de expressões faciais (ver Tabela 9). As mesmas bases também são utilizadas para treinamento e validação das CNNs na Seção 4.3. O experimento desta seção tem como premissa que existe uma face em todas as imagens analisadas. Os experimentos foram realizados na plataforma Google Colab[5], uma plataforma gratuita da Google para executar experimentos científicos, com a seguinte configuração: *GPU NVIDIA TESLA T4* que possui *16GB* de memória com processamento de *8TFLOPS* em precisão simples, *Intel(R) Xeon(R) CPU @ 2.30GHz* com 2 núcleos e *12GB de RAM*.

Os resultados são apresentados na Tabela 4. A coluna "Detectou" destaca a quantidade de imagens que o método detectou uma face e a coluna "Não detectou" o contrário. É notório que o desempenho dos quatro métodos foram superiores nas seguintes bases: CK+, KDEF, JAFFE e RAFD, com exceção do Viola Jones que não foi bem na RAFD. Coincidentemente essas bases de dados pertence ao conjunto de imagens capturadas em laboratório. De certa forma, isto é um resultado esperado pois imagens capturadas em laboratório são mais fáceis de detectar face do que em ambientes externos ou natural. Neste experimento, de acordo com a Seção das Métricas de Avaliação de Desempenho (ver seção 2.10), somente faz sentido calcular a métrica de precisão, pois temos a premissa que todas as imagens são de faces, então o que deseja-se medir é a fração de faces que o método não conseguiu detectar ou o quão bem o método trabalhou. Portanto, analisando a pre-

[1] Implementação disponível em: https://bit.ly/2RIY0Nd
[2] Implementação disponível em: https://bit.ly/2FG7Iew
[3] Implementação disponível em: https://bit.ly/2RQNgMQ
[4] Implementação disponível em: https://bit.ly/2FG7Iew
[5] Google Colab: https://colab.research.google.com/

Tabela 4 – Avaliação dos métodos para detecção de face usando bases com expressões faciais

Base de Dados	**Métodos**	**Classificação**		**Precisão**	**Tempo (s)**
		Detectou	Não detectou		
CIFE	Viola Jones	4433	10324	30.04%	309.59
	HoG SVM	12478	2279	84.56%	139.35
	MMOD-CNN	14016	741	94.98%	35.8
	MTCNN	13615	1142	92.26%	250.08
CK+	Viola Jones	2999	19	99.37%	66.53
	HoG SVM	3018	0	100%	30.27
	MMOD-CNN	2989	29	99.04%	7.8
	MTCNN	3013	5	99.83%	60.4
FER	Viola Jones	9894	25993	22.57%	721.74
	HoG SVM	24410	11477	68.02%	33.81
	MMOD-CNN	33931	1956	94.55%	86.39
	MTCNN	25998	9889	72.44%	627.71
KDEF	Vila Jones	2381	559	80.99%	63.16
	HoG SVM	2916	24	99.18%	28.13
	MMOD-CNN	2936	4	99.86%	7.61
	MTCNN	2933	7	99.76%	47.94
Nova Emotions	Viola Jones	151	33524	0.45%	664.35
	HoG SVM	10274	23401	30.51%	310.39
	MMOD-CNN	29473	4202	87.52	82.27
	MTCNN	25839	7836	76.73%	451.56
JAFFE	Viola Jones	213	0	100%	4.94
	HoG SVM	213	0	100%	2.02
	MMOD-CNN	213	0	100%	0.53
	MTCNN	213	0	100%	4.29
RAFD	Viola Jones	1724	3100	35.74%	101.32
	HoG SVM	4287	537	88.87%	50.54
	MMOD-CNN	4820	4	99.92%	12.56
	MTCNN	4824	0	100%	87.92

cisão, no geral, os métodos que alcançaram melhores desempenhos foram: MMOD-CNN, MTCNN, HoG-SVM e Viola Jones. Sendo que o MMOD-CNN foi o método mais estável tanto para bases de laboratório como para ambientes e poses naturais. Vale destacar que parte do MMOD-CNN é executada em GPU, e a outra parte, uma operação pesada de pré-processamento, é executada em CPU. Desta forma, este método demanda mais recursos computacionais que os demais, dificultando seu emprego em aplicações de tempo real ou quando há uma grande demanda de imagens para processar.

Um segundo experimento foi realizado com a intenção de medir o desempenho dos métodos na base VOC-2007 (Everingham et al., 2010). Esta amostra de imagens possui conteúdos diversos como: (i) animais: pássaro, gato, vaca, cachorro, ovelha e cavalo; (ii) veículos: avião, bicicleta, barco, ônibus, carro, motocicleta e trem e (iii) ambientes

Tabela 5 – Matriz de confusão e acurácia para detecção de face usando a base VOC-2007

Métodos	**Matriz de Confusão**			**Acurácia**
		Face	Outros	
Viola Jones	Face	507	1500	68.90%
	Outros	40	2905	
HoG SVM	Face	711	1296	73.22%
	Outros	30	2915	
MMOD-CNN	Face	829	1178	75.94%
	Outros	13	2932	
MTCNN	Face	944	1063	76.55%
	Outros	98	2847	

Tabela 6 – Matriz de confusão juntando a base de dados de expressões faciais e a VOC-2007

Métodos	**Matriz de Confusão**		
		Face	Outros
Viola Jones	Face	22302	75019
	Outros	40	2905
HoG-SVM	Face	58847	39014
	Outros	30	2915
MMOD-CNN	Face	89207	8114
	Outros	13	2932
MTCNN	Face	77379	19942
	Outros	98	2847

internos: garrafa, cadeira, mesa, televisão e outros, além de pessoas. Na Tabela 5 é apresentado o resultado desse experimento. Foi considerado que a classe "face" é equivalente às imagens da classe pessoa na VOC-2007 e as demais classes são denominadas como "outros". Analisando os resultados pela acurácia podemos eleger que o melhor método foi o MTCNN alcançando 76.55%, embora o MMOD-CNN tenha-se aproximado deste resultado com 75.94% e em seguida o HoG-SVM com 73.22% de acurácia. Novamente, o pior método foi o Viola Jones com 68.90% de acurácia.

Na Tabela 7 é apresentado às métricas de avaliação de desempenho usando tanto a base de expressões faciais como a VOC-2007. Para calcular estas métricas foi utilizada a matriz de confusão condensada entre as bases de expressões faciais e a VOC-2007 apresentada na Tabela 6. Vale ressaltar que a base VOC-2007 tem 4.952 imagens enquanto que as bases de expressões faciais somam 95.314 imagens. Analisando os resultados verificamos que todos os métodos foram excelentes na métrica de revocação, isso se explica pelo fato de haver muito mais imagens de face do que de não face (outros), ou seja, é um conjunto desbalanceado. Desta forma, a quantidade de erros que houve ao detectar "outros" como "face" (ver Tabela 6) é irrisória sobre a quantidade de faces corretamente identificadas.

Tabela 7 – Avaliação dos métodos para detecção de face utilizando base de expressões faciais e VOC-2007

	Revocação	**Precisão**	**F1-score**	**Acurácia**	**Tempo (s)**
Viola Jones	99.82%	22.92%	37.27%	25.14%	1931.63
HoG-SVM	99.95%	60.13%	75.09%	61.27%	594.51
MMOD-CNN	99.99%	91.66%	95.64%	91.89%	232.96
MTCNN	99.87%	79.51%	88.54%	80.01%	1529.9

A coluna precisão apresenta a porcentagem de faces que o método não conseguiu identificar. É conclusivo que o Viola Jones é o pior método neste quesito. Em contrapartida, o MMOD-CNN destacou-se com 91.66% de precisão, significando que realmente detecta as faces. A métrica de F1-score é uma média ponderada entre precisão e revocação, sendo uma maneira mais justa de eleger o melhor método que foi o MMOD-CNN com 95.64%, logo em seguida vem o MTCNN com 88.54% e o HoG-SVM com 75.09%, enquanto que o Viola Jones está bem atrás dos demais com 37.27%. A coluna acurácia retrata a precisão geral do método observando tanto a "face" como "outros". Na acurácia repete-se a sequência da F1-score em que o melhor método é o MMOD-CNN seguido do MTCNN e assim por diante. Analisando a coluna de tempo na unidade de segundos, o método mais veloz foi o MMOD-CNN com 232.96, seguido de HoG-SVM com 594.51, na sequência aproximadamente 6 vezes e meio mais lento que o MMOD-CNN vem o MTCNN, e por último, o método mais lento de todos foi o Viola Jones com 1931.63 segundos.

Portanto, diante destes experimentos pode-se concluir que quando dispõe de muitos recursos computacionais o método a ser empregado como detecção de face é o MMOD-CNN. Em outro cenário, quando os recursos são modestos em que não é possível executar de forma satisfatória o MMOD-CNN, mas há uma GPU disponível, o método a ser escolhido é o MTCNN. No entanto, quando há somente CPU sem a disponibilidade de GPU, o método selecionável é o HoG-SVM. De acordo com os experimentos, não é aconselhável usar o Viola Jones em hipótese nenhuma. Vale destacar que o Viola Jones foi um método estado-da-arte e revolucionário para sua época, porém isto ocorreu décadas atrás.

4.3 Rede Neural de Convolução

A rede neural de convolução (RNC) é a parte central desta abordagem. Inicialmente, necessita-se de uma etapa de treinamento, em que as imagens de expressões faciais são processadas a fim de aprender a extrair os contornos, padrões, formas e características relevantes para diferenciar as emoções. Portanto, após o treinamento, gera-se um modelo

especializado em processar imagens de expressões faciais emocionais capaz de traduzir uma imagem tridimensional em um vetor unidimensional com a informação que representa a emoção na imagem ou o conjunto de características extraídas. Além disso, o vetor unidimensional tem o tamanho bem inferior ao tridimensional. Também, o modelo tem uma camada final de decisão que é padrão em redes neurais de convolução para fins de classificação denominado *Softmax*. Esta camada final recebe o vetor unidimensional para decidir ou prever qual emoção está na imagem. Portanto, o modelo é um classificador.

Um estudo experimental foi realizado a fim de comparar as seguintes arquiteturas de RNCs: VGGNet, ResNet-50, Inception-V3, InceptionResNet-V2 e a MobileNet-V2. Este estudo tem como objetivo eleger a arquitetura que gerou o melhor modelo avaliando a precisão, revocação, f1-score e a acurácia, tanto para rodar em sistemas com enorme recursos (computação em nuvem), como para executar em hardware com recursos limitado comum em sistemas embarcados. As arquiteturas mencionadas receberam como entrada dois tamanho de imagens quadradas em 185 e 210 *pixels*.

4.3.1 Preparação dos Dados

A preparação dos dados consistiu na formação de duas bases de dados: treino e validação. A base de treino foi utilizada para treinar os algoritmos e, geralmente, é a base que tem a quantidade majoritária de instâncias. Uma base de treino formada erroneamente reflete nos modelos ocasionando *underfitting* (não aprendeu a resolver o problema) ou *overfitting* (não aprendeu a generalizar o problema ou somente funciona com a base de treino caracterizando um super vício). Modelos com *underfitting* ou *overfitting* não são confiáveis e possuem problemas de acurácia. Diante disso, a preparação dos dados é essencial para obtenção de modelos com bons resultados. Já a base de validação é usada para validar os modelos que estão sendo gerados durante o treinamento. Pois, ao fim de cada época calcula-se a função de perda (ou custo) e a acurácia na base de validação. Depois do treinamento de cada arquitetura, a base de validação é utilizada para calcular métricas de avaliação de desempenho a fim de verificar qual modelo alcançou melhores resultados.

Um dos objetivos da revisão sistemática (ver Seção A) foi encontrar as bases de dados mais populares utilizadas pela comunidade científica. A Tabela 8 apresenta as bases de dados localizadas na revisão sistemática. As bases de dados mais populares foram: CK+, FER e a JAFFE. Os dados advindos dessas bases foram usados neste estudo experimental. Partindo do pressuposto que as bases de dados aplicadas em aprendizagem de máquina devem ser bastantes diversificadas para gerar modelos com bons resultados. As bases de treino e validação foram formadas a partir da concatenação de todas as bases a fim de gerar a diversificação nos dados na seguinte divisão: 80% para treino e 20% para validação. Uma etapa de limpeza foi executada em todas as bases com intuito de

retirar amostras com baixa representatividade. Esta etapa de limpeza consistiu no uso do *MMOD-CNN* (veja Seção 4.2), pois há várias imagens com forte ruídos, por exemplo, faces com rotações de grau elevado, incompletas, obstruídas e imagens de personagens de animação (não humanos). Quando o *MMOD-CNN* não detecta a face é um indício de que a imagem sofre de algum tipo de ruído já mencionado. Após o processo de limpeza, as imagens resultantes foram distribuídas conforme a Tabela 9, sendo divididas por base de treino e validação. As bases FER e NovaEmotions são as que tem maior quantidade de amostra, enquanto que a JAFFE o contrário. A Tabela 10 indica a distribuição das imagens referentes as emoções (classes). Vale ressaltar que a emoção alegria é a classe que tem mais instâncias representativas, ao contrário da emoção medo que possui a menor quantidade de amostra.

Tabela 8 – Bases de dados encontradas na literatura

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

4.3.2 Materiais

Para a realização desse estudo experimental foi necessário o uso de alguns materiais. O *framework Tensorflow* (`https://www.tensorflow.org/`) e o *Keras* (`https://keras.io/`) foram utilizados para implementar e treinar as RNCs. A biblioteca *OpenCV*

Tabela 9 – As bases de dados foram concatenadas e divididas em duas bases: treino e validação. Na seguinte porcentagem: 80% para treino e 20% para validação.

Base de Dados	**B. de Treino**	**B. de Validação**	**Total de Imagens**
CIFE	10678	2673	13351
CK	2413	605	3018
FER	25213	6306	31519
JAFFE	170	43	213
KDEF	2348	588	2936
NOVAEMOTIONS	25837	6460	32297
RAFD	3857	967	4824
Total de Imagens	70516	17642	88158

Disponível em: `https://bit.ly/2YdgiaW`

Tabela 10 – Distribuição das emoções (classes) nas bases de treino e validação. As emoções também foram divididas em: 80% para treino e 20% para validação.

Classe	B. de Treino	B. de Validação	Total de Imagens
Raiva	6207	1553	7760
Desgosto	4150	1039	5189
Medo	5702	1428	7130
Alegria	23506	5878	29384
Tristeza	8556	2141	10697
Surpresa	10629	2660	13289
Neutralidade	11766	2943	14709
Total de Imagens	70516	17642	88158

Disponível em: https://bit.ly/2YdgiaW

(https://opencv.org/) foi usado para manipular as imagens. Os experimentos foram realizados em dois computadores com a seguinte configuração: *GPU NVIDIA GTX 1060-TI* que possui *6GB* de memória com processamento de *4TFLOPS* em precisão simples, *Intel Core-i7* e *16 GB de RAM*, e na plataforma Google Colab[6], com a seguinte configuração: *GPU NVIDIA TESLA T4* que possui *16GB* de memória com processamento de *8TFLOPS* em precisão simples, *Intel(R) Xeon(R) CPU @ 2.30GHz* com 2 núcleos e *12GB de RAM*.

4.3.3 Treinamento

O treinamento consistiu no uso dos materiais descritos na Seção 4.3.2 e das bases de treino e validação descritas nas Tabelas 9 e 10. Inicialmente, os pesos da RNC estão em um estado inicial aleatório. Depois, a base de treinamento é dividida em *batches* para serem processados individualmente e a cada *batch* processado é contado uma interação e atualizado os pesos da RNC. Após a base de treinamento ter sido processada por completo (depois de rodar todos os *batches*) é considerado uma época e um novo modelo. O melhor modelo é aquele que atingir a menor taxa de função de perda (ou custo) na base de validação. A Figura 15 ilustra este conceito. Portanto, a estratégia adotada durante a fase de treinamento foi monitorar a função de perda na base de validação ao fim de cada época registrando a menor taxa. Quando ocorre a diminuição da menor taxa da função de perda já registrada, o modelo é imediatamente salvo no disco, caso não haja a diminuição, obviamente esse modelo é pior do que já registrado anteriormente não fazendo sentido salvá-lo.

Outros artifícios foram usados no treinamento das RNCs com intuito de gerar modelos com maior acurácia. Vale ressaltar a inclusão dos pesos nas classes. Este parâmetro funciona adicionando um peso maior nas classes minoritárias e o contrário nas classes majoritárias de forma proporcional ao tamanho da amostra. Este peso atua punindo de maneira mais contundente na função de perda quando o modelo erra uma instância da

[6] Google Colab: https://colab.research.google.com/

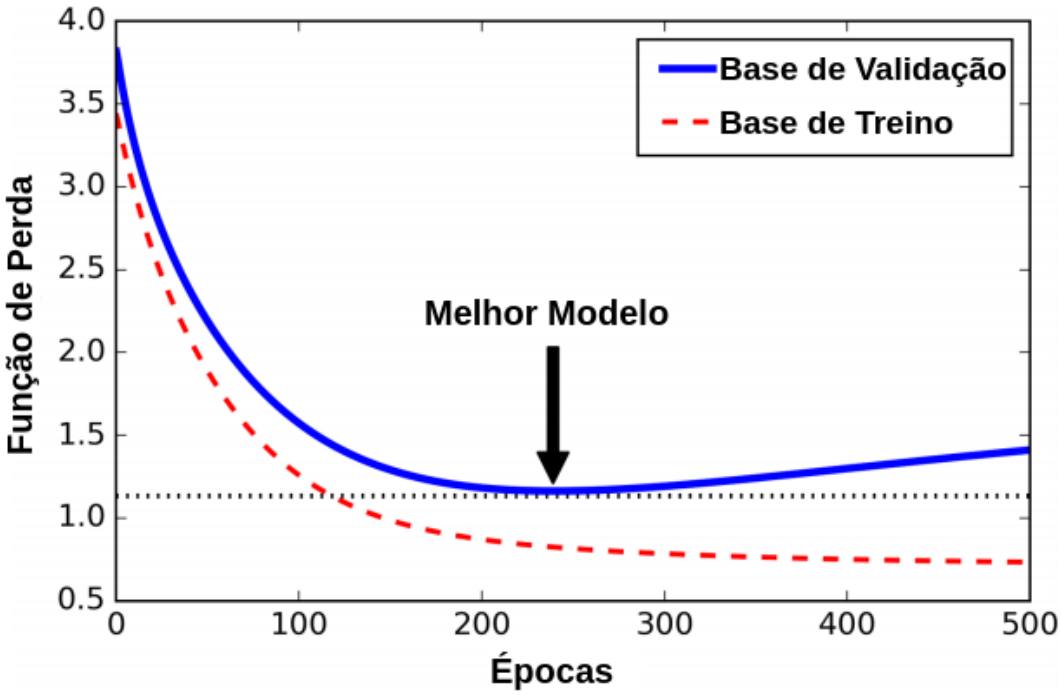

Figura 15 – Gráfico da função de perda durante o treinamento

Fonte: (Géron, 2017)

classe minoritária. Portanto, de acordo com a Tabela 10 errar medo e desgosto pesa mais negativamente para o aumento da função de perda do que errar alegria ou neutralidade. Desta forma, a função de treinamento possui mais um artifício para evitar *overfitting* nas classes majoritárias forçando a aquisição de aprendizado também nas classes minoritárias.

A taxa de aprendizado (do inglês: *Learning Rate*) é uma configuração importante no treinamento. Esta taxa especifica para o otimizador o tamanho do passo que os pesos do modelo devem ser atualizados. Em outras palavras, funciona como uma espécie de acelerador veicular. Para exemplificar a importância deste parâmetro, quando há uma alta taxa de aprendizado ocorre que a atualização dos pesos é tão forte que provoca *underfitting*, pois a atualização foi agressiva e o melhor modelo ficou para trás. Então o contrário, quando a taxa de aprendizado é baixa, a atualização é tão suave que as épocas vão se passando e o aprendizado retido é baixo ocasionando a saturação dos neurônios da RNC, isto é, tais neurônios não aprendem mais nada. Logo, em ambos os casos, não é possível gerar um modelo interessante. O ideal é que no estágio inicial de treinamento, pelo fato de o modelo não possuir nenhum aprendizado, tenha uma taxa um pouco alta para acelerar o aprendizado inicial e ao decorrer das épocas essa taxa vai decrescendo, pois quando se aproxima do modelo ideal a taxa de aprendizado deve ter um valor baixo. Este comportamento da taxa de aprendizado ser dinâmico durante o treinamento também está empregado neste trabalho.

A parada antecipada (do inglês: *Early Stopping*) também foi empregada no treina-

mento. Este recurso está intrinsecamente ligado ao fato de monitorar a função de perda na base de validação como mostra a Figura 15. Pois, o comportamento da função de perda é a seguinte: o início do treinamento está com um valor bem alto, significa que o modelo ainda não aprendeu nada. Com o passar das épocas, o modelo vai treinando e aprendendo, e consequentemente, vai reduzindo o valor de perda na validação até atingir o piso que é o marco do melhor modelo a ser gerado no treino. Após encontrar o piso, a função de perda na validação tem seu valor acrescido no decorrer das épocas demonstrando que não haverá outro modelo melhor. Enquanto que paralelamente o valor de perda no treino tende a zero representando que não há mais aprendizagem a ser retida. Desta forma, a parada antecipada funciona salvando o marco do piso e conta quantas épocas faz que o piso não é atualizado. Existe um parâmetro denominado de paciência (do inglês: *patience*) que é um limiar onde, caso a contagem de atualização do piso seja superior ao valor de paciência, o treinamento é automaticamente encerrado. Sendo assim, a parada antecipada caracteriza-se como um recurso inteligente pois automaticamente detecta que é o momento correto de encerrar o treinamento e gerando vantagens como: economia de recursos computacionais, garantia que o treinamento alcançou o piso na função de perda e a não incidência de encerrar precocemente o treino.

4.3.4 Execução do Treinamento

Inicialmente, de maneira empírica, foi verificada várias configurações de parâmetros. As configurações que retornavam resultados fracos foram descartadas. O otimizador (gradiente descendente) que funcionou melhor para cada arquitetura foi o Adadelta (Zeiler, 2012). Os seguintes otimizadores foram testados: SGD (Robbins e Monro, 1951), RMSProp (Tieleman e Hinton, 2012) e o Adam (Kingma e Ba, 2014), porém todos foram descartados com resultados pouco satisfatórios. Por isso, o Adadelta foi o otimizador aplicado para todas as arquiteturas. Outros parâmetros a serem definidos foram referentes à taxa de aprendizagem (TA): inicial, final e a redução. A TA inicial é o valor atribuído quando o treinamento começa. A cada época esse valor vai reduzindo até atingir o piso que seria a TA final. A TA de redução subtrai a porcentagem correspondente do atual valor da TA na época. Por exemplo, é definido que a TA inicial 1.0, TA final é 0.75 e a TA redução 15%. O treinamento inicia com TA no valor de 1.0, após o fim da primeira época é reduzido 15%, ou seja, para a segunda época TA passaria a ser 0.85 e na terceira época TA seria 0.72. Entretanto, como este valor é menor que a TA final, então a partir da terceira época o valor de TA não se alteraria mais congelado no valor de piso que é 0.75. A Tabela 11 apresenta os parâmetros finais utilizados para o treinamento por arquitetura. A paciência é a configuração da parada antecipada (ver Seção 4.3.3). Outro parâmetro definido foi o tamanho de *batch* sendo igual a 28. Infelizmente, por limitações de *hardware* não foi possível variar o *batch*, então congelamos no maior valor que o *hardware* suportou sendo múltiplo de 7 que é o total de expressões faciais (classes) a serem treinadas.

Tabela 11 – Definição de parâmetros para o treinamento das redes neurais de convolução

Arquiteturas	**Otimizador**	**TA Inicial**	**TA Final**	**TA Redução**	**Paciência**
InceptionResNetV2	Adadelta	0.08	0.01	15%	4
InceptionV3	Adadelta	0.08	0.01	15%	3
MobileNetV2	Adadelta	0.1	0.05	10%	3
ResNet50	Adadelta	0.08	0.01	15%	3
VGG19	Adadelta	0.08	0.01	15%	3

TA = Taxa de Aprendizagem

As arquiteturas VGG19, ResNet-50, Inception-V3, InceptionResNet-V2 e a MobileNet-V2 foram treinadas usando dois tamanhos de imagens quadradas na entrada: 185 e 210 *pixels*. A Figura 16 ilustra função de perda na base de validação no treino com as imagens em 185 *pixels* e a Figura 17 demonstra para 210 *pixels*. Em ambas as figuras, analisando a função de perda para as arquiteturas InceptionResNetV2 e ResNet50, aparenta que a taxa de aprendizado está muito alta, entretanto foram testados valores menores e não foi possível obter resultados melhores. De uma forma global, as arquiteturas VGG19 e InceptionV3 obtiveram os melhores resultados. Todas as arquiteturas com exceção da MobileNetV2 precisaram de até 5 épocas para alcançar o piso na função de perda demarcado pelo círculo no gráfico. Enquanto que a MobileNetV2 precisou de 14 e 23 épocas, respectivamente, com imagens quadradas de 185 e 210 *pixels*.

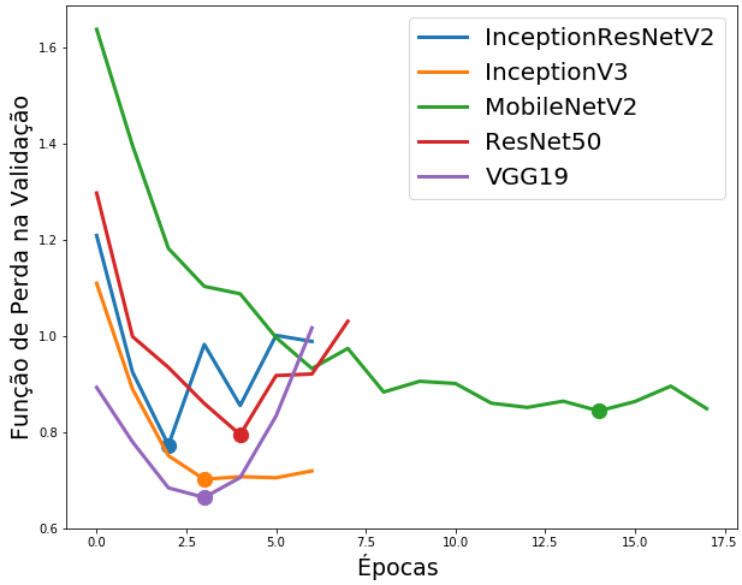

Figura 16 – Gráfico da função de perda na base de validação com imagens em 185 *pixels*

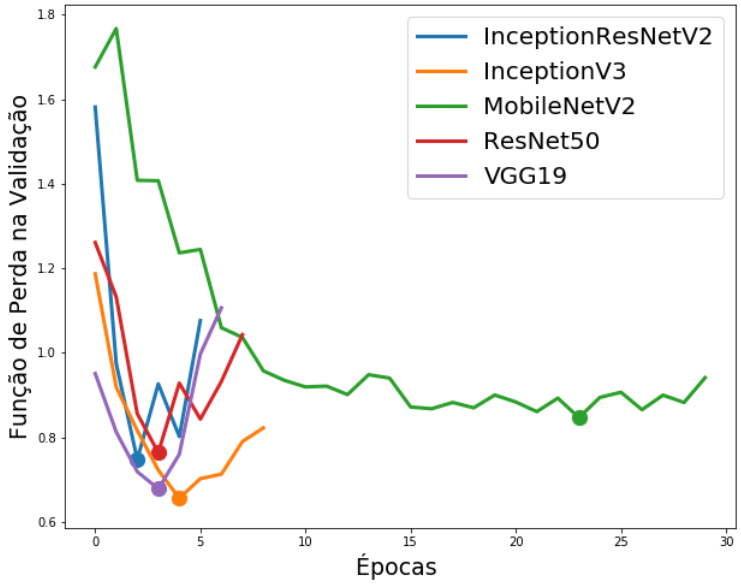

Figura 17 – Gráfico da função de perda na base de validação com imagens em 210 *pixels*

A Figura 18 ilustra a acurácia na base de validação calculada a partir do treino, época por época, para as imagens quadrada de 185 *pixels*, enquanto que a Figura 19 para 210 *pixels*. O círculo no gráfico demarca o melhor modelo selecionado pela função de perda na validação. Lembrando que esta acurácia é considerando o classificador *Softmax* na última camada que é o padrão nas arquiteturas experimentadas. Analisando as imagens com tamanho em 185 *pixels*, o modelo (demarcado pelo círculo) que atingiu o maior nível de acurácia foi a VGG19, enquanto que em 210 *pixels* deu um empate entre a InceptionV3 e a VGG19. A pior arquitetura em acurácia em ambos os tamanhos de imagem foi a MobileNetV2. Vale ressaltar que a MobileNetV2 é a rede projetada para sistemas embarcados com intuito de consumir menos recursos computacionais e é esperado que tenha menor acurácia do que as demais arquiteturas. Embora a acurácia tenha aumentado após a época do melhor modelo, este fato é enganoso pois significa que o modelo começou o *overfitting* nas classes majoritárias. Por isso, paralelamente a função de perda (veja Figuras 16 e 17) começa a subir demonstrando que o modelo está errando as classes minoritárias. Deve-se ter cuidado ao analisar acurácia, pois quando a avaliação é através de uma base muito desbalanceada, o fato de acertar somente as classes majoritárias já traduz em uma alta acurácia, em contrapartida, o modelo não aprendeu e erra as classes minoritárias.

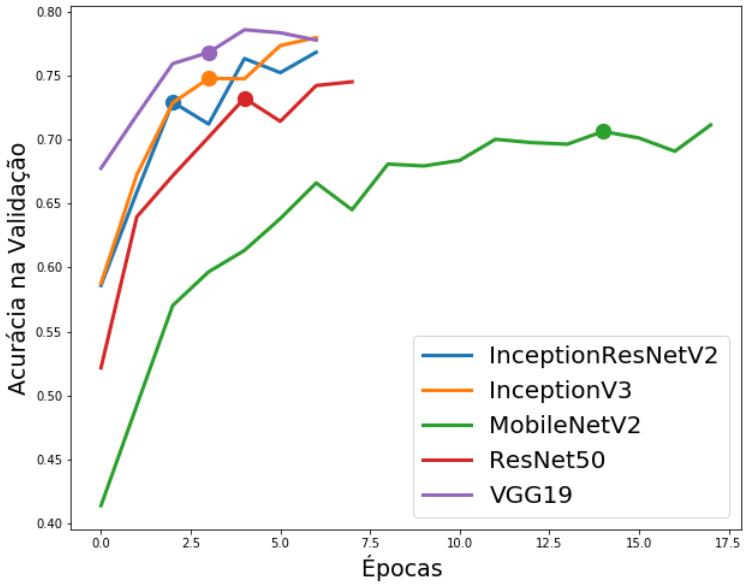

Figura 18 – Gráfico de acurácia na base de validação com imagens em 185 *pixels*

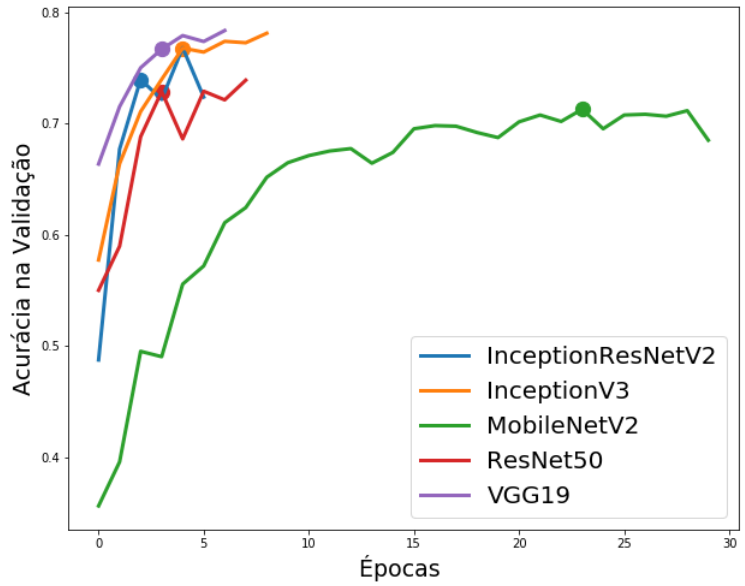

Figura 19 – Gráfico de acurácia na base de validação com imagens em 210 *pixels*

A Tabela 12 apresenta a quantidade de camadas e um resumo da contagem de parâmetros por arquitetura. Os parâmetros representam a quantidade de neurônios e pesos a serem ajustados nas camadas e suas conexões durante o treinamento. A arquitetura que tem a maior quantidade de parâmetros é a VGG19 totalizando 83.957.575 enquanto que a MobileNetV2 tem a menor quantidade com 209.303. A diferença de parâmetros é tão relevante que a MobileNetV2 chega a ter somente 0.25% do total de parâmetros da VGG19. Essa diferença também é ressaltada no tamanho do modelo para armzenamento em que a VGG19 possui 1.3GB enquanto que a MobileNetV2 tem apenas 3MB. Todos os modelos estão disponíveis em: `https://bit.ly/2Ym9qMl`.

Tabela 12 – Resumo da profundidade e contagem de parâmetros por arquitetura

Arquitetura	**Parâmetros Treináveis**	**Parâmetros Não Treináveis**	**Total de Parâmetros**	**Total de Camadas**	**Tamanho do Modelo**
InceptionResNetV2	54.286.951	60.544	54.347.495	782	654MB
InceptionV3	21.782.695	34.432	21.817.127	313	262MB
MobileNetV2	199.303	10.000	209.303	158	3MB
ResNet50	23.548.935	53.120	23.602.055	177	283MB
VGG19	83.957.575	0	83.957.575	23	1.3GB

4.3.5 Extração de Características

A extração de características é uma etapa fundamental em problemas de classificação de imagem. Esta fase é responsável por retirar da imagem de entrada as principais informações (características) relevantes capazes de diferenciar as classes (emoções) para enviar ao classificador. A RNC é projetada para ser um exímio extrator de características bioinspirada na visão dos mamíferos (Géron, 2017). O foco é identificar as zonas da imagem que são características daquela classe e representar em uma informação numérica. Por exemplo, ao extrair as características de uma expressão facial de alegria, que geralmente é acompanhada de um sorriso, neste caso, a missão é conseguir transformar essa característica (sorriso) em uma informação numérica e o que não for importante para classificar alegria seja ignorado. A RNC chama atenção pelo seu poder de redução de dimensionalidade demonstrada na Tabela 13. A coluna de entrada representa a quantidade de características ou informações que a camada de entrada da RNC recebe. Por exemplo, neste trabalho estamos avaliando dois conjuntos de imagem, quadradas em RGB com 185 e 210 *pixels*, respectivamente. Portanto, temos para imagens com tamanho quadrado de 185 *qtd_características_entrada = 185 * 185 * 3* igual a 102675 enquanto que para 210 temos *qtd_características_entrada = 210 * 210 * 3* igual a 132300. A camada de saída é o tamanho do vetor com as características extraídas que a RNC retorna. Então, das arquiteturas avaliadas neste trabalho, a MobileNetV2 é a que retorna a menor quantidade de características, seguido da InceptionResNetV2, InceptionV3, ResNet50, e por último, a VGG19 que retorna a maior quantidade de características. A princípio retornar a menor quantidade características é positivo, pois significa que existe menos informações para o classificador processar e aprender. A coluna de redução na Tabela 13 mostra em porcentagem a proporção da redução de dimensionalidade realizada nas características de entrada comparada às características de saída. Este dado faz alusão ao poder de redução de dimensionalidade da RNC que, para MobileNetV2 chega a 99.03% quando o tamanho da imagem é 210 *pixels*.

Tabela 13 – Contagem de características na camada de entrada e saída por arquitetura

Arquitetura	**Entrada**		**Saída**	**Redução**	
	185 *pixels*	210 *pixels*		185 *pixels*	210 *pixels*
InceptionResNetV2	102675	132300	1536	98.50%	98.83%
InceptionV3	102675	132300	2048	98%	98.45%
MobileNetV2	102675	132300	1280	98.75%	99.03%
ResNet50	102675	132300	2048	98%	98.45%
VGG19	102675	132300	4096	96.01%	96.90%

4.4 Classificador

O classificador é responsável por determinar qual é a emoção (classe) analisando as características extraídas pela RNC. O conjunto de características extraídas é um vetor que possui uma quantidade de elementos bastante inferior à imagem original. O conteúdo do vetor consiste no conjunto de informações representativas para diferenciação das emoções que a rede neural aprendeu a extrair no treinamento. Então, posteriormente, o classificador também é treinado para separar ou definir as emoções usando as características extraídas da imagem.

Um estudo experimental foi realizado a fim de comparar os seguintes classificadores: Softmax, RandomForest, SVM e o KNN. Este estudo consistiu em adaptar cada classificador recebendo as características de cada arquitetura treinada. Em outras palavras, cada classificador foi colocado na última camada de cada arquitetura de RNC, respectivamente, tanto para 185 como para 210 *pixels*. Todos os classificadores experimentados têm a propriedade de estimar a probabilidade para cada emoção: neutralidade, raiva, alegria, tristeza, desgosto, medo e surpresa. E essa estimativa é distribuída entre as classes (neutralidade: 0.95, alegria: 0.025 e surpresa: 0.025), em que o somatório das probabilidades é igual a 1, e a emoção elegida é a que tem maior probabilidade, nesse caso, a neutralidade com 0.95.

4.4.1 Treinamento

Inicialmente, todos os modelos de RNC treinados na Seção 4.3.4 foram reunidos. Tais modelos possuem na última camada um classificador que é o Softmax, portanto durante o treinamento da RNC também treinou-se o Softmax. Desta forma, os classificadores RandomForest, SVM e o KNN são os que faltam a serem treinados. As bases de dados para treino e validação dos classificadores também foram as mesmas usadas pelas RNCs descritas na Tabela 9. Porém, considera-se um cenário diferente, pois as RNCs treinam com as imagens, e os classificadores são treinados com as características extraídas das imagens. Sendo assim, para cada imagem da Tabela 9 deve-se extrair as características usando cada modelo de RNC armazenando em uma nova base de dados. Treinou-se 5

Tabela 14 – Parâmetros aplicados nos classificadores durante o treinamento

Classificador	Parâmetro	Valor Inicial	Valor Final	Intervalo
KNN	K	3	18	3
Random Forest	D	3	27	3
SVM	C	0.007	3	0.03

modelos (InceptionResNetV2, InceptionV3, MobileNetV2, ResNet50 e VGG19) usando imagens quadradas com 185 *pixels*, e depois, com 210 *pixels*. Portanto, criou-se no total 10 modelos de RNCs, consequentemente, será necessário gerar outras 10 bases de dados somente com as características para treinar os classificadores. Vale ressaltar que, o tamanho da instância (características extraída) de entrada para cada classificador está na Tabela 13. Por exemplo, para cada classificador que estiver na última camada da InceptionResNetV2 vai receber um vetor de 1536 valores, enquanto que para VGG19 um de 4096 posições.

Cada classificador tem seus parâmetros particulares para serem ajustados. O experimento consistiu em variar os principais parâmetros para encontrar qual funcionou melhor. Por exemplo, para o KNN tem-se o valor de K, que representa a quantidade de vizinhos que mais aproximam da instância. Para o RandomForest tem-se o D, que é atribuído ao tamanho da profundidade das árvores. Enquanto que para o SVM tem o parâmetro denominado de C, que é justamente a penalidade aplicada durante a função de otimização interna deste método no treino. A Tabela 14 apresenta os parâmetros aplicados nos classificadores durante o treinamento. A coluna Valor Inicial significa que o primeiro experimento começa com este valor, por exemplo, para o K é igual a 3, no próximo experimento esse K passaria a ser 6, pois vai intervalando de 3 em 3 (valor da coluna Intervalo) até alcançar o valor final 18. O mesmo vale para os demais parâmetros com seus respectivos valores. Todos os modelos estão disponíveis em: `https://bit.ly/2Ym9qMl`.

4.4.2 Resultados

O estudo experimental adotado gerou um conjunto vasto de resultados. Lembrando que há cinco arquiteturas de RNCs, quatro classificadores com vinte e seis variações e dois tamanhos diferentes de imagens. Desta forma, gerou-se duzentos e sessenta resultados a serem analisados. Diante de tantos resultados, o foco foi avaliar qual o melhor classificador por arquitetura, e posteriormente, qual a melhor combinação entre arquitetura e classificador.

A Tabela 15 apresenta os resultados do melhor classificador por cada arquitetura e ressalta a melhor combinação entre arquitetura e classificador. Vale destacar que a base geral tem 17642 imagens, em que a emoção com maior amostragem é a alegria com 5878, ao contrário de desgosto que possui a menor taxa de amostragem com 1039, estes dados

podem ser consultados na Tabela 10. Os resultados referentes as demais arquiteturas treinadas pelas imagens quadradas de 185 *pixels* estão no Anexo B, enquanto que para 210 *pixels* estão no Anexo C.

Das cinco arquiteturas de RNC o Random Forest foi o melhor classificador em três oportunidades. Exatamente com o mesmo parâmetro de profundidade das árvores equivalente a 27. O kNN e SVM ambos contabilizaram como o melhor classificador uma vez para ResNet50 e VGG19, respectivamente. Em nenhuma arquitetura o Softmax foi o melhor classificador. Este fato reforça a ideia deste trabalho em utilizar a RNC como extrator de características, e também, em testar outros classificadores na última camada da RNC. Todas as arquiteturas foram superiores com as imagens no tamanho de 185 *pixels*, com exceção da InceptionV3 que o tamanho 210 *pixels* atingiu melhor resultado.

A melhor combinação no geral foi a VGG19 recebendo na entrada imagens de 185 *pixels* com SVM. Esta combinação atingiu a melhor acurácia com 78.95% e média de 74% de f1-score. As duas emoções que tiveram os melhores resultados foram alegria e surpresa com 91% e 85% de f1-score, respectivamente. Em contrapartida, as piores emoções foram medo e raiva com 58% e 67% de f1-score nesta ordem. Coincidentemente, alegria e surpresa são duas das três emoções com maior taxa de amostragem, enquanto que raiva e medo estão entre as três menores amostragem (veja Tabela 10). Deve-se salientar que a VGG19 foi a arquitetura com maior quantidade de parâmetros, consequentemente, a que produziu o modelo mais pesado chegando à 1.3GB (consultar Tabela 12). Portanto, sua utilização exige mais recursos computacionais do que as demais.

A pior combinação no geral foi a MobileNetV2. Esta arquitetura juntamente com o Random Forest alcançou 73.39% de acurácia chegando a ser inferior cerca de 5.56% à VGG19 com SVM. Esse resultado é esperado, pois o foco da MobileNet é a utilização eficiente dos recursos. O que ocorreu neste estudo, pois a MobileNetV2 possui somente 0.25% dos parâmetros total da VGG19. O modelo da MobileNetV2 exige somente 3MB para armazenamento, enquanto que a VGG19 necessita de 1.3GB. Ou seja, a MobileNet chega à ser 401 vezes menor do que a VGG19. Analisando por este ponto de vista, a MobileNetV2 é uma rede poderosa e cumpre a premissa em ser eficiente. Embora que a utilização da MobileNetV2 perca em 6.5% de acurácia, em contrapartida, ganha-se na eficiência dos recursos. Portanto, a MobileNetV2 não deve ser descartada e seu uso é adequado para sistemas embarcados que possui recursos limitados.

Tabela 15 – Resultados experimentais do melhor classificador por arquitetura avaliando a base de validação geral

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
InceptionResNetV2 (Entrada = 185) + Random Forest (D = 27)	Raiva	0.68	0.63	0.65	0.7787
	Desgosto	0.87	0.67	0.76	
	Medo	0.64	0.49	0.56	
	Alegria	0.89	0.92	0.91	
	Tristeza	0.64	0.69	0.67	
	Surpresa	0.87	0.85	0.86	
	Neutralidade	0.66	0.75	0.7	
	Média/Total	0.75	0.72	0.73	
InceptionV3 (Entrada = 210) + Random Forest (D = 27)	Raiva	0.68	0.66	0.67	0.7868
	Desgosto	0.86	0.72	0.78	
	Medo	0.65	0.52	0.58	
	Alegria	0.9	0.92	0.91	
	Tristeza	0.65	0.7	0.67	
	Surpresa	0.87	0.86	0.86	
	Neutralidade	0.67	0.75	0.71	
	Média/Total	0.76	0.73	0.74	
ResNet50 (Entrada = 185) + kNN (k = 15)	Raiva	0.61	0.6	0.6	0.7526
	Desgosto	0.84	0.67	0.75	
	Medo	0.56	0.5	0.53	
	Alegria	0.88	0.9	0.89	
	Tristeza	0.63	0.63	0.63	
	Surpresa	0.85	0.83	0.84	
	Neutralidade	0.64	0.72	0.67	
	Média/Total	0.72	0.69	0.70	
VGG19 (Entrada = 185) + SVM (C = 0.007)	**Raiva**	**0.67**	**0.67**	**0.67**	**0.7886**
	Desgosto	**0.82**	**0.74**	**0.78**	
	Medo	**0.63**	**0.55**	**0.58**	
	Alegria	**0.9**	**0.92**	**0.91**	
	Tristeza	**0.69**	**0.68**	**0.68**	
	Surpresa	**0.86**	**0.85**	**0.85**	
	Neutralidade	**0.7**	**0.74**	**0.72**	
	Média/Total	**0.75**	**0.74**	**0.74**	
MobileNetV2 (Entrada = 185) + Random Forest (D = 27)	Raiva	0.58	0.55	0.56	0.7239
	Desgosto	0.8	0.63	0.7	
	Medo	0.54	0.42	0.47	
	Alegria	0.84	0.9	0.87	
	Tristeza	0.57	0.57	0.57	
	Surpresa	0.83	0.83	0.83	
	Neutralidade	0.62	0.66	0.64	
	Média/Total	0.68	0.65	0.66	

Entrada = Tamanho de *pixels*

4.5 Integração

A abordagem proposta consiste nos seguintes componentes: coleta de dados, detecção de face, extração de características e classificação. A coleta de dados monitora o indivíduo capturando imagens continuamente. Desta forma, gera-se um *buffer* de imagens a serem processadas. A detecção de face recebe uma imagem e computa a quantidade e a localização (coordenadas) das faces existentes. O método a ser utilizado é o MMOD-CNN. Depois, é executado um processo de recorte da face caracterizando-se como uma etapa de pré-processamento. A face recortada é enviada para um verificador de posicionamento da face. Esta etapa analisa se o posicionamento da face está totalmente dentro da imagem, se positivo, a face é enviada para RNC extrair as características, caso não esteja o processo é encerrado. Quando dispõe de muitos recursos a RNC a ser utilizada é a VGG19, já o contrário quando dispõe de poucos recursos e, pretende-se usar a abordagem nestes sistemas a opção é a MobileNetV2. A RNC tem sua maneira particular, também chamada de *black box*, para extrair as características, em que a imagem da face atravessa milhares, no caso da MobileNetV2, ou milhões de parâmetros, quando VGG19, distribuída em dezenas ou centenas de camadas. A extração de características resulta em um vetor unidimensional com tamanho bastante inferior da imagem original, cerca de 4% do tamanho total da imagem quando VGG19 e 1.25% para MobileNetV2, para o classificador analisar. Quando o extrator de características for a VGG19 o classificador é o SVM, entretanto, quando for a MobileNetV2 a escolha é o Random Forest. Após averiguação do classificador, tem-se a decisão que é mapeada para um vetor de probabilidades distribuída para cada emoção, onde a probabilidade mais alta é a emoção detectada nesta imagem. Portanto, o método proposto retorna as coordenadas das faces localizadas com a emoção correspondente para a aplicação consumidora. Desta forma, é possível saber de qual face se trata e qual foi a emoção detectada. Este processo é ilustrado na Figura 20. Para concluir, é importante salientar que em uma imagem com múltiplas faces o processo consiste em classificar uma por vez, pois é mais simples reconhecer a emoção de uma única face do que em todas ao mesmo tempo.

4.6 Resumo

A abordagem proposta foi descrita neste capítulo. As principais etapas são: coleta de dados, detecção da face, extração de características e classificação. Um estudo experimental acerca dos seguintes detectores de faces foi realizado: MMOD-CNN, MTCNN, HoG-SVM e Viola Jones. O estudo experimental incluiu analisar imagens de faces e não faces. O MMOD-CNN seguido do MTCNN foram as melhores técnicas de detecção facial. Em contrapartida, o Viola Jones foi o pior e não deve ser usado em nenhum caso neste contexto. Um componente importante da abordagem é a extração de características que é

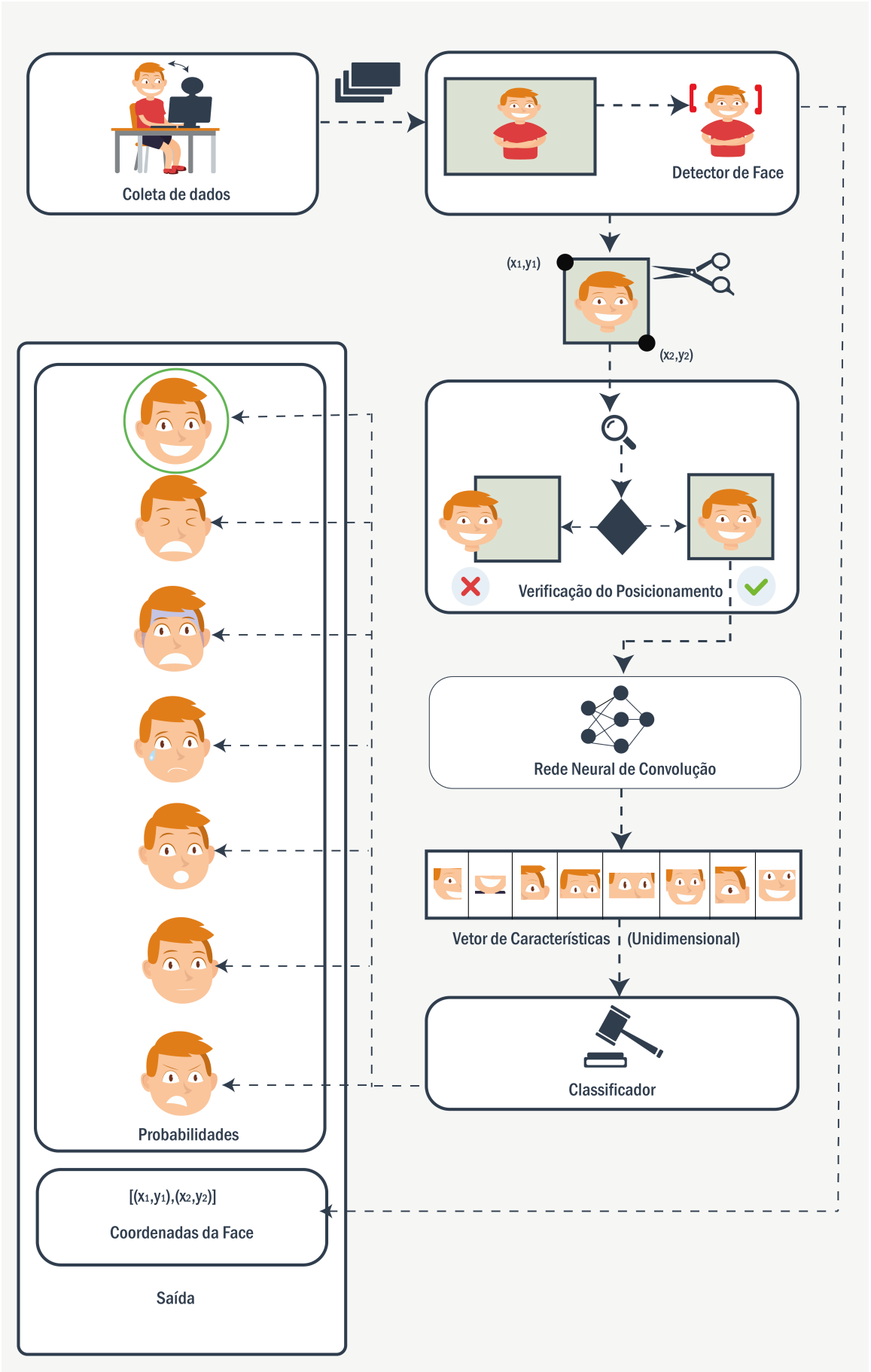

Figura 20 – Abordagem Proposta

baseada em RNCs. Cinco arquiteturas de RNCs foram treinadas utilizando vários recursos que maximizaram a aprendizagem do modelo. As arquiteturas foram: InceptionResNetV2, InceptionV3, MobileNetV2, ResNet50 e VGG19. A VGG19 foi a arquitetura que gerou o modelo mais pesado. Enquanto que a MobileNet50 mostrou-se 401 vezes menor em quantidade de parâmetros e tamanho de armazenamento do que a VGG19, caracterizando-se como uma opção para sistemas embarcado. Por fim, um estudo experimental foi realizado para analisar quatro classificadores na última camada de cada arquitetura de RNC. Os classificadores foram: Softmax, SVM, Random Forest e KNN. A melhor combinação entre arquitetura e classificador foi da VGG19 com o SVM na última camada alcançando 78.86% de acurácia. Enquanto que a MobileNetV2 com Random Forest chegou à 72.39%. Este estudo avaliou uma composição de 7 bases de dados que contém imagens de expressões faciais.

5 Resultados

Neste capítulo os resultados são apresentados divididos nas seguintes seções. A Seção 5.1 apresenta um estudo comparativo entre a abordagem proposta que será denominada como *Single-Shot Facial Expression Recognition* (SSFER) contra a API de reconhecimento de emoções da *Microsoft Cognitive Services* (MCS). A Seção 5.2 destaca o estudo de caso baseado na coleta de estados emocionais de estudantes correlacionando com seu desempenho no teste, e por fim, a Seção 5.3 um resumo do capítulo.

5.1 Estudo Comparativo: Abordagem Proposta (SSFER) e *Microsoft Cognitive Services* (MCS)

É importante comparar o SSFER com outro método de reconhecimento de emoção. Entretanto, comparar o SSFER com a literatura, observando somente a métrica de acurácia nas bases de dados, configura-se como uma comparação injusta. Pois, há pouca informação sobre como foi o treinamento e a validação, abrindo a possibilidade de um super vício (*overfitting*) nas bases utilizadas atingindo alta acurácia. Desta forma, quando o método receber imagens de um cenário real pode não funcionar tão bem e a super acurácia alcançada em testes, não se traduz como verdade nestas condições. O objetivo deste trabalho é propor uma abordagem que alcance a maior generalização possível aumentando as chances de funcionar corretamente em cenário real. Portanto, o SSFER não deve ser viciado nas bases de dados da literatura.

Sendo assim, para executar uma comparação mais justa foi necessário escolher um método de reconhecimento de emoção e processar a mesma base de validação usada na Seção 4.3 para experimentar o SSFER. Atualmente, há vários *Software as a Service* (SaaS) disponívieis para reconhecer emoção. Para este estudo comparativo foi selecionada a API da *Microsoft Cognitives Services* (MCS)[1]. Esta API é um serviço na nuvem que oferece um conjunto de aplicações cognitivas, inclusive um reconhecedor de emoções via expressão facial. A escolha pela Microsoft foi por esta gozar de prestígio, por ser uma gigante no mercado mundial de SaaS, e uma das pioneiras em inteligência artificial.

A Tabela 16 apresenta os resultados experimentais da MCS e SSFER avaliando a base de validação geral. A emoção que a MCS foi melhor avaliando pela f1-score, que é a métrica harmônica entre precisão e revocação, foi alegria e surpresa. Enquanto que medo e desgosto foram as piores em f1-score. Coincidentemente, a abordagem proposta por este trabalho, o SSFER, também teve como as melhores emoções alegria e surpresa e,

[1] Disponível em: https://bit.ly/2xcqd3N

Tabela 16 – Resultados experimentais da *Microsoft Cognitive Services*(MCS) e abordagem proposta (SSFER) avaliando a base de validação geral

Métodos	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
MCS Microsoft Cognitive Services	Raiva	0.72	0.44	0.55	0.6912
	Desgosto	0.84	0.28	0.42	
	Medo	0.87	0.20	0.32	
	Alegria	0.81	0.96	0.88	
	Tristeza	0.68	0.45	0.54	
	Surpresa	0.82	0.66	0.74	
	Neutralidade	0.46	0.84	0.59	
	Média/Total	0.74	0.55	0.58	
SSFER **VGG19 + SVM**	**Raiva**	**0.67**	**0.67**	**0.67**	**0.7886**
	Desgosto	**0.82**	**0.74**	**0.78**	
	Medo	**0.63**	**0.55**	**0.58**	
	Alegria	**0.9**	**0.92**	**0.91**	
	Tristeza	**0.69**	**0.68**	**0.68**	
	Surpresa	**0.86**	**0.85**	**0.85**	
	Neutralidade	**0.7**	**0.74**	**0.72**	
	Média/Total	**0.75**	**0.74**	**0.74**	

em contrapartida, medo e raiva como as piores. Porém, o SSFER para todas as emoções analisando pelo f1-score atingiu resultados superiores ao MCS. Por exemplo, o SSFER em f1-score para alegria e surpresa alcançou 0.91 e 0.85, respectivamente. Enquanto que o MCS, teve para alegria 0.88 e surpresa com 0.74 de f1-score. Analisando a acurácia geral, o SSFER alcançou 0.7886, enquanto que o MCS apenas 0.6912. Totalizando uma diferença significativa de 0.0974. Em outras palavras, quando traduzida para porcentagem, podemos afirmar que o SSFER foi superior em 9.74% de acurácia em relação ao MCS na base de validação geral. Na MCS, os valores de precisão e revocação são sempre muito diferentes, de onde é possível concluir que a MCS não apresenta um equilíbrio entre precisão e revocação quando comparada ao SSFER.

Vale ressaltar que a base de validação geral usada no estudo experimental é a composição de várias bases da literatura (consultar Tabela 9) e possui imagens com alto grau de variabilidade dos dados ilustrada pela Figura 21. Há imagens coloridas, em escala de cinza, com alta resolução e outras que apresentam *pixelização* ou desfoque. Essa variabilidade dos dados mostra que a base de validação geral não é uma base fácil para ser classificada. A matriz de confusão dos resultados experimentais na base de validação geral é apresentada na Tabela 17. A matriz de confusão tanto de SSFER como da MCS ajuda a compreender a diferença significativa de 9.74% entre os métodos na métrica de acurácia. É notório que o MCS apresenta um vício (*overfitting*) nas emoções de neutralidade e alegria, especialmente, na neutralidade. Enquanto que o SSFER os erros já estão distribuídos de uma forma mais uniforme. A diagonal principal da matriz de confusão representa os acertos do método. O SSFER computou 13927 acertos, enquanto que MCS alcançou 11430

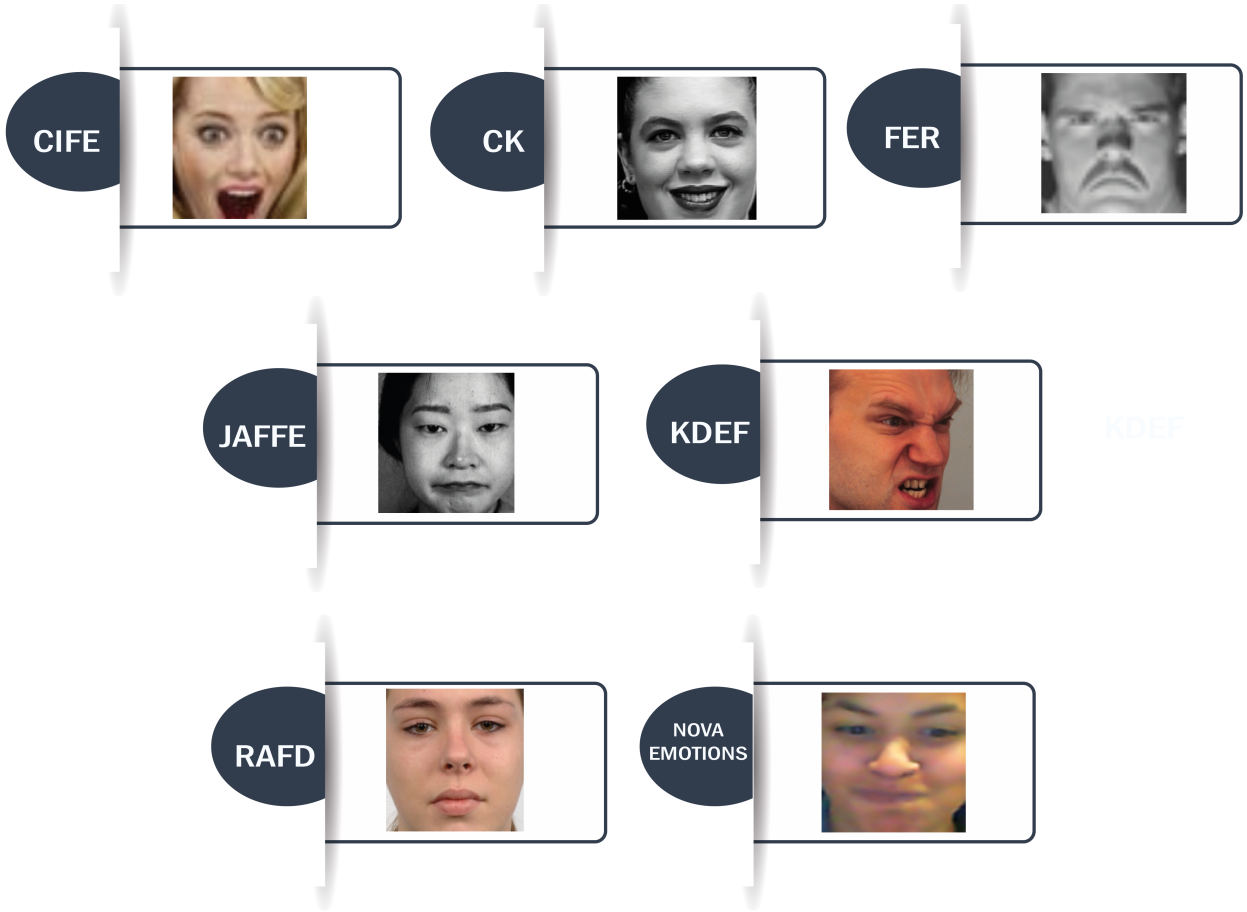

Figura 21 – Representação da Base de Validação Geral

acertos, do total de 17642 imagens. Desta forma, o SSFER tem um saldo de acertos em 2497 imagens. Analisando as demais colunas da matriz de confusão, podemos concluir que o comportamento da MCS é mais previsível, pois seus erros são tendenciosos para neutralidade e alegria. Em contrapartida, o SSFER os erros são imprevisíveis podendo ser para qualquer emoção.

5.2 Estudo de Caso: Coleta de Estados Emocionais de Estudantes

Um estudo de caso aplicado na área de educação foi realizado. O estudo de caso consistiu na realização de um simulado do Exame Nacional do Ensino Médio em uma plataforma educacional. O objetivo foi coletar imagens das expressões faciais dos estudantes e toda a interação e cliques com a plataforma foram armazenadas.

A concepção do estudo de caso foi estimulado pelos resultados da revisão sistemática (ver Capítulo A) em que mostrou uma carência de trabalhos de reconhecimento de emoção em cenários reais. Desta forma, a utilização em cenários reais passa a ser um ponto de contribuição, atestando que essas tecnologias estão amadurecidas. Além disso,

Tabela 17 – Matriz de Confusão da *Microsoft Cognitive Services* (MCS) e abordagem proposta (SSFER) avaliando a base de validação geral

Método	**Matriz de Confusão**							
		Raiva	Desgosto	Medo	Alegria	Tristeza	Surpresa	Neutralidade
MCS	Raiva	620	25	5	48	71	43	599
	Desgosto	106	273	3	169	76	3	361
	Medo	89	9	251	61	184	256	415
	Alegria	1	2	2	5392	10	28	176
	Tristeza	24	6	2	58	857	19	947
	Surpresa	9	2	25	509	13	1681	290
	Neutralidade	11	7	0	381	47	13	2356
SSFER		Raiva	Desgosto	Medo	Alegria	Tristeza	Surpresa	Neutralidade
	Raiva	1043	39	94	56	149	32	140
	Desgosto	53	763	17	45	47	22	92
	Medo	137	17	775	51	166	153	129
	Alegria	24	21	38	5433	56	98	208
	Tristeza	146	33	123	79	1453	16	291
	Surpresa	32	3	123	137	21	2262	82
	Neutralidade	124	42	65	243	224	47	2198

evidenciar que as tecnologias de reconhecimento de emoção é útil para integrar aos outros sistemas, principalmente para tomadas de decisão, recomendações, análises de comportamento, dentre outros. Contudo, uma abordagem de reconhecimento de emoção alcançar acurácia em cenário real, não é uma tarefa trivial, visto que o mundo real é desafiador com alta variabilidade do ambiente e dos indivíduos.

5.2.1 Metodologia Experimental

Um experimento foi realizado com 27 alunos do Ensino Médio de uma escola de tempo integral que, na época, estavam se preparando para o Exame Nacional do Ensino Médio (ENEM). O experimento consistiu em um simulado do exame contendo 40 questões de múltipla escolha.

A coleta de dados se deu a partir de uma plataforma educacional que executou o questionário de múltipla escolha em um *tablet* ou *notebook*. A plataforma digital selecionada foi a proposta por Leitão (2017). Durante a interação com a plataforma foram armazenados todos os cliques efetuados pelo estudante. Além disso, a cada intervalo entre 2 a 5 segundos capturava-se uma foto via câmera frontal do dispositivo. Os assuntos escolhidos foram: matemática, língua portuguesa, química, raciocínio lógico, geografia e história. O simulado teve duração de duas horas e cada questão possuía dois níveis de dificuldade (fácil ou difícil), além de ter cinco respostas alternativas.

A seleção dos estudantes para participar do experimento ocorreu voluntariamente.

O grupo final foi heterogêneo, onde 53% consideravam até o momento seu desempenho na escola como bom ou ótimo e, 47% como regular ou ruim; além disso, 30% deles consideravam a sua preparação para o vestibular como boa ou ótima e, 70% como regular ou fraca. Os alunos selecionados foram de 5 turmas do ensino médio do segundo e terceiro ano.

Neste estudo de caso, as imagens coletadas foram processadas pelo SSFER e MCS. Sendo assim, as emoções detectadas dos estudantes durante a execução do simulado são comparadas entre ambas abordagens de reconhecimento de emoção. Vale ressaltar que, foi assinado um termo de consentimento em que concordamos em não disponibilizar ou publicar qualquer imagem dos estudantes.

5.2.2 Preparação dos Dados

O módulo de coleta de imagens pela câmera frontal para a plataforma educacional foi desenvolvida sob medida para este experimento. Desta forma, enquanto os 27 estudantes estavam sendo monitorados houve algumas perdas de dados e mal uso da plataforma, pois ocorreu sobrecarregamento da rede. Uma limpeza de dados foi executada para excluir os dados defeituosos. Após a operação de limpeza, sobraram os dados confiáveis de 25 estudantes para as análises. O conjunto coletado totalizou 31038 imagens e 3920 de cliques. O conjunto de 31038 imagens para as próximas seções será referenciado como base do cenário real.

5.2.3 Emoções Detectadas

O primeiro passo foi processar as 31038 imagens coletadas pelo SSFER e MCS para detectar as emoções. Das 31038 imagens da base do cenário real, o MCS detectou face em 23197 imagens, enquanto que o SSFER detectou face em 26762 imagens. Portanto, também, no quesito de detecção de face, o SSFER foi superior à MCS. Vale destacar, que é compreensível o alto número de imagens com faces não detectadas, pois os estudantes quando estão rabiscando papel deixam de olhar para a plataforma ficando cabisbaixo e esta posição realmente não há face para detectar. Diante deste contexto, foi executada uma limpeza de dados para uma comparação justa, em que as imagens válidas para o estudo foram as que ambos os métodos conseguiram detectar face. Após a operação de limpeza de dados, restaram o conjunto de 22846 imagens na base do cenário real.

A Tabela 18 denota as detecções de emoções na base do cenário real. É importante ressaltar que as imagens do cenário real não estão rotuladas, então não é possível calcular as métricas de acurácia, precisão, revocação e f1-score. A MCS detectou a quantidade de 0, 15 e 23 para medo, raiva e desgosto, respectivamente. Logo, a detecção deste sub conjunto de emoções obteve uma baixa representatividade. Em sequência, temos 355 amostras para

Tabela 18 – Detecção de estados emocionais na base do cenário real

Método	**Limiar**	**Emoção**						
		Raiva	Desgosto	Medo	Alegria	Tristeza	Surpresa	Neutralidade
MCS	0.0	15	23	0	355	399	133	21921
SSFER	0.0	1092	543	330	1117	2014	335	17415
SSFER	<30%	932	437	221	977	1806	263	18210
SSFER	<40%	631	191	103	676	1070	155	20020
SSFER	<50%	402	43	35	449	418	96	21403

alegria e 399 para tristeza. Para a MCS, que é um método viciado em neutralidade, 95.95% das imagens dos estudantes são de neutralidade. Isso significa que somente em 4.05% das imagens houveram uma emoção diferente da neutralidade, logo, um número baixo. A alta amostragem de neutralidade em ambiente educacionais faz sentido. O trabalho de D'Mello e Calvo (2013) diz que emoções não-básicas como: engajamento, tédio, confusão e frustração, ocorrem até cinco vezes mais do que as emoções básicas. Embora vale destacar que as evidências literárias apontam o reconhecimento por expressão facial somente para o grupo das emoções básicas e neutralidade. Sendo assim, de forma autônoma não é possível reconhecer o grupo das emoções não-básicas somente pela análise das expressões faciais.

A abordagem proposta SSFER mostrou-se adaptável durante os experimentos. Motivada pela alta quantidade de neutralidade encontrada pela MCS foi adicionado um recurso chamado limiar. Lembrando que a saída do SSFER é um vetor de probabilidades e a princípio a emoção detectada é a que possui maior probabilidade. Há casos em que a maior probabilidade não é tão alta assim, ou seja, caracterizando um caso em que o SSFER não tem tanta certeza da emoção. Portanto, o limiar funciona atacando os casos quando o SSFER não tem tanta certeza atribuindo para o estado de neutralidade. Quando a maior probabilidade do vetor for menor que o limiar é atribuído o estado de neutralidade. Pois, durante a troca de estados emocionais a expressão de neutralidade antecede em algum instante de tempo as emoções básicas (Ekman, 1999).

O SSFER com limiar igual a 0 detectou mais diferentes emoções do que o MCS. Para 76.22% das imagens coletadas foram expressadas neutralidade. As emoções de tristeza, alegria e raiva tiveram destaque com uma amostragem de 2014, 1117 e 1092, respectivamente. A medida que o limiar aumenta, o SSFER passa a detectar mais neutralidade reduzindo a sensibilidade do método para reconhecer as demais emoções. O limiar de 50% aplicado faz a taxa de detecção de neutralidade subir para 93.68% das imagens. E tristeza, alegria e raiva reduz drasticamente sua amostragem de detecção para 418, 449 e 402, respectivamente. Com limiar em 50% a alegria torna-se a segunda maior amostragem. Enquanto que quando o limiar é igual à 0, tristeza é a segunda maior amostragem. Em ambos os casos a neutralidade prevalece como a maior amostragem.

D'Mello e Calvo (2013) afirmou que não vale tanto a pena detectar emoção básica, pois as emoções não-básicas chegam a ser até 5 vezes mais frequentes. Este trabalho contra-argumenta que depende do caso, por exemplo, neste estudo de caso, um cenário de educação, das 96 imagens detectadas como surpresa pelo SSFER, 37 imagens apresentaram bocejos que segundo a medicina é um indicativo de sono. Detectar que o estudante manifesta sono durante atividade educacional é bastante relevante para outros sistemas inteligentes atuar ou adaptar conteúdo para acordar o estudante e retomar o nível de engajamento. Desta forma, este trabalho reuniu evidências que é possível detectar bocejo por meio da surpresa. Das 96 imagens de surpresa, houveram dois estudantes mastigando chiclete e quando abria-se de forma contundente a boca, também detectou-se como a expressão facial de surpresa, isto ocorreu em 45 imagens. Além da surpresa, a expressão de alegria funciona bem para o cenário de educação, em que na maioria dos casos os estudantes expressavam sorrisos com alto grau de abertura, caracterizando de fato alegria e, pode ser um estado interessante caso haja um interesse em detectar se um determinado conteúdo consegue engajar o aluno. As expressões faciais detectadas como tristeza e raiva apresentavam os músculos com bastante movimentação e que, portanto, não é possível que essas informações sejam desprezíveis nos estudos comportamentais e da computação afetiva.

5.2.4 Concordância entre a Abordagem Proposta (SSFER) e Microsoft Cognitive Services (MCS)

Uma informação interessante a ser derivada a partir dos dados da Tabela 18 é o grau de concordância entre o SSFER e a MCS. Este dado é interessante porque, como já mencionado anteriormente, os dados da base do cenário real não estão rotulados, desta forma, não é possível saber quais das duas abordagens tem maior quantidade de acertos.

A Tabela 19 apresenta um estudo referente à concordância entre SSFER e a MCS. A coluna da Emoção mostra as contagens de quando ambos os métodos concordaram. Em nenhum caso foi concordado a emoção medo, o que é óbvio pois o MCS não detectou medo em nenhuma imagem. A emoção de desgosto foi concordado para uma única imagem. As expressões faciais de neutralidade, alegria e surpresa foram as que houveram maior concordância. Em um contexto geral, a medida que vai subindo o limiar de probabilidades no SSFER, as duas abordagens vão concordando mais na expressão de neutralidade e a proporção geral de concordância sobe de 76.79% até alcançar 92.07%. Embora haja um decréscimo de concordância nas demais emoções.

O coeficiente de Kappa (Cohen, 1960) também foi calculado na Tabela 19. Esta métrica é especializada em medir concordância. O Kappa varia de -1 a 1 e quando o número é positivo significa que há concordância. Entre 0 e 0.20 a concordância é denominada como fraca, entre 0.21 e 0.40 é julgada como moderada. O melhor caso de concordância baseado

Tabela 19 – Concordância entre a Abordagem Proposta (SSFER) e a Microsoft Cognitive Service (MCS) na base de cenário real

Limiar	**Emoção**							**Total**	**Concordância**	**Coeficiente Kappa**
	RAI	DES	MED	ALE	TRI	SUR	NEU			
0.0	11	1	0	251	96	71	17091	17521	76.69%	0.12
<30%	11	1	0	243	87	62	17828	18232	79.80%	0.13
<40%	8	1	0	215	56	46	19499	19825	86.78%	0.16
<50%	8	1	0	190	31	33	20771	21034	92.07%	0.21

no coeficiente de Kappa entre o SSFER e a MCS foi quando o limiar é menor que 50%. Assim, o coeficiente atinge 0.21, embora em 92.07% das imagens há concordância. Abre-se o questionamento, pois para 92.07% dos casos há concordância, por que o Kappa foi tão baixo? A explicação é de que há pouca concordância nas outras emoções e há muita concordância no estado de neutralidade. Essa baixa concordância nas emoções levou o Kappa a ter um valor baixo.

5.2.5 Correlação das emoções detectadas com o desempenho no teste

A Tabela 20 apresenta os dados de proporção das emoções detectadas pela nota tradicional (score) no teste. Em geral, os alunos foram ruim no teste. Apenas 5 estudantes conseguiram nota acima de 6, enquanto que 9 estudantes receberam nota menor que 4.

Analisando pela API da MCS o grupo que atingiu a maior nota (acima de 6), isto é, um total de 5 alunos, tiveram a maior taxa de neutralidade do que os demais, totalizando a proporção de 97.5% de neutralidade. Uma hipótese a ser levantada é que esse grupo teve maior nível de concentração durante o teste. Visto que a neutralidade pode ser associada ao estado de concentração (Cruz et al., 2017). O contrário também é válido, pois o grupo que atingiu a menor nota foi o que teve menor proporção de neutralidade com 93.7% e, também, a maior taxa de surpresa e a segunda maior taxa de tristeza. Lembrando que a surpresa pode ser associada ao bocejo, consequentemente, ao estado de sono.

Em contrapartida, as hipóteses levantadas pelo MCS não são confirmadas pelo SSFER. Com exceção da hipótese da emoção surpresa, em que o grupo que atingiu a melhor nota, isto é, acima de 6 no teste, teve 0.1% de proporção total das imagens foram de surpresa, o menor valor de todos os grupos. Isto significa que tal grupo foi o que bocejou menos, sendo assim, manifestou menos sonolência. O grupo com maior nota teve a menor taxa de neutralidade, desta forma, foi o grupo que mais manifestou as outras emoções. Uma hipótese para este cenário são que estudantes com maiores proporções de acertos, tem um engajamento maior com o teste, e durante o teste acertando ou errando as questões transmitem emoções.

Tabela 20 – Proporção da emoção detectada por nota tradicional (score) no simulado

Método	**Nota**	**Emoção**							**Total Alunos**
		RAI	DES	MED	ALE	TRI	SUR	NEU	
MCS	<3	0.0	0.1	0.0	2.0	2.80	1.5	93.7	3
	<4	0.0	0.0	0.0	2.4	3.30	0.2	94.0	6
	<5	0.1	0.0	0.0	1.09	1.3	0.3	97.3	4
	<6	0.1	0.2	0.0	1.5	0.70	0.70	96.7	9
	>6	0.1	0.0	0.0	0.8	1.3	0.2	97.5	5
SSFER	<3	0.0	0.0	0.0	4.10	4.2	0.6	91.0	3
	<4	0.70	0.0	0.1	2.0	1.9	0.5	94.8	6
	<5	0.4	0.1	0.2	0.8	1.09	0.3	97.1	4
	<6	0.4	0.1	0.2	1.3	1.3	0.5	96.3	9
	>6	6.9	0.70	0.2	2.19	1.40	0.1	88.5	5

5.3 Resumo

Os resultados foram discutidos a partir de duas visões. A primeira foi comparando a Abordagem Proposta (SSFER) contra um serviço em nuvem a Microsoft Cognitive Services (MCS) na base de validação geral. O SSFER ganhou da MCS em 9.74% de acurácia. A MCS mostrou-se ser um método com vício nos estados de neutralidade e de alegria. Além disso, até mesmo na detecção de face o SSFER mostrou ser mais eficaz do que a MCS. A segunda etapa dos resultados foi por intermédio de um estudo de caso executado em um cenário educacional. Por meio de uma plataforma digital, reuniu-se 27 estudantes para realizar um simulado do ENEM e, enquanto respondia o teste suas expressões faciais foram capturadas, assim como, todos os cliques. Depois, houve uma análise offline das 31 mil expressões capturadas por meio do SSFER e MCS. Verificou-se que a neutralidade foi o estado mais transmitido pelos estudante durante os testes e que as outras emoções ocorreram em casos raros. Entretanto, constatou-se que a surpresa pode ser relacionado ao bocejo, movimento natural de indicativo de sonolência. Posteriormente, foi medido o grau de concordância entre o SSFER e o MCS na base coletada das expressões faciais dos estudantes. Ambos os métodos concordaram em 92.07% das imagens. Por fim, houve um cruzamento das emoções pela nota alcançada no teste. Conclui-se que ambos os métodos não apresentaram as mesmas hipóteses entre as emoções e o desempenho da nota tradicional. Com exceção da emoção surpresa, em que ambos os métodos apontaram que os estudantes com notas acima de 6 foram os que transmitiram menores taxas de surpresa, isto é, tiveram menores ocorrências de bocejo, isto é, manifestação de sono.

6 Considerações Finais

Esta dissertação apresentou uma abordagem para reconhecer emoção por meio da expressão facial utilizando redes neurais de convolução (RNC). O diferencial desta abordagem é reunir os principais elementos identificados na literatura para reconhecer emoções. Além disso, este trabalho encontrou uma solução para reconhecer emoções tanto em nuvem (VGG19) como na computação embarcada (MobileNet).

Inicialmente, este trabalho experimentou diversos detectores de faces. Em que o vencedor também é uma RNC: o MMOD-CNN. Este método de detecção de face demonstrou ser mais eficaz que o da Microsoft Cognitive Service (MCS). Depois, treinou-se cinco arquiteturas de RNCs alternando quatro classificadores na última camada com duas entradas diferentes de dados. A melhor combinação foi a VGG19 com SVM. Podemos concluir que quando há muitos recursos disponíveis usar a VGG19 juntamente com a SVM é a melhor combinação. Em contrapartida, quando deseja-se um método com menor consumo de recursos computacionais, a combinação ideal é a MobileNet com Random Forest. A MobileNet chega a ser 403 vezes menor do que a VGG19, que corresponde a uma economia relevante de memória e processamento. A abordagem proposta foi denominada como *Single-Shot Facial Expression Recognition* (SSFER).

O SSFER foi avaliado experimentalmente e alcançou 78.86% de acurácia. Foi necessário a comparação do SSFER contra um outro método de reconhecimento de emoção a Microsoft Cognitive Services (MCS). O SSFER foi superior ao MCS em cerca de 9.74%. A MCS demonstrou sua deficiência em possuir vício no estado de neutralidade e emoção alegria. As emoções de alegria e surpresa foram as que obtiveram melhor precisão. Entretanto, as emoções de medo e raiva foram as que tiveram menores taxas de precisão.

Por fim, um estudo de caso em um cenário real foi executado. Reuniu-se 27 estudantes para um simulado da prova do Exame Nacional do Ensino Médio (ENEM). A prova ocorreu por meio de uma plataforma educacional digital. Os estudantes usavam um *tablet* ou *notebook* para responder as questões e, enquanto interagia com a plataforma, suas expressões faciais foram coletadas. Identificou-se que a neutralidade foi o estado mais frequente em cerca de 95% das imagens. Constatou-se que a emoção de surpresa pode ser relacionada ao bocejo, e assim, viabilizar a detecção de sonolência. Houve destaque na emoção de alegria, pois quando detectada, de fato, o estudante estava com um sorriso largo. Enquanto que em raiva e desgosto os estudantes apresentavam músculos faciais com fortes movimentações.

Verificou-se também que os estudantes que obtiveram as melhores notas, tanto pela MCS quanto pelo SSFER, eram o grupo que menos transmitiu a emoção surpresa,

desta forma, conclui-se o grupo com menor manifestação de sono. Além disso, um estudo sobre a concordância entre MCS e SSFER proporcionalmente alcançou 92.07% do total das imagens. Portanto, o SSFER demonstrou ser eficaz para reconhecer emoções básicas em cenário real de uso. Inclusive batendo a MCS em mais de 9% na base de validação geral.

Este trabalho tem como objetivo reconhecer emoções humanas por meio da expressão facial. Todavia, os trabalhos de Darwin (1965) e Ekman e Davidson (1994) apontaram que somente o grupo das emoções básicas (raiva, alegria, tristeza, desgosto, medo e surpresa) são transmitidas por meio da expressão facial. Portanto, este trabalho se limita a somente reconhecer as emoções básicas.

As pesquisas na área de reconhecimento de emoções usando expressões faciais ainda permanece com alguns desafios. Consequentemente, este tópico ainda tem oportunidades de melhoria, tais como:

- **Analisar sequência de imagens**: Atualmente, a abordagem analisa somente uma imagem para reconhecer as emoções. Desta forma, está sendo ignorada a característica temporal e apenas uma imagem é considerada para definir as emoções. Integrar uma sequência de imagens, na forma de uma série temporal, pode ser relevante para melhorar o reconhecimento, justamente por analisar um conjunto de imagens amostrada em uma fração de tempo para definir as emoções;
- **Técnicas de verificação da face**: Aplicar técnicas para verificação de face com intuito de detectar uma face com grande quantidade de ruído para ser descartada e não emitir qualquer classificação;
- **Avaliar em cenários de uso reais**: Coletar *big data* de expressões faciais e dados do ambiente para depois processar pelo SSFER e fazer análises de dados entre os dados do ambiente e das emoções detectadas, seja em educação ou em outros campos;
- **Integração com sistemas de recomendação**: Integrar o SSFER com sistemas de recomendação para prover adaptação de conteúdo durante a interação com a plataforma educacional;
- **Reconhecimento de emoção multimodal**: Reconhecer emoção através da combinação dos dados sensoriais, de voz, movimentação e interação com a plataforma, com a finalidade de identificar emoções secundária como: tédio, confusão, engajamento e timidez.

Referências

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

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

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

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

Boser, B. E., Guyon, I. M., e Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, páginas 144–152. ACM.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cruz, A., Leitão, G., Colonna, J., Silva, E., Barreto, R., e 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, página 997.

Dalal, N. e Triggs, B. (2005). Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, páginas 886–893. IEEE Computer Society.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., e Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, páginas 248–255. Ieee.

D'Mello, S. e Calvo, R. A. (2013). Beyond the basic emotions: what should affective computing compute? In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, páginas 2287–2294. ACM.

D'Mello, S., Kappas, A., e Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, 10(2):174–183.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Ekman, P. e Friesen, W. V. (1977). *Facial action coding system*. Consulting Psychologists Press, Stanford University, Palo Alto.

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

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., e Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. "O'Reilly Media, Inc.".

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

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

He, K., Zhang, X., Ren, S., e Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., e Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, Y. e Lu, H. (2016). Deep learning driven hypergraph representation for image-based emotion recognition. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, páginas 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. e 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. e Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243.

Jaques, P. A. e 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., e Ahn, C. (2015). Development of deep learning-based facial expression recognition system. In *Frontiers of Computer Vision (FCV), 2015 21st Korea-Japan Joint Workshop on*, páginas 1–4. IEEE.

Kim, B.-K., Dong, S.-Y., Roh, J., Kim, G., e Lee, S.-Y. (2016). 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*, páginas 48–57.

King, D. E. (2015). Max-margin object detection. *arXiv preprint arXiv:1502.00046*.

Kingma, D. P. e Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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

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

Leitão, G. d. S. (2017). *Uma plataforma de suporte ao docente no contexto da Educação Digital*. Universidade Federal do Amazonas.

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

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

Mafra, S. N. e 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., e Pai, M. M. (2016). Automatic facial expression recognition using dcnn. *Procedia Computer Science*, 93:453–461.

Nasoz, F., Alvarez, K., Lisetti, C. L., e 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., e 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*, páginas 443–449. ACM.

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

Robbins, H. e Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, páginas 400–407.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., e Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, páginas 4510–4520.

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

Shan, K., Guo, J., You, W., Lu, D., e 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*, páginas 123–128. IEEE.

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

Shojaeilangari, S., Yau, W.-Y., Li, J., e Teoh, E.-K. (2014). Multiscale analysis of local phase and local orientation for dynamic facial expression recognition. *Journal ISSN*, 1(1).

Simonyan, K. e 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., e Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 1–9.

Taigman, Y., Yang, M., Ranzato, M., e Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 1701–1708.

Tieleman, T. e Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Viola, P. e 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, páginas I–I. IEEE.

Vo, D. M. e 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*, páginas 80–84. IEEE.

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

Witten, I. H., Frank, E., Hall, M. A., e Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yu, X., Huang, J., Zhang, S., e Metaxas, D. N. (2015). 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., e Metaxas, D. N. (2014). Consensus of regression for occlusion-robust facial feature localization. In *European Conference on Computer Vision*, páginas 105–118. Springer.

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

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, K., Zhang, Z., Li, Z., e Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

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 e 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 21 – 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.

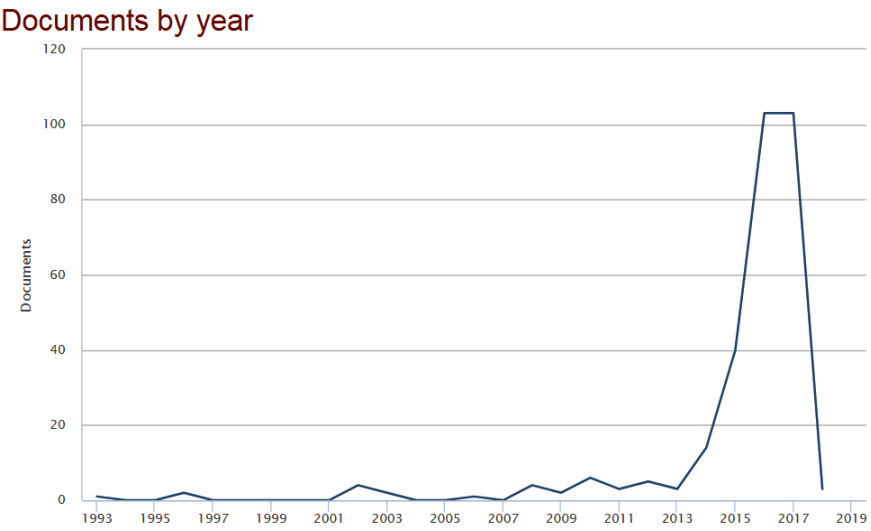

Figura 22 – Artigos por ano retornados pela *string* de busca

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.

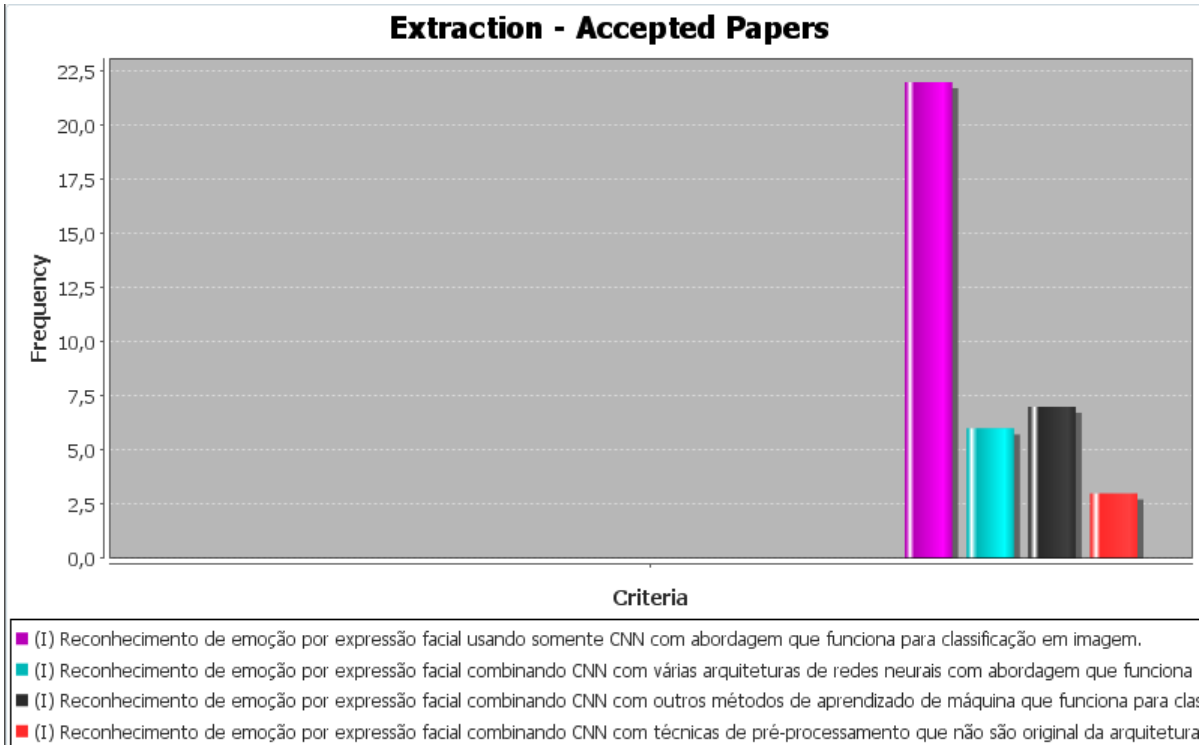

Figura 23 – Gráfico representando a frequência dos critérios de inclusão para os artigos aceitos do segundo filtro

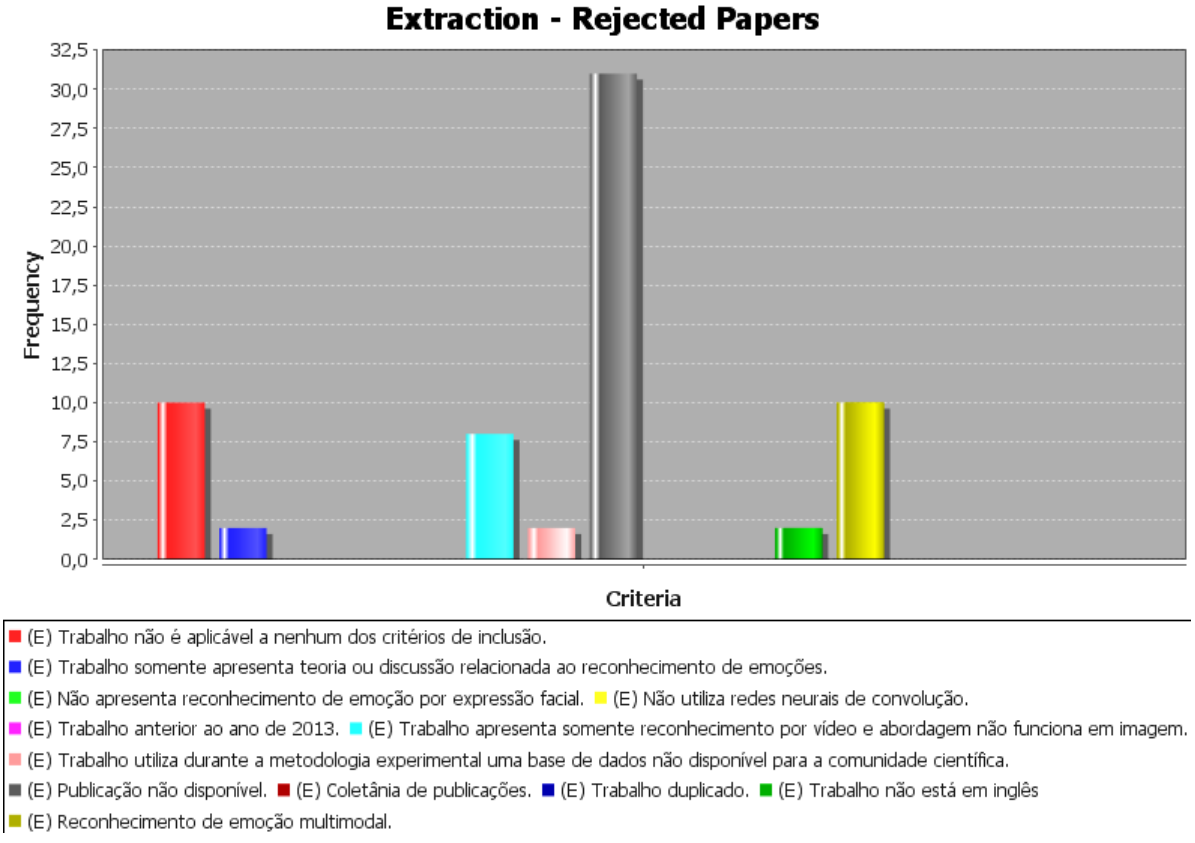

Figura 24 – Gráfico representando a frequência dos critérios de exclusão para os artigos rejeitados do segundo filtro

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, alegria, surpresa, tristeza, raiva, desgosto e medo. Alguns trabalhos como Barsoum et al. (2016) reconhece também nojo.

A.3.2 Q2: Quais tipos de pré-processamento têm sido aplicados nas imagens?

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., 2017; Li et al., 2015; Mayya et al., 2016; Ng et al., 2015; Shan et al., 2017; Shin et al., 2016; Vo e Le, 2016);
- **Normalização de Brilho (Equalização de histograma)**: Transformada para realce do contraste (Kim et al., 2016; Shan et al., 2017; Shin et al., 2016);
- **Normalização Min e Max**: Transformação linear baseada no valor mínimo e máximo da imagem (Kim et al., 2016);
- **Pontos da Face (pontos geométricos)**: Extração de pontos da face e a distância entre os pontos (Yu et al., 2016);
- **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., 2016);
- **Filtro de Difusão Isotópica**: (Shin et al., 2016);
- **Normalização DCT**: Transformada discreta do cosseno utilizada em compressão de dados e eventualmente evidenciando informações relevantes da imagem (Shin et al., 2016);
- **Alinhamento da Face**: Utilização de uma rede neural *autoencoder* para alinhar a face no centro (Kim et al., 2016).

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

Tabela 22 – 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 22 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 e Davidson (1994) que são: neutralidade, alegria, medo, desgosto, raiva, surpresa e tristeza, isto significa que

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

Tabela 23 – Bases de Dados

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

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., 2016; Chen et al., 2017; Liu et al., 2016; Wen et al., 2017);
- **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., 2017; Mayya et al., 2016; Wen et al., 2017);
- **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., 2015),;
- **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., 2015; Shin et al., 2016);

- **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 e Le, 2016; Yu et al., 2016).

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 **?** 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 normali-

zação 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.

ANEXO B – Avaliação experimental na base de validação geral em imagens quadradas de 185 *pixels*

Neste anexo são apresentados os resultados dos modelos avaliando a base de validação geral com imagens quadradas em 185 *pixels*.

B.1 InceptionResNetV2

Tabela 24 – Arquitetura InceptionResNetV2 avaliando a base de validação geral com imagens quadradas em 185 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
InceptionResNetV2 (Entrada = 185) + Softmax	Raiva	0.67	0.55	0.60	0.7292
	Desgosto	0.84	0.60	0.70	
	Medo	0.82	0.23	0.36	
	Alegria	0.94	0.84	0.89	
	Tristeza	0.46	0.82	0.59	
	Surpresa	0.72	0.91	0.81	
	Neutralidade	0.69	0.66	0.67	
	Média/Total				
InceptionResNetV2 (Entrada = 185) + kNN (k = 15)	Raiva	0.63	0.66	0.64	0.7737
	Desgosto	0.86	0.66	0.75	
	Medo	0.6	0.5	0.54	
	Alegria	0.9	0.91	0.91	
	Tristeza	0.67	0.68	0.68	
	Surpresa	0.86	0.84	0.85	
	Neutralidade	0.66	0.74	0.7	
	Média/Total				
InceptionResNetV2 (Entrada = 185) + Random Forest (D = 27)	**Raiva**	**0.68**	**0.63**	**0.65**	**0.7787**
	Desgosto	**0.87**	**0.67**	**0.76**	
	Medo	**0.64**	**0.49**	**0.56**	
	Alegria	**0.89**	**0.92**	**0.91**	
	Tristeza	**0.64**	**0.69**	**0.67**	
	Surpresa	**0.87**	**0.85**	**0.86**	
	Neutralidade	**0.66**	**0.75**	**0.7**	
	Média/Total				
InceptionResNetV2 (Entrada = 185) + SVM (C = 0.007)	Raiva	0.66	0.65	0.65	0.7783
	Desgosto	0.82	0.72	0.77	
	Medo	0.6	0.52	0.56	
	Alegria	0.9	0.92	0.91	
	Tristeza	0.66	0.68	0.67	
	Surpresa	0.85	0.85	0.85	
	Neutralidade	0.69	0.73	0.7	
	Média/Total				

B.2 InceptionV3

Tabela 25 – Arquitetura InceptionV3 avaliando a base de validação geral com imagens quadradas em 185 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
InceptionV3 (Entrada = 185) + Softmax	Raiva	0.73	0.57	0.64	0.7477
	Desgosto	0.72	0.73	0.72	
	Medo	0.45	0.55	0.49	
	Alegria	0.90	0.90	0.90	
	Tristeza	0.59	0.71	0.65	
	Surpresa	0.85	0.80	0.82	
	Neutralidade	0.68	0.62	0.65	
	Média/Total				
InceptionV3 (Entrada = 185) + kNN (k = 15)	Raiva	0.66	0.65	0.65	0.7708
	Desgosto	0.84	0.67	0.75	
	Medo	0.58	0.48	0.53	
	Alegria	0.9	0.91	0.91	
	Tristeza	0.66	0.66	0.66	
	Surpresa	0.85	0.83	0.84	
	Neutralidade	0.65	0.76	0.7	
	Média/Total				
InceptionV3 (Entrada = 185) + Random Forest (D = 27)	**Raiva**	**0.69**	**0.65**	**0.67**	**0.7787**
	Desgosto	**0.85**	**0.69**	**0.76**	
	Medo	**0.63**	**0.49**	**0.55**	
	Alegria	**0.9**	**0.92**	**0.91**	
	Tristeza	**0.64**	**0.71**	**0.67**	
	Surpresa	**0.87**	**0.85**	**0.86**	
	Neutralidade	**0.66**	**0.73**	**0.7**	
	Média/Total				
InceptionV3 (Entrada = 185) + SVM (C = 0.007)	Raiva	0.67	0.65	0.66	0.7769
	Desgosto	0.82	0.7	0.75	
	Medo	0.61	0.5	0.55	
	Alegria	0.9	0.91	0.91	
	Tristeza	0.65	0.68	0.66	
	Surpresa	0.85	0.85	0.85	
	Neutralidade	0.67	0.74	0.7	
	Média/Total				

B.3 ResNet50

Tabela 26 – Arquitetura ResNet50 avaliando a base de validação geral com imagens quadradas em 185 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
ResNet50 (Entrada = 185) + Softmax	Raiva	0.58	0.61	0.59	0.7322
	Desgosto	0.94	0.48	0.64	
	Medo	0.69	0.31	0.43	
	Alegria	0.83	0.93	0.88	
	Tristeza	0.67	0.55	0.61	
	Surpresa	0.76	0.85	0.80	
	Neutralidade	0.60	0.72	0.66	
	Média/Total				
ResNet50 (Entrada = 185) + kNN (k = 15)	**Raiva**	**0.61**	**0.6**	**0.6**	**0.7526**
	Desgosto	**0.84**	**0.67**	**0.75**	
	Medo	**0.56**	**0.5**	**0.53**	
	Alegria	**0.88**	**0.9**	**0.89**	
	Tristeza	**0.63**	**0.63**	**0.63**	
	Surpresa	**0.85**	**0.83**	**0.84**	
	Neutralidade	**0.64**	**0.72**	**0.67**	
	Média/Total				
ResNet50 (Entrada = 185) + Random Forest (D = 27)	Raiva	0.63	0.58	0.6	0.7520
	Desgosto	0.87	0.68	0.76	
	Medo	0.58	0.46	0.51	
	Alegria	0.87	0.91	0.89	
	Tristeza	0.63	0.63	0.63	
	Surpresa	0.85	0.83	0.84	
	Neutralidade	0.62	0.72	0.67	
	Média/Total				
ResNet50 (Entrada = 185) + SVM (C = 0.007)	Raiva	0.62	0.61	0.61	0.7516
	Desgosto	0.8	0.7	0.75	
	Medo	0.54	0.5	0.52	
	Alegria	0.88	0.9	0.89	
	Tristeza	0.63	0.63	0.63	
	Surpresa	0.83	0.83	0.83	
	Neutralidade	0.66	0.69	0.67	
	Média/Total				

B.4 VGG19

Tabela 27 – Arquitetura VGG19 avaliando a base de validação geral com imagens quadradas em 185 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
VGG19 (Entrada = 185) + Softmax	Raiva	0.70	0.61	0.65	0.7679
	Desgosto	0.77	0.70	0.74	
	Medo	0.66	0.46	0.54	
	Alegria	0.90	0.91	0.90	
	Tristeza	0.75	0.57	0.65	
	Surpresa	0.78	0.89	0.83	
	Neutralidade	0.60	0.78	0.68	
	Média/Total				
VGG19 (Entrada = 185) + kNN (k = 15)	Raiva	0.65	0.66	0.66	0.7802
	Desgosto	0.85	0.67	0.75	
	Medo	0.65	0.47	0.55	
	Alegria	0.89	0.92	0.91	
	Tristeza	0.65	0.71	0.68	
	Surpresa	0.86	0.85	0.85	
	Neutralidade	0.68	0.74	0.71	
	Média/Total				
VGG19 (Entrada = 185) + Random Forest (D = 27)	Raiva	0.69	0.67	0.68	0.7885
	Desgosto	0.88	0.69	0.78	
	Medo	0.66	0.5	0.57	
	Alegria	0.9	0.93	0.91	
	Tristeza	0.67	0.7	0.68	
	Surpresa	0.86	0.85	0.86	
	Neutralidade	0.68	0.76	0.72	
	Média/Total				
VGG19 (Entrada = 185) + SVM (C = 0.007)	**Raiva**	**0.67**	**0.67**	**0.67**	**0.7886**
	Desgosto	**0.82**	**0.74**	**0.78**	
	Medo	**0.63**	**0.55**	**0.58**	
	Alegria	**0.9**	**0.92**	**0.91**	
	Tristeza	**0.69**	**0.68**	**0.68**	
	Surpresa	**0.86**	**0.85**	**0.85**	
	Neutralidade	**0.7**	**0.74**	**0.72**	
	Média/Total				

B.5 MobileNetV2

Tabela 28 – Arquitetura MobileNetV2 avaliando a base de validação geral com imagens quadradas em 185 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
MobileNetV2 (Entrada = 185) + Softmax	Raiva	0.52	0.59	0.55	0.7062
	Desgosto	0.73	0.58	0.65	
	Medo	0.48	0.41	0.45	
	Alegria	0.81	0.91	0.86	
	Tristeza	0.61	0.50	0.55	
	Surpresa	0.82	0.79	0.80	
	Neutralidade	0.64	0.61	0.63	
	Média/Total				
MobileNetV2 (Entrada = 185) + kNN (k = 15)	Raiva	0.55	0.57	0.56	0.7169
	Desgosto	0.77	0.61	0.68	
	Medo	0.49	0.41	0.44	
	Alegria	0.84	0.9	0.87	
	Tristeza	0.58	0.55	0.56	
	Surpresa	0.82	0.82	0.82	
	Neutralidade	0.62	0.65	0.64	
	Média/Total				
MobileNetV2 (Entrada = 185) + Random Forest (D = 27)	**Raiva**	**0.58**	**0.55**	**0.56**	**0.7239**
	Desgosto	**0.8**	**0.63**	**0.7**	
	Medo	**0.54**	**0.42**	**0.47**	
	Alegria	**0.84**	**0.9**	**0.87**	
	Tristeza	**0.57**	**0.57**	**0.57**	
	Surpresa	**0.83**	**0.83**	**0.83**	
	Neutralidade	**0.62**	**0.66**	**0.64**	
	Média/Total				
MobileNetV2 (Entrada = 185) + SVM (C = 0.007)	Raiva	0.56	0.55	0.55	0.7131
	Desgosto	0.72	0.63	0.67	
	Medo	0.52	0.35	0.42	
	Alegria	0.84	0.9	0.87	
	Tristeza	0.56	0.56	0.56	
	Surpresa	0.79	0.82	0.8	
	Neutralidade	0.63	0.65	0.64	
	Média/Total				

ANEXO C – Avaliação experimental na base de validação geral em imagens quadradas de 210 *pixels*

Neste anexo são apresentados os resultados dos modelos avaliando a base de validação geral com imagens quadradas em 210 *pixels*.

C.1 InceptionResNetV2

Tabela 29 – Arquitetura InceptionResNetV2 avaliando a base de validação geral com imagens quadradas em 210 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
InceptionResNetV2 (Entrada = 210) + Softmax	Raiva	0.65	0.61	0.63	0.7394
	Desgosto	0.84	0.61	0.71	
	Medo	0.46	0.56	0.51	
	Alegria	0.92	0.86	0.89	
	Tristeza	0.87	0.40	0.55	
	Surpresa	0.84	0.84	0.84	
	Neutralidade	0.55	0.85	0.67	
	Média/Total				
InceptionResNetV2 (Entrada = 210) + kNN (k = 10)	Raiva	0.63	0.67	0.65	0.7732
	Desgosto	0.85	0.68	0.75	
	Medo	0.58	0.53	0.55	
	Alegria	0.89	0.92	0.9	
	Tristeza	0.68	0.65	0.66	
	Surpresa	0.87	0.83	0.85	
	Neutralidade	0.67	0.73	0.7	
	Média/Total				
InceptionResNetV2 (Entrada = 210) + Random Forest (D = 27)	Raiva	0.69	0.64	0.66	0.7778
	Desgosto	0.86	0.68	0.76	
	Medo	0.63	0.51	0.57	
	Alegria	0.88	0.92	0.9	
	Tristeza	0.65	0.68	0.66	
	Surpresa	0.87	0.85	0.86	
	Neutralidade	0.67	0.74	0.7	
	Média/Total				
InceptionResNetV2 (Entrada = 210) + SVM (C = 0.007)	**Raiva**	**0.66**	**0.66**	**0.66**	**0.7785**
	Desgosto	**0.81**	**0.7**	**0.75**	
	Medo	**0.6**	**0.51**	**0.55**	
	Alegria	**0.9**	**0.92**	**0.91**	
	Tristeza	**0.67**	**0.67**	**0.67**	
	Surpresa	**0.86**	**0.85**	**0.85**	
	Neutralidade	**0.68**	**0.74**	**0.71**	
	Média/Total				

C.2 InceptionV3

Tabela 30 – Arquitetura InceptionV3 avaliando a base de validação geral com imagens quadradas em 210 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
InceptionV3 (Entrada = 210) + Softmax	Raiva	0.62	0.70	0.66	0.7676
	Desgosto	0.83	0.71	0.77	
	Medo	0.52	0.58	0.55	
	Alegria	0.89	0.92	0.90	
	Tristeza	0.66	0.64	0.65	
	Surpresa	0.87	0.80	0.83	
	Neutralidade	0.71	0.66	0.69	
	Média/Total				
InceptionV3 (Entrada = 210) + kNN (k = 10)	Raiva	0.66	0.67	0.67	0.7821
	Desgosto	0.84	0.71	0.77	
	Medo	0.63	0.53	0.57	
	Alegria	0.9	0.91	0.91	
	Tristeza	0.69	0.66	0.67	
	Surpresa	0.85	0.85	0.85	
	Neutralidade	0.67	0.75	0.71	
	Média/Total				
InceptionV3 (Entrada = 210) + Random Forest (D = 27)	**Raiva**	**0.68**	**0.66**	**0.67**	**0.7868**
	Desgosto	**0.86**	**0.72**	**0.78**	
	Medo	**0.65**	**0.52**	**0.58**	
	Alegria	**0.9**	**0.92**	**0.91**	
	Tristeza	**0.65**	**0.7**	**0.67**	
	Surpresa	**0.87**	**0.86**	**0.86**	
	Neutralidade	**0.67**	**0.75**	**0.71**	
	Média/Total				
InceptionV3 (Entrada = 210) + SVM (C = 0.007)	Raiva	0.68	0.66	0.67	0.7853
	Desgosto	0.83	0.75	0.79	
	Medo	0.62	0.55	0.58	
	Alegria	0.91	630	0.91	
	Tristeza	0.67	0.68	0.68	
	Surpresa	0.85	0.85	0.85	
	Neutralidade	0.69	0.74	0.71	
	Média/Total				

C.3 ResNet50

Tabela 31 – Arquitetura ResNet50 avaliando a base de validação geral com imagens quadradas em 210 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
ResNet50 (Entrada = 210) + Softmax	Raiva	0.64	0.58	0.61	0.7287
	Desgosto	0.69	0.68	0.68	
	Medo	0.56	0.43	0.49	
	Alegria	0.86	0.90	0.88	
	Tristeza	0.55	0.64	0.59	
	Surpresa	0.85	0.79	0.82	
	Neutralidade	0.63	0.63	0.63	
	Média/Total				
ResNet50 (Entrada = 210) + kNN (k = 10)	**Raiva**	**0.62**	**0.63**	**0.62**	**0.7467**
	Desgosto	**0.81**	**0.64**	**0.71**	
	Medo	**0.55**	**0.48**	**0.51**	
	Alegria	**0.87**	**0.9**	**0.89**	
	Tristeza	**0.63**	**0.61**	**0.62**	
	Surpresa	**0.85**	**0.82**	**0.84**	
	Neutralidade	**0.64**	**0.69**	**0.66**	
	Média/Total				
ResNet50 (Entrada = 210) + Random Forest (D = 27)	Raiva	0.64	0.6	0.62	0.7464
	Desgosto	0.85	0.64	0.73	
	Medo	0.61	0.42	0.5	
	Alegria	0.86	0.91	0.88	
	Tristeza	0.6	0.62	0.61	
	Surpresa	0.85	0.83	0.84	
	Neutralidade	0.62	0.7	0.66	
	Média/Total				
ResNet50 (Entrada = 210) + SVM (C = 0.007)	Raiva	0.62	0.62	0.62	0.7440
	Desgosto	0.77	0.68	0.72	
	Medo	0.55	0.46	0.5	
	Alegria	0.87	0.9	0.88	
	Tristeza	0.61	0.6	0.61	
	Surpresa	0.83	0.84	0.83	
	Neutralidade	0.64	0.68	0.66	
	Média/Total				

C.4 VGG19

Tabela 32 – Arquitetura VGG19 avaliando a base de validação geral com imagens quadradas em 210 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
VGG19 (Entrada = 210) + Softmax	Raiva	0.64	0.66	0.65	0.7666
	Desgosto	0.70	0.75	0.72	
	Medo	0.64	0.47	0.54	
	Alegria	0.92	0.89	0.90	
	Tristeza	0.61	0.70	0.65	
	Surpresa	0.83	0.86	0.85	
	Neutralidade	0.69	0.70	0.69	
	Média/Total				
VGG19 (Entrada = 210) + kNN (k = 10)	Raiva	0.64	0.65	0.65	0.7788
	Desgosto	0.81	0.68	0.74	
	Medo	0.67	0.47	0.55	
	Alegria	0.90	0.92	0.91	
	Tristeza	0.65	0.70	0.67	
	Surpresa	0.87	0.85	0.86	
	Neutralidade	0.67	0.75	0.71	
	Média/Total				
VGG19 (Entrada = 210) + Random Forest (D = 27)	**Raiva**	**0.67**	**0.65**	**0.66**	**0.7869**
	Desgosto	**0.87**	**0.69**	**0.77**	
	Medo	**0.68**	**0.52**	**0.59**	
	Alegria	**0.9**	**0.92**	**0.91**	
	Tristeza	**0.67**	**0.7**	**0.68**	
	Surpresa	**0.87**	**0.85**	**0.86**	
	Neutralidade	**0.67**	**0.75**	**0.71**	
	Média/Total				
VGG19 (Entrada = 210) + SVM (C = 0.007)	Raiva	0.67	0.65	0.66	0.7850
	Desgosto	0.8	0.74	0.77	
	Medo	0.61	0.55	0.58	
	Alegria	0.9	0.92	0.91	
	Tristeza	0.68	0.68	0.68	
	Surpresa	0.85	0.85	0.85	
	Neutralidade	0.69	0.73	0.71	
	Média/Total				

C.5 MobileNetV2

Tabela 33 – Arquitetura MobileNetV2 avaliando a base de validação geral com imagens quadradas em 210 *pixels*

Arquitetura + Classificador	**Emoção**	**Precisão**	**Revocação**	**F1-score**	**Acurácia**
MobileNetV2 (Entrada = 210) + Softmax	Raiva	0.55	0.56	0.56	0.7127
	Desgosto	0.75	0.62	0.68	
	Medo	0.53	0.35	0.42	
	Alegria	0.83	0.91	0.87	
	Tristeza	0.55	0.61	0.58	
	Surpresa	0.81	0.81	0.81	
	Neutralidade	0.64	0.60	0.62	
	Média/Total				
MobileNetV2 (Entrada = 210) + kNN (k = 15)	Raiva	0.57	0.55	0.56	0.7182
	Desgosto	0.75	0.62	0.68	
	Medo	0.48	0.42	0.45	
	Alegria	0.85	0.89	0.87	
	Tristeza	0.58	0.58	0.58	
	Surpresa	0.82	0.81	0.82	
	Neutralidade	0.62	0.65	0.64	
	Média/Total				
MobileNetV2 (Entrada = 210) + Random Forest (D = 27)	**Raiva**	**0.59**	**0.54**	**0.57**	**0.7235**
	Desgosto	**0.78**	**0.63**	**0.70**	
	Medo	**0.51**	**0.43**	**0.47**	
	Alegria	**0.85**	**0.9**	**0.87**	
	Tristeza	**0.57**	**0.61**	**0.59**	
	Surpresa	**0.83**	**0.81**	**0.82**	
	Neutralidade	**0.61**	**0.66**	**0.64**	
	Média/Total				
MobileNetV2 (Entrada = 210) + SVM (C = 0.007)	Raiva	0.57	0.54	0.55	0.7130
	Desgosto	0.73	0.61	0.66	
	Medo	0.51	0.37	0.43	
	Alegria	0.85	0.9	0.87	
	Tristeza	0.56	0.58	0.57	
	Surpresa	0.8	0.82	0.81	
	Neutralidade	0.61	0.64	0.63	
	Média/Total				