**Reconhecimento de face e de “prova de vida” com Tensorflow
para criação de um sistema de segurança voltado a
residências e a ambientes de acesso restrito**

Sama Rouhani

Dissertação de Mestrado do Programa de Pós-Graduação em
Matemática Estatística e Computação Aplicadas à Indústria (PPG-
MECAI)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Sama Rouhani

**Reconhecimento de face e de “prova de vida” com
Tensorflow para criação de um sistema de segurança
voltado a residências e a ambientes de acesso restrito**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *EXEMPLAR DE DEFESA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Fernando Santos Osório

**USP – São Carlos
Março de 2019**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

R856r Rouhani, Sama
Reconhecimento de face e de prova de vida com
Tensorflow para criação de um sistema de segurança
voltado a residências e a ambientes de acesso
restrito. / Sama Rouhani; orientador Fernando
Santos Osório. -- São Carlos, 2019.
96 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Mestrado Profissional em Matemática, Estatística
e Computação Aplicadas à Indústria) -- Instituto de
Ciências Matemáticas e de Computação, Universidade
de São Paulo, 2019.

1. Reconhecimento Facial. 2. Prova de vida. 3.
Inteligência Artificial. 4. Segurança. 5. Redes
Neurais. I. Santos Osório, Fernando, orient. II.
Título.

Sama Rouhani

Face recognition and "proof of life" with Tensorflow to
create a security system to houses and restricted access
environments

Master dissertation submitted to the Institute of
Mathematics and Computer Sciences – ICMC- USP, in
partial fulfillment of the requirements for the degree of
the Master – Professional Masters in Mathematics,
Statistics and Computing Applied to Industry.
EXAMINATION BOARD PRESENTATION COPY

Concentration Area: Mathematics, Statistics and
Computing

Advisor: Prof. Dr. Fernando Santos Osório

USP – São Carlos
Março 2019

AGRADECIMENTOS

A presente dissertação de mestrado não poderia chegar a bom porto sem o precioso apoio de várias pessoas.

Agradeço primeiramente a Deus por me dar força intelectual e emocional em todas as etapas do mestrado.

Agradeço ao meu orientador Prof. Dr Fernando Santos Osório, por acreditar no meu potencial desde o primeiro semestre do curso, pela paciência e confiança na ideia do projeto, pelo empenho prático com que me orientou. Agradeço-o profundamente nessa fase da vida.

Agradeço à minha esposa Sandra Regina Vendramini Rouhani, pelo elevado grau de paciência e amor comigo no período do mestrado, por ser uma mulher forte, determinada e guerreira na vida, pela sua doçura e amor contagiantes, por ser uma mãe e esposa excelente.

Desejo igualmente agradecer a todos os meus colegas do MECAL, especialmente ao João Carlos Batista, Maria Rita, Renato Damasceno, Marcelo Pereira de Santana, Cláudio Vinícius Gonçalves e Cecília Pereira Sassi cujos apoio e amizade estiveram presentes em todos os momentos dentro e fora da sala de aula.

DEDICATÓRIA

Dedico essa dissertação à minha esposa Sandra, pelo apoio incondicional e à constante incentivo, paciência e amor.

Dedico essa dissertação ao meu orientador Prof Dr Fernando Santos Osório, por acreditar no meu potencial e me apoiar na ideia e criação do projeto, pelo incentivo, amizade e excelente orientação.

Dedico também aos meus pais Iraj e Mehrangiz por me apoiarem em vários aspectos nessa jornada longa e marcante da minha vida.

Dedico à minha esposa Sandra, pelo incentivo e força para continuar e terminar o mestrado, nessa fase da vida.

Dedico essa dissertação à ciência e aos futuros pesquisadores da área de Reconhecimento de Face e Prova de Vida com a ferramenta Tensorflow.

RESUMO

ROUHANI, S. Reconhecimento de Face e de “prova de vida” com Tensorflow para criação de um sistema de segurança voltado a residências e a ambientes de acesso restrito. 2019. 96p Dissertação (Mestrado - Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

O reconhecimento de faces é aplicado em várias tecnologias de segurança, como desbloqueios de celular, autenticação de entradas e saídas de estabelecimentos, meios de pagamentos via aplicativos móveis e outros. Para melhorar a confiança nos sistemas de segurança, foi desenvolvido também um reconhecedor de vida, com o uso do espectrograma da voz do usuário. Esse projeto propõe a criação de um sistema de segurança de autenticação de portas de residências e de acesso a salas e ambientes restritos (p.ex. laboratórios de pesquisa), com a aplicação das técnicas de reconhecimento de faces e de “prova de vida”, através de *Deep Learning* e Aprendizagem de Máquina, com o auxílio da ferramenta Tensorflow (um framework de aprendizado de máquina bastante utilizado no mercado), e com isso, criar um protótipo do sistema de reconhecimento facial e de vida, que pode ser facilmente convertido em um produto comercial. O projeto foi dividido em quatro redes neurais(RN-1, 2, 3, 4). A primeira, RN-1, consiste em treinar a rede com a base LFW (*Face Dataset*) para obter a melhor arquitetura de Aprendizado de Máquina. Então a aplicação realiza a captura da imagem e da voz do usuário. Na segunda rede, RN-2, utilizou-se dessa arquitetura selecionada para treinar a rede em base própria de reconhecimento de face (de complexidade 1:N), que foi criada especificamente para este trabalho. A terceira, RN-3, corresponde ao Reconhecimento da “Prova de Vida” (de complexidade 1:N), pois o sistema verifica se a voz (áudio) se refere à mesma pessoa da RN-2, evitando que o sistema seja enganado com a simples apresentação de uma foto/imagem no lugar da face humana real ou mesmo de um gravador de voz. Isso aumenta muito o grau de confiabilidade da aplicação, pois a verificação é dupla. A quarta rede, RN-4,(de complexidade 1:1) verifica também a voz da pessoa, porém há somente 2 classes: a do indivíduo em questão e a classe “outros” (outros usuários), portanto, haverá uma rede separada para cada usuário. Essa análise aumenta ainda mais o grau de confiabilidade da aplicação. Os resultados mostraram que: dentre 9 arquiteturas testadas, na RN-1, apenas a mobilenet_1.0_128 obteve os melhores resultados; a acurácia da rede, com a base própria na RN-2, obteve resultados muito melhores do que na RN-1 com a base LFW; o espectrograma de voz analisado na RN-3 obteve resultados muito satisfatórios com acurácia chegando a 92%; já na RN-4 as acurácias das redes chegaram a 100%.

Palavras-chave: Reconhecimento Facial, Prova de Vida, Segurança, Redes Neurais, Inteligência Artificial,.

ABSTRACT

ROUHANI, S. Face recognition and "proof of life" with Tensorflow to create a security system to houses and restricted access environments. 2019. 96p Dissertation (Master - Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Face recognition is applied in lot of security technologies, such as mobile unlocking, authentication of station inputs and outputs, payment options by mobile applications, and more. To improve confidence in security systems, a life recognizer has also been developed with the use of the user's voice spectrogram. This project proposes creation of a door entry security system and access to classrooms and restricted environments, with an application of face recognition and "Life-Proof" techniques, through the Deep Learning and Machine Learning, using Tensorflow tool, a framework of machine learning, the most famous of the market, with the objective of creating a prototype to a commercial product. The project was divided into four neural networks (RN-1, 2, 3, 4). The first, RN-1, consists in training a network with LFW (Face Dataset) to get the best architecture. Then the application makes a capture of the image and the voice of the user. In the second network, RN-2, we used this selected architecture to train the network based on its own face recognition dataset (1: N complexity), which was created specifically for this project. The third, RN-3, corresponds to the "Life-Proof" Recognition (of complexity 1: N), since the system checks if the voice refers to the same person as the RN-2, with the simple presentation of a photo / image in place of the actual human face or even of a voice recorder. This greatly increases the degree of reliability of the application, since the verification is double. The fourth network, RN-4 (1: 1 complexity) also verifies the person's voice, but there are only 2 classes: one of the individual in question and the class "other" (other users), so there will be a separate network for each person. This analysis improve the degree of reliability of the application. The results showed the best architecture was the mobilenet_1.0_128; the network with the own dataset (RN-2) obtained much better results than the network with the LFW base (RN-1); the voice spectrogram analyzed in RN-3 obtained very satisfactory results with accuracy reaching 92%; already in RN-4 the accuracy of the networks reached 100%.

Key Words: Facial Recognition, Life-Proof, Security, Neural Network, Artificial Intelligence

LISTA DE SIGLAS

CF	Características Faciais
ABIN	Agência Brasileira de Inteligência
3D	Tridimensional
IA	Inteligência Artificial
VIP	Very Important People
CNN	Convolutional Neural Network (Redes Neurais Convolucionais)
RNAs	Redes Neurais Artificiais
ANNs	Artificial Neural Networks
MLP	Multi Layer Perceptron (Perceptron Multicamadas)
OCR	Optical Character Recognition
ROI	Região de Interesse
SIADs	Sistemas Inteligentes de Apoio à Decisão
ML	Aprendizado de Máquina (<i>Machine Learning</i>)
RL	<i>Representation learning</i>
DCNN	Redes Convolucionais Profundas
GPU	Graphic Processing Unit
LPQ	Quantização de Fase Local
API	<i>Application Programming</i> Interface
LFW	Labeled Faces in the Wild
SVM	Support Vector Machine
AM	Aprendizado de Máquina
DL	Deep Learning
RNN	Recurrent Neural Networks
DBN	Deep Belief Networks
ASR	Reconhecimento Automático de Fala
HMM	CNN com Modelo Escondido de Markov

LISTA DE TABELAS

Tabela 2.1	Datasets de reconhecimentos de face.....	36
Tabela 4.1	Modelo de formato de análise das arquiteturas Mobilenet.....	55
Tabela 4.2	Modelo de imagens utilizadas na RN-1.....	56
Tabela 4.3	Modelo de imagens das RN-1 e RN-2.....	58
Tabela 4.4	Tabela hipotética dos limiares da Pessoa X para cada RN.....	63
Tabela 5.1	Matriz de confusão da base de teste para TOP 1 da RN-1.....	70
Tabela 5.2	Matriz de confusão da base de teste para TOP 2 da RN-1.....	71
Tabela 5.3	Matriz de confusão da base de teste para TOP 1 da RN-2.....	73
Tabela 5.4	Matriz de confusão da base de teste para TOP 2 da RN-2.....	73
Tabela 5.5	Matriz de confusão da base de teste para TOP 1 da RN-3.....	76
Tabela 5.6	Matriz de confusão da base de teste para TOP 2 da RN-3.....	76
Tabela 5.7	Resultados da RN-4.1 para o indivíduo A.....	78
Tabela 5.8	Resultados da RN-4.2 para o indivíduo B.....	78
Tabela 5.9	Resultados da RN-4.3 para o indivíduo C.....	78
Tabela 5.10	Resultados da RN-4.4 para o indivíduo D.....	79
Tabela 5.11	Resultados da RN-4.5 para o indivíduo E.....	79
Tabela 5.12	Limiar da Pessoa A para cada RN.....	81
Tabela 5.13	Limiar da Pessoa B para cada RN.....	82
Tabela 5.14	limiar da Pessoa D para cada RN.....	83
Tabela 5.15	Limiar da Pessoa E para cada RN.....	83

LISTA DE FIGURAS

Figura 1.1	Reconhecendo faces em multidões.....	02
Figura 2.1	Exemplo de neurônio.....	12
Figura 2.2	Exemplo de neurônio artificial.....	13
Figura 2.3	Esquema de um sistema de visão computacional.....	22
Figura 2.4	Partes de um algoritmo de visão computacional.....	22
Figura 2.5	Hiperplano gerado pelo SVM.....	25
Figura 2.6	Camadas de redes neurais.....	30
Figura 2.7	Exemplo dos 68 pontos de referência (landmarks).....	34
Figura 4.1	Esquema das Fases e Etapas do Sistema Proposto.....	52
Figura 4.2	Espectrogramas de voz da frase.....	61

LISTA DE GRÁFICOS

Gráfico 5.1	RNs aplicadas para classificar 2 classes.....	66
Gráfico 5.2	RNs aplicadas para classificar 3 classes.....	67
Gráfico 5.3	RNs aplicadas para classificar 4 classes.....	68
Gráfico 5.4	RNs aplicadas para classificar 5 classes.....	69
Gráfico 5.5	Gráfico de comparação dos TOP 1 e TOP 2 da RN-1.....	72
Gráfico 5.6	Gráfico de comparação dos TOP 1 e TOP 2 da RN-2.....	74
Gráfico 5.7	Gráfico de comparação dos TOP 1 e TOP 2 da RN-3.....	77
Gráfico 5.8	resultado das cinco redes neurais de reconhecimento de vida com complexidade 1:1.....	80

SUMÁRIO

1 INTRODUÇÃO	01
1.1 Contextualização.....	01
1.2 Motivação.....	05
1.3 Objetivo.....	08
1.4 Estrutura da Dissertação.....	09
2 FUNDAMENTAÇÃO TEÓRICA	11
2.1 Redes Neurais Artificiais.....	11
2.1.2 Redes Neurais Aplicadas à Detecção de Faces.....	15
2.2 Computação Visual.....	18
2.2.1 Visão Computacional.....	19
2.3 Inteligência Artificial (IA).....	23
2.3.1 Aprendizado de Máquina (AM).....	24
2.3.2 Support Vector Machine (SVM).....	25
2.3.3 Deep Learning.....	26
2.3.3.1 Convolutional Neural Network(CNN).....	28
2.4 Reconhecimento Facial.....	32
2.4.1 Detecção de Rostos.....	33
2.4.2 Alinhamento.....	33
2.4.3 Extração de Características.....	34
2.4.4 Treinamento.....	35
2.5 Reconhecimento de Vida.....	37

2.5.1 Espectrograma da Voz.....	38
2.6 Tensorflow	39
2.7 Considerações Finais.....	40
3 TRABALHOS RELACIONADOS	41
3.1 Considerações Finais.....	47
4 METODOLOGIA	49
4.1 Contextualização.....	49
4.2 Esquema do Projeto.....	51
4.3 RN-1 - Escolha da melhor arquitetura para o reconhecimento de face e de vida.	54
4.4 RN-2 - Reconhecimento de face com base própria com o uso da arquitetura vencedora da RN-1.....	57
4.4.1 Análise e Comparação das Imagens Utilizadas.....	58
4.5 Reconhecimento de Vida parte 1: RN-3	59
4.5.1 Reconhecimento de Vida parte 2: RN-4.....	61
4.6 Medidas dos Limiares.....	63
5 RESULTADOS E DISCUSSÃO	65
5.1 Resultados da RN-1 - Escolha da melhor arquitetura para o reconhecimento de face e de vida.....	65
5.2 RN-2 - Reconhecimento de face utilizando a própria base e a arquitetura mobilenet_1.0_128.....	72
5.3 RN-3 - Reconhecimento de vida com complexidade 1:N.....	75
5.3.1 RN-4 - Reconhecimento de vida com complexidade 1:1.....	77
5.4 Análise dos Limiares das Redes Neurais.....	80

5.4.1 Limiares da Pessoa A para cada RN.....	81
5.4.2 Limiares da Pessoa B para cada RN.....	81
5.4.3 Limiares da Pessoa C para cada RN.....	82
5.4.4 Limiares da Pessoa D para cada RN.....	82
5.4.5 Limiares da Pessoa E para cada RN.....	83
 6 CONCLUSÕES E TRABALHOS FUTUROS	 84
 7 REFERÊNCIAS BIBLIOGRÁFICAS	 86

CAPÍTULO 1

INTRODUÇÃO

1.1 Contextualização

Desde a sua infância o ser humano adquire e desenvolve a sua capacidade de reconhecimento facial, que é uma característica peculiar relacionada à visão, é o alicerce das relações sociais, onde atua como um dos fatores primordiais na evolução da espécie humana. As principais características da face humana foram representadas e relatadas há cerca de 27000 anos atrás (período neolítico) e, desde então, vem sendo representadas por inúmeros artistas e historiadores (Biometria: Reconhecimento Facial, 2017).

Há estudos sobre a automação do reconhecimento facial desde a década de 60. Os projetos iniciais nessa área dependiam do administrador para encontrar manualmente as características faciais (CF) nas imagens (olhos, nariz, orelha e boca) para somente então, o sistema calcular as distâncias entre as CF e comparar as suas dimensões normalizadas com as referenciadas. Segundo KIRBY & SIROVICH, 1987, a exata codificação da face, alinhada e normalizada, pode ser obtida com menos de cem pontos, sendo um fator importante para classificar as faces com o intuito de permitir o reconhecimento automatizado.

Uma grande aplicação da tecnologia de reconhecimento facial é para fins de vigilância e segurança de ambientes públicos (Figura 1). A identificação de criminosos procurados, de terroristas suspeitos, ou a localização de crianças desaparecidas, são apenas algumas das possíveis aplicações. As maiores vantagens desta tecnologia incluem:

- A capacidade de capturar faces em locais públicos (o que pode acelerar buscas policiais);
- A captura ser feita à distância (sem contato físico);
- A fácil integração com sistemas de vigilância já existentes (p.ex. CFTV);
- A fácil análise de bases de dados já existentes (por exemplo, registros criminais).
- Reconhecimento de várias faces simultaneamente, importante para ganhar tempo e acionar as autoridades em casos suspeitos.

Figura 1.1 – Reconhecendo faces em multidões

Fonte: Retirado de Reconhecimento facial e tecnologia para vigilância nacional.

Existem no mercado alguns sistemas de reconhecimento de face para destravar o celular, como o Samsung S8, o S8 Plus e o S9, mas alguns destes aparelhos ainda podem ser “enganados” com a apresentação de fotos/imagens no lugar da face. O iPhone X também possui um destravamento do aparelho via reconhecimento de face, porém esse sistema também pode ser falho em casos de indivíduos gêmeos. No entanto, o iPhone X possui um software e hardware dedicados que são muito eficientes para este tipo de aplicação (sensor 3D), sendo reconhecido como um dos melhores sistemas de reconhecimento facial do mercado.

No Brasil, por exemplo, desde 2016 os passageiros de voos internacionais que chegam a qualquer dos 14 aeroportos brasileiros são identificados por meio do reconhecimento das características faciais. O sistema permite também a integração com bancos de dados de outros órgãos de controle de fronteira e segurança, como Polícia Federal, Interpol e Agência Brasileira de Inteligência (ABIN), ampliando consequentemente a atuação da Receita Federal no combate a crimes transnacionais, tais como tráfico internacional de drogas, lavagem de dinheiro, comércio clandestino de pedras preciosas e obras de arte de alto valor (RECONHECIMENTO FACIAL EM AEROPORTOS, 2016).

O uso do reconhecimento facial pelas empresas tem crescido, principalmente nas ações de marketing. O cruzamento da imagem do rosto de clientes com um banco de dados pode permitir que varejistas, hotéis e bancos identifiquem rapidamente aqueles candidatos a um tratamento VIP, por exemplo. Além disso, reconhecer a expressão facial está se tornando uma ferramenta poderosa de *Business Intelligence*. Um Shopping Center na Finlândia, chamado Rajalla Pää Gränsen, entrou para a vanguarda do varejo mundial depois de adotar tecnologias que analisam o gênero, idade aproximada e até expressões faciais de quem circula pelos corredores.

No Brasil, o Santander, a Mastercard e a Dafiti iniciaram os testes do *Identity Check Mobile*, solução que faz a autenticação dos pagamentos online com o uso do reconhecimento facial. O objetivo é verificar, de forma simples e segura, a identidade do portador do cartão, sem a necessidade de digitar a senha. A rede americana de *fast-food*, KFC, lançou em setembro na China, um sistema de pagamento via reconhecimento facial.

Assim, se compreende a importância fundamental do uso do reconhecimento facial na área de segurança. Nesse projeto, este recurso também será usado com esse intuito, na autenticação de usuários para abertura e fechamento de portas de residências e ambientes de acesso restrito (p.ex. pequenas empresas e laboratórios de pesquisa). Além do reconhecimento facial, esse projeto também fará o reconhecimento da “prova de vida”, conferindo maior garantia na identificação de indivíduos.

No mercado de produtos de Biometria existem diversas tecnologias de reconhecimento de vida (*liveness detection*), como as baseadas em geometria da mão (3D), impressão digital com sensor de temperatura, leitura da íris, reconhecimento de face, reconhecimento de voz e reconhecimento de caligrafia. No projeto em questão, a prova de vida foi abordada pela análise do espectrograma da voz (“assinatura espectral” obtida por análise da voz de uma frase fornecida por cada pessoa).

O espectrograma pode ser definido como um gráfico que mostra a intensidade por meio do escurecimento ou coloração do traçado, as faixas de frequência no eixo vertical e o tempo no eixo horizontal (PONTES PAL, 2002). A identificação da voz pode ser distorcida por diferentes variações como ruído de fundo, canais de transmissão, emoções na voz e rouquidão, fatores que afetam significativamente a frequência, largura de banda e duração do segmento (HOLLIEN, 1974).

O reconhecimento de voz vem sendo utilizado na área forense, onde é uma das mais novas armas tecnológicas na guerra contra operações fraudulentas em bancos e para combate de acesso indevido a dados sigilosos.

Neste projeto a validação por voz foi realizada para que se tenha uma maior acurácia dos resultados e consequentemente uma maior segurança na identificação e autenticação do indivíduo para o sistema de controle de acesso. O reconhecimento de voz foi feito na etapa seguinte ao reconhecimento de face (GLOBAL IDENTITY MANAGEMENT; SPRINGER; BIOMETRICS).

1.2 Motivação

Existem atualmente vários *frameworks* por trás do Reconhecimento Facial, que utilizam recursos da Inteligência Artificial (I.A.), de *Deep Learning* e Visão Computacional, tais como o OpenCV, Facenet, OpenFace, Tensorflow e outros. A evolução destas tecnologias, ferramentas e métodos de reconhecimento facial têm permitido obter altos índices de acerto no reconhecimento de faces, tornando possível o uso deste tipo de sistemas em aplicações comerciais.

Segue abaixo uma breve descrição desses conceitos relacionados ao Reconhecimento Facial e como eles estão associados ao projeto em questão.

Deep Learning é uma técnica de Redes Neurais Artificiais com muitas camadas escondidas e que automaticamente “descobre” as representações necessárias para detecção ou classificação dos dados e padrões complexos (BENGIO, 2009).

O reconhecimento facial é uma subárea da Visão Computacional que ganhou muita atenção devido ao seu uso em sistemas de segurança e monitoramento de áreas públicas (ZHAO et al., 2003).

O TensorFlow (TENSORFLOW) é uma biblioteca de aprendizado de máquina criada pelos pesquisadores da *Brain Team* da Google e está agora com código aberto sob a licença Apache 2.0. O TensorFlow é um framework de redes neurais profundas (*Deep Learning*), que pode ser associado ao OpenFace que é uma biblioteca de reconhecimento de faces. O trabalho em questão, visa realizar o reconhecimento facial e a prova de vida usando esta ferramenta.

Um sistema de IA é capaz de fazer tanto o armazenamento e manipulação de dados e informações, quanto a aquisição, representação e manipulação de conhecimentos. Essa manipulação inclui a capacidade de deduzir ou inferir novos conhecimentos - novas relações sobre fatos e conceitos - a partir dos conhecimentos existentes e utilizar métodos matemáticos, estatísticos e computacionais de representação e manipulação de dados para resolver problemas complexos que não são, na maioria das vezes, exclusivamente quantitativos e nem qualitativos por natureza. Uma das ideias mais úteis que emergiram das pesquisas em IA, é que fatos e regras - conhecimento declarativo - podem ser representados separadamente dos algoritmos de decisão - conhecimento procedimental. Isto teve um efeito profundo, tanto na maneira dos cientistas abordarem os problemas, quanto nas técnicas de engenharia utilizadas para produzir Sistemas Inteligentes (SCHUTZER, 1987).

Já a Visão Computacional vem se destacando como um novo campo da Matemática Aplicada, que integra de forma efetiva as áreas de Visão Computacional, Visualização, Modelagem, Processamento de Imagens, Vídeo Digital, e Animação. O objetivo principal é desenvolver sistemas computacionais capazes de entender a cena como um todo ou elementos específicos em uma imagem. Dentre as aplicações destacam-se a criação de formas tridimensionais na cena a partir de imagens da mesma, reconhecimento de medidas físicas (distância, volume e densidade) e objetos a partir de imagens, visão artificial de sistemas robóticos e auxílio na interpretação de

exames médicos por imagens e no diagnóstico de doenças. Essa integração tornou possível grandes avanços no uso de modelos de alta complexidade baseados em dados reais (i.e. imagens) e a consequente criação de novas aplicações, tais como panoramas digitais, ambientes imersivos de realidade aumentada, interfaces visuais inteligentes, e reconhecimento de faces (VELHO, 2016).

Nos últimos anos, estas áreas, da IA e da Visão Computacional, vêm sendo muito estudadas e utilizadas em sistemas públicos de segurança e autenticação de usuários de empresas. As companhias que controlam mainframes, grandes servidores, agências de inteligências, sistemas governamentais entre outros, vêm investindo muito em diversos tipos de segurança além do reconhecimento de face, bem como o reconhecimento de digitais, da íris, de voz, de batimentos cardíacos, temperatura e outros parâmetros biométricos. Porém, essas empresas possuem grandes recursos financeiros e equipamentos que facilitam o desenvolvimento e aplicação destes tipos de tecnologias mais avançadas, que são de alto custo.

Um grande aspecto motivacional nesta área de pesquisa e desenvolvimento é a parte financeira, pois o mercado de reconhecimento facial deverá crescer em média 21,3% nos próximos quatro anos, para US\$ 9,6 bilhões até 2022, *segundo a Allied Market Research*.

O grande diferencial desse trabalho é a criação, com poucos recursos em termos de Hardware e Software (baixo custo), de um sistema de autenticação com o uso das excelentes ferramentas, de uso gratuito, disponíveis no mercado, e posteriormente podendo gerar um produto comercial. O projeto em questão, vai além do reconhecimento facial, pois atinge o reconhecimento/detecção de vida através do processamento da fala (reconhecimento de uma imagem contendo a “assinatura” do espectrograma da voz com a fala de uma pessoa) e, posteriormente, realiza a análise dos dados para chegar a uma verificação de autenticidade (validação do

reconhecimento facial, visando a autenticação de um usuário específico, denominada de “relação 1:1”). Essa aplicação é ideal em ambientes fechados como residências e ambientes de acesso restrito, como pequenas empresas e laboratórios de pesquisa, onde possa ser aplicado o reconhecimento facial e de vida para um conjunto limitado e bem definido de usuários, que passam a ter acesso a ambientes mais restritos.

1.3 Objetivo

O principal objetivo deste projeto de pesquisa e desenvolvimento é o de criar um sistema de reconhecimento facial e de prova de vida robusto e de baixo custo, voltado para aplicações que visam a liberação do acesso a ambientes restritos. Em termos mais específicos, o sistema proposto visa tratar um conjunto limitado de usuários, em torno de 5 usuários (de 3 a 10 pessoas cadastradas seria considerado um número adequado de usuários), que são autorizados a acessar um determinado ambiente com o acesso controlado pelo sistema. O sistema visa, portanto, alcançar uma alta taxa de confiabilidade no controle de acesso, ou seja, um número muito baixo de Falsos Positivos (acesso indevido) e um alto número de reconhecimentos corretos. Além disto, o sistema deve ser robusto às tentativas de ser enganado, e para isto a autenticação da pessoa será feita em duas etapas: a primeira por reconhecimento de face e a segunda por verificação da “prova de vida” através da análise do espectrograma de voz.

O reconhecimento de face e principalmente o de detecção de vida é algo muito buscado para aumentar o grau de certeza de autenticação dos sistemas de segurança. Atualmente a “prova de vida” é um grande desafio para ser superado. Dentro deste contexto, não foram encontrados muitos trabalhos relacionados ao reconhecimento de vida, principalmente com o uso de imagens do espectrograma da voz, que podem ser reconhecidos através do uso de uma ferramenta de *Deep Learning*.

Portanto, neste trabalho é proposto um método de reconhecimento de faces e de prova de vida, por meio da aplicação de *Deep Learning* e do Aprendizado de Máquina pela ferramenta Tensorflow. São utilizadas imagens para o reconhecimento de faces e o reconhecimento de voz faz uso do espectrograma da voz (representado também por uma imagem com a “assinatura” da voz do usuário) no reconhecimento de vida. Este trabalho tem o intuito de usar esta metodologia como uma aplicação em sistemas de segurança de ambientes restritos como pequenas empresas, laboratórios e residências.

1.4 Estrutura da Dissertação

A presente dissertação está dividida em seis capítulos organizados da seguinte forma:

Capítulo 1 - Introdução: esclarece sobre a área que é o tema principal deste estudo em questão, para contextualizar o problema. Define-se também o objetivo principal e os objetivos mais específicos deste trabalho.

Capítulo 2 - Fundamentação Teórica: Apresenta uma breve fundamentação teórica, tendo como objetivo principal mostrar teorias a respeito dos temas abordados nesta pesquisa. Neste contexto, será apresentada uma explanação a respeito de Redes Neurais Artificiais (RNAs ou ANNs), Redes Profundas (*Deep Learning*) e Redes Neurais Convolucionais (CNN ou DCNN), Computação Visual, Inteligência Artificial (IA ou AI), Reconhecimento Facial, Reconhecimento de Vida (*Liveness Detection* ou *Proof of Life*) e Tensorflow.

Capítulo 3 - Trabalhos Relacionados: Apresenta os trabalhos correlatos, contemplando trabalhos que possuem foco semelhante ou técnicas similares ao deste projeto.

Capítulo 4 - Metodologia: Apresenta a solução tecnológica desenvolvida durante o projeto de mestrado, que consiste na criação de um sistema voltado à área de segurança de controle de acesso de residências e ambientes de acesso restrito, que envolve reconhecimento de face e posteriormente detecção de vida com o uso de uma ferramenta muito utilizada no mercado, o Tensorflow.

Capítulo 5 - Resultados e Discussão: Apresenta os resultados obtidos da metodologia proposta e aplicando-os na prática do dia a dia de um sistema de controle de acesso.

Capítulo 6 - Conclusões e Trabalhos Futuros: Apresenta as conclusões deste trabalho, levando em consideração as contribuições, limitações e trabalhos futuros. Por fim, são apresentadas as referências bibliográficas utilizadas na realização deste documento.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

2.1 Redes Neurais Artificiais

As Redes Neurais Artificiais são modelos matemáticos inspirados nas Redes Neurais Cerebrais. As Redes Neurais Cerebrais são constituídas de milhares de neurônios interligados, direta ou indiretamente. Estes neurônios, transmitem pulsos eletroquímicos entre si, difundem estes pulsos através de excitação mútua e interpretam estes sinais, para que nós possamos agir. A partir dos estudos realizados sobre a maneira de pensar e agir, foi constatado que a teoria de que o cérebro poderia ser comparado a um computador digital convencional (antigamente chamado de “cérebro eletrônico”) estava equivocada. O processamento das informações no cérebro é realizado de forma distribuída, fortemente interconectada, e com alto grau de paralelismo entre as unidades de processamento (neurônios).

É possível afirmar que nem o melhor dos computadores da atualidade pode realizar determinados processos que são realizados pelo cérebro (p.ex. controle-motor, reconhecimento de padrões e percepção), com tanta precisão, flexibilidade e agilidade. Constatou-se também, que o cérebro, diferente do computador digital, trabalha de uma forma não linear e paralela, conseguindo assim organizar seus conhecimentos junto aos neurônios de forma a realizar seus processos de maneira única (HAYKIN, 1999).

Levando em consideração tais conhecimentos sobre a capacidade das redes neurais, despertou-se um enorme interesse, em desenvolver Redes Neurais Artificiais (RNAs) ou *Artificial Neural Nets* (ANNs) que pudessem modelar as redes neurais biológicas. As RNAs, segundo (FAUSETT, 1995), são sistemas de processamento de informações que tem como princípio alguma inspiração numa rede neural biológica. As Redes Neurais Artificiais têm sido desenvolvidas como generalizações de modelos matemáticos da cognição humana neural.

Os neurônios biológicos (Figura 2.1) basicamente são compostos por Dendritos e Axônios. Dendritos são filamentos que recebem as informações de outros neurônios ou de entradas externas (sensores). Não tem conexão física com os outros dendritos, e a ligação entre os dendritos é realizada através dos pulsos eletroquímico que são denominados de sinapses. A partir dos sinais recebidos pelos dendritos, o Corpo Celular (SOMA) realiza uma série de transformações e processos internos e, os resultados advindos destes processos são propagados aos outros neurônios da rede. O Axônio propaga, através das sinapses, o resultado dos processos do corpo celular aos neurônios subsequentes.

Figura 2.1 – Exemplo de neurônio

Fonte: Bianchini (2001)

O modelo de neurônio computacional (Figura 2.2) se assimila ao neurônio biológico. Em sua estrutura são encontrados os seguintes elementos: O conjunto de Sinapses, o Somador e a Função de Ativação. Assim como no modelo biológico, a sinapse consiste na entrada dos sinais para o neurônio (x_i), porém neste caso existem pesos diferentes associados às determinadas entradas (w_i), em que serão realizadas as operações. É importante ressaltar que os pesos serão positivos caso a sinapse seja excitatória e negativos se ela for inibitória (BIANCHINI, 2001).

O Somador ($\sum x_i * w_i - \theta$) realiza as operações de soma dos resultados advindos dos sinais, enquanto que a Função de Ativação ($g()$) ajusta e repassa os valores vindos do Somador limitando a amplitude de saída do neurônio (y).

Figura 2.2 – Exemplo de neurônio artificial

Fonte: Silva 2010

De uma maneira geral, as RNAs são redes maciças de unidades de processamento, paralelamente distribuídas e constituídas. Estas redes armazenam os conhecimentos que são empiricamente adquiridos, e os disponibilizam para usos posteriores, se assemelhando, assim, ao cérebro humano. A rede é treinada através de processos que ocorrem em seu ambiente e os conhecimentos são armazenados através de pesos sinápticos - x_i (HAYKIN, 1999).

As RNAs se dividem em diversos modelos. Alguns exemplos mais conhecidos, são descritos a seguir: Perceptron, MLP e Rede Neural Recorrente. O Perceptron foi o primeiro modelo implementado de RNA. Generalizando, o Perceptron é composto por duas camadas, sendo a camada de entrada responsável por executar a função de repassar à camada seguinte as variáveis de interesse, sem executar nenhuma operação e a camada de saída possui um neurônio, que corresponde à saída desejada. O Perceptron atua através do aprendizado de uma função de tomada de decisão para classificar dois conjuntos linearmente separáveis (MEHROTRA KISHAN e MOHAN, 2000).

Perceptron Multicamadas (Multi Layer Perceptron - MLP) é uma rede que conecta Perceptrons em camadas, sendo geralmente adotadas redes de 3 camadas. De maneira superficial, uma MLP pode ser vista como uma extensão natural do Perceptron, na qual se coloca outras camadas, que não têm ligação com as camadas de entrada e saída. A principal diferença do MLP é que os neurônios que são inseridos na RNA têm como função de ativação uma função do tipo não linear, ou seja, é aplicável a problemas independentes das classes serem linearmente separáveis. Uma de suas principais aplicações está relacionado ao reconhecimento de padrões (BIANCHINI, 2001).

A Rede Neural Recorrente se diferencia das redes MLP por ter pelo menos um *loop* de realimentação, ou seja, esta rede possui um sistema de *feedback*. Pode haver uma única camada de neurônio, onde apenas um neurônio reabastece as entradas dos demais neurônios (ARBIB, 2002).

As RNAs têm uma ampla gama de áreas de aplicação, como por exemplo, aplicações militares (ex. processamento de sinais, identificação de alvos, etc.), processamento de imagem (ex. reconhecimento de padrões, análise imagens, etc.), reconhecimento de voz, manufatura (prognóstico de consumo de energia), controle de

processos (identificação e controle de processos), robótica, dentre outras. Devido ao bom nível de eficiência das RNAs no reconhecimento de padrões, foi proposto o seu uso para colaborar na solução do problema de reconhecimento de faces.

2.1.2 Redes Neurais Aplicadas à Detecção de Faces

GOUVEIA & PAIVA, (2009) apresentaram um algoritmo de detecção de faces utilizando MLP, onde os atributos são extraídos a partir da segmentação em imagens coloridas. A rede MLP apresenta excelentes resultados na classificação de elementos como face ou não face. É interessante notar que, neste caso, as RNAs são utilizadas como suporte à segmentação de imagem, uma das etapas mais complexas no processo de visão computacional, o que faz com que o processo de classificação receba atributos mais precisos.

Dentre as diversas metodologias aplicadas na detecção de faces, muitas se baseiam em amostras de padrões de determinadas classes, como por exemplo, métodos baseados em características invariantes, métodos baseados em conhecimento, entre outros.

Nas abordagens tratadas acima, o treinamento das RNAs é simples, pois, utilizam padrões passados em cada classe, para calcular os parâmetros das funções de decisão das mesmas. Estes elementos têm sua essência fundamentada em padrões computacionais não lineares, que se dão pelos neurônios da rede (GONZALES & WOODS, 2010).

Segundo BIANCHINI, 2001; GOUVEIA & PAIVA, 2009, a utilização de RNAs multicamadas, para a implementação e melhoria dos problemas encontrados para o desenvolvimento das tecnologias voltadas para detecção de faces, produz excelentes resultados.

O avanço obtido pela utilização das RNAs multicamadas (MLPs), nas tecnologias de detecção de faces, se dá em função da possibilidade de treinar as diversas camadas de neurônios para aprender uma característica específica e detectá-la em uma face, solucionando assim, uma das principais dificuldades no desenvolvimento deste tipo de tecnologia, ligada a segmentação da face. É possível implementar uma camada, por exemplo, para o reconhecimento dos olhos, outra camada para o reconhecimento específico do nariz, outra para o reconhecimento da boca e inclusive uma camada específica para a medida de proporções áureas da face humana.

As características advindas das proporções áureas são aspectos biométricos presentes somente nos seres vivos, portanto presente nos seres humanos. Estas características estão ligadas diretamente a harmonia ou ausência da mesma em nossas faces independentemente de suas expressões. O treinamento das RNAs multicamadas para a classificação de formatos, independente da interferência de ruídos nas amostras, se dá através de formação de vetores de padrões que são gerados a partir das assinaturas dos formatos uniformemente espaçados (GONZALES & WOODS, 2010).

Para isso, uma rede é treinada com a inicialização aleatória dos pesos, e ajustados de acordo com os vetores de características correspondentes às amostras. Enquanto o treinamento da rede é realizado, os nós de saída podem ser monitorados e, somente então, pode-se julgar que as camadas da rede tenham aprendido a realizar as tarefas a elas atribuídas, com relação aos formatos pré-estabelecidos. Devido à possibilidade das RNAs multicamadas apresentarem resultados satisfatórios para o reconhecimento facial, algumas delas, projetadas especificamente para outros tipos de reconhecimento, acabaram apresentando, também, boas respostas para o problema de reconhecimento facial. É o caso, por exemplo, da Rede Neural Neocognitron, a qual foi inicialmente proposta para o reconhecimento de escrita (OCR - Optical Character

Recognition) e que posteriormente foi utilizada, também, para o reconhecimento facial. Esta rede utiliza o conceito de várias camadas neurais e foi baseado no modelo de Hubel e Wiesel, sendo possível notar uma fase de treinamento não supervisionado e uma fase de reconhecimento (FUKUSHIMA, 1995).

A utilização das RNAs para detecção de faces funciona basicamente com a aplicação de filtros sobre a janela de padrões definida. No entanto, os dados retirados destas aplicações podem não ser tão confiáveis, devido aos conjuntos de detecções sobrepostos e às falsas detecções que podem ocorrer. Para isto, foram desenvolvidas técnicas direcionadas à solução destes problemas. O processamento de imagens e a extração de atributos das imagens, denominada de pré-processamento das entradas, pode contribuir para melhorar o reconhecimento de faces. Neste caso, a primeira fase é responsável por realizar um pré-processamento na imagem de entrada, verificando se realmente existe uma face, logo após são realizadas correções de iluminação e a equalização do histograma e ainda nesta fase é realizado o janelamento da imagem da face.

Após esta primeira etapa, a janela de padrões é passada como entrada para a rede e propagada para suas camadas e subcamadas, que realizarão a detecção de características como, olhos, boca, nariz, pele, cabelo e outras. Para que a rede seja treinada e atue como um filtro, pode-se, por exemplo, estipular padrões a partir de *templates* variados e rotulados, após detectadas as características comuns. A detecção de uma característica pré-estabelecida é dada como saída e, então, é armazenada. Em seguida, realiza-se um cálculo central de cada detecção incomum, para que os pontos passem por um processo de *spreading out* (expansão de saída) e, só então, a partir destes dados, é estipulado um limiar, que delimita um número mínimo de pontos detectados, para que se considere uma determinada imagem como uma face. Ainda existem outros métodos para que se possa realizar a detecção de uma face, como, por

exemplo, a utilização de múltiplas redes, arbitrando suas saídas para uma decisão final.

Por um longo período, da década de 80 até o início dos anos 2000, as redes MLP eram constituídas de poucas camadas, usualmente de 3 a 4 camadas, sendo denominadas de Redes Neurais Rasas (*Shallow Networks*). No final dos anos 90, começaram a aparecer redes com mais camadas, e por volta de 2010 surgem as Redes Neurais Profundas (*Deep Learning*), que começaram a se tornar viáveis e a se consolidar como uma abordagem robusta para o reconhecimento de padrões complexos em imagens.

2.2 Computação Visual

A Computação Visual vem se configurando como um novo campo da Matemática Aplicada Computacional que integra de forma efetiva as áreas de Visão Computacional, Visualização, Modelagem, Processamento de Imagens, Vídeo Digital e Animação. Para este trabalho, a atenção maior foi para Visão Computacional e Processamento de imagens.

O próximo passo na evolução da Computação Visual será uma resposta aos avanços sem precedentes que estão ocorrendo na indústria de hardware, em particular nos equipamentos de entrada e saída gráfica. De fato, com os efeitos da lei de Moore, estamos chegando a um ponto em que mudanças quantitativas certamente levarão a uma mudança qualitativa. Com isso teremos novas mídias inteligentes e amigáveis, que estarão integradas entre si de forma dinâmica e transparente, nas quais o papel do computador será indissociável da sua própria natureza, onde a computação visual atuará na orquestração das aplicações. Para realizar plenamente esses desígnios, um dos grandes desafios da Computação Visual compreende a efetivação de um canal entre os modelos matemáticos, os métodos computacionais e as demandas das aplicações. Isso envolve a análise e síntese a partir de “coleções de imagens” que

traduzam dados do mundo real e permitam a elaboração de modelos probabilísticos não-paramétricos. O segredo está no emprego de processos estocásticos associados a métodos computacionais de alta escalabilidade e que levem em conta aspectos perceptuais.

2.2.1 Visão Computacional

Visão Computacional é a área da ciência que estuda e procura desenvolver teorias e métodos voltados à extração automática de informações úteis contidas em imagens. Tais imagens são capturadas por dispositivos como câmeras de vídeo e scanners (CROWLEY & CHRISTENSEN, 1995).

Por este motivo, a recente evolução tecnológica dos computadores e dispositivos de imagem, a preços acessíveis ao consumidor, tem possibilitado a crescente aplicação de visão computacional nas mais diversas áreas. Como por exemplo, análise automática de sêmen humano, medição computadorizada do dimensional de peças, monitoramento de animais experimentais, análise morfológica de células, reconhecimento e sintetização de faces humanas, entre outras (FAUGERAS, 1993).

A criação de sistemas de visão computacional, seja para aplicações em inspeção industrial ou para navegação de robôs móveis, envolve, quase sempre, a execução de um determinado conjunto de transformações em dados obtidos de sensores como câmeras e sonares. Para esta classe de problemas é possível a determinação de uma arquitetura de software que suporte as principais etapas do desenvolvimento deste tipo de sistema, enfatizando o reaproveitamento de código e minimizando o esforço despendido em atividades repetitivas (HARTLEY; ZISSERMAN, 2003).

O objetivo da construção de máquinas inteligentes é expandir os mecanismos (sentidos) através dos quais o computador pode se comunicar com o mundo exterior. A utilização da visão da máquina aumenta as aplicações em computadores, como por exemplo, navegação móvel por robô, tarefas complexas de manufatura, análise de imagens de satélites e processamento de imagens médicas.

A visão computacional procura oferecer uma vasta quantidade de informações ao sistema computacional. A partir de um conjunto de imagens, de modo mais eficiente possível. O reconhecimento de padrões está relacionado a área da visão computacional com atuações e perspectivas importantes para obter máquinas inteligentes (JÄHNE, 2002).

Para todo sistema de visão computacional, como mostrado nas Figuras 4 e 5, são necessárias algumas etapas de processamento que variam de acordo com a aplicação. Na maioria dos casos, ocorrem as seguintes etapas:

- **Aquisição de imagem:** todas as imagens digitais são criadas por um ou mais sensores. De acordo com cada sensor pode-se ter vários tipos de imagens, como bidimensional, tridimensional ou ainda uma sequência de imagens. Cada pixel tem um valor agregado e cada valor é responsável por uma intensidade de luz em uma ou em várias faixas de cor (MORRIS, 2004). A faixa de cores está associada ao espectro eletromagnético, visível ou não, dentro de uma determinada frequência. As imagens adquiridas possuem uma amostragem (que define sua resolução) e uma quantização (que define o valor medido para cada canal de cor - sendo usualmente amostradas e quantizadas em 3 canais: RGB).
- **Pré-processamento:** antes da aplicação de qualquer método de visão computacional é necessário verificar se a imagem satisfaz as condições do método a ser utilizado. Para isso podemos citar alguns exemplos:

remapeamento (para assegurar o sistema de coordenadas), redução de ruídos (para assegurar que as informações sejam mais próximas e consistentes com os dados reais) e aumento de contraste (para assegurar que as informações relevantes sejam detectadas e destacadas). No monitoramento de animais, por exemplo, o aumento de contraste foi muito importante, com o intuito de melhor detectar os objetos de interesse (MORRIS, 2004).

- **Extração de características:** existem várias características matemáticas que se pode obter em uma imagem com diversos níveis de complexidade. Como exemplo básico inclui-se a detecção de bordas, de cantos ou de pontos. Exemplos mais complexos incluem detecção de textura, formato e movimento (MORRIS, 2004).
- **Detecção e segmentação:** durante uma parte deste processo sempre se toma uma determinada decisão em relação às regiões de interesse (ROI) que serão processadas posteriormente, como por exemplo, seleção de regiões de interesse específicos e segmentação de uma ou mais regiões que contém um objeto de interesse (MORRIS, 2004).
- **Processamento de alto nível:** nesta etapa a entrada é definida como um conjunto pequeno de dados, ou seja, no caso do monitoramento de animais, o dado é apenas o objeto de interesse já detectado. O processo seguinte caracteriza-se pela verificação da satisfação dos dados, pela estimativa de parâmetros sobre a imagem e pela classificação dos objetos detectados em diferentes categorias (MORRIS, 2004).

Figura 2.3: Esquema de um sistema de visão computacional

Fonte: DAMASCENO SILVA, 2008

Um sistema de visão computacional definido para o reconhecimento de faces e de pessoas em imagens digitais deve ter como objetivo a extração de características de pessoas ou de faces. Um único indivíduo pode variar suas características, o que o difere da detecção de objetos. Como exemplos dessa variação podemos citar as roupas, os óculos, o cabelo, a barba, a expressão facial da pessoa (alegre, triste, zangada, etc.), entre diversas outras características, fatores esses que se tornam complicadores para o reconhecimento. As características que podem ser usadas na detecção de pessoas em uma imagem, por meio de visão computacional, são por exemplo a face, a cabeça, todo o corpo, incluindo as pernas e a pele.

Figura 2.4: Partes de um algoritmo de visão computacional, a extração de características está implícita junto com o pré-processamento.

Fonte AMARO, R; DIEGUEZ, A., 2000.

2.3 Inteligência Artificial (IA)

Para LUGER e STUBBLEFIELD, 2009, Inteligência Artificial é a área da computação voltada para a automação de comportamentos inteligentes. Winston (citado em FERNANDES, 2005, p. 2) define essa área como o estudo da computação que torna possível perceber, raciocinar e agir. Diz ainda, que IA é o estudo das ideias que permitem aos computadores serem inteligentes (FERNANDES, 2005, p. 2)

A Inteligência Artificial é a capacidade que um sistema computacional tem de raciocinar, planejar, processar ideias, compreender linguagem e adquirir conhecimento. O conceito de IA pode ser sintetizado na capacidade do homem em desenvolver sistemas computacionais, que são capazes de simular o raciocínio humano e de ser inteligente (FERNANDES, 2005; LORENZI; SILVEIRA, 2011).

As técnicas de IA processam os dados para criar relatórios gerenciais para melhor interpretação dos dados e informações. A IA muitas vezes está atrelada a sistemas maiores de Big Data e a grandes banco de dados. Desta forma, diferentes técnicas de IA (como *Data Mining*, Árvores de Decisão, Sistemas Especialistas e Redes Neurais Artificiais) podem ser empregadas no desenvolvimento de SIADs (Sistemas Inteligentes de Apoio à Decisão).

As aplicações utilizam várias técnicas de IA, como classificação, regressão, detecção de desvios, regras de associação, padrões sequenciais, agrupamento e sumarização. Atualmente, grande parte destas aplicações requer a utilização de Algoritmos de Aprendizado de Máquina para que possam ser criados modelos eficientes de tratamento de dados baseados no aprendizado.

2.3.1 Aprendizado de Máquina (AM)

A principal característica da inteligência humana é o aprendizado, essa representa os meios básicos para a obtenção de conhecimentos. Segundo HUA, 2009, o processo de aprendizagem humana envolve a memória, o pensamento, a percepção, o sentimento, e outras atividades mentais relacionadas. Comparado à aprendizagem humana, o Aprendizado de Máquina (AM) é mais rápido, o acúmulo de conhecimento é facilitado e os resultados da aprendizagem são mais fáceis de demonstrar. Entretanto, esse processo depende diretamente da ação humana, isto implica que todo o progresso do ser humano no campo de AM, vai aumentar a capacidade dos computadores em aprender, auxiliando no melhoramento do processamento das informações.

Segundo HUA 2009, o aprendizado é uma atividade que processa a informação de fora para dentro. As informações do ambiente externo são processadas e geram conhecimento, que posteriormente é armazenado em um repositório, guardando muitos princípios gerais que norteiam uma parte da ação de execução. O aprendizado dependerá muito da qualidade dessas informações fornecidas pelo ambiente, tornando a aprendizagem fácil e organizada ou difícil e desordenada.

Segundo ABRAMSON, (1963), as máquinas vão identificar e atribuir os objetos e elementos às classes, definindo um reconhecimento automático de padrões. Esse estudo atua na compreensão dos conceitos de como ocorre o Aprendizado de Máquina, que tem sido uma das principais técnicas utilizadas para este fim na atualidade.

Nesta dissertação iremos tratar de um tipo específico de Aprendizado de Máquina, o aprendizado supervisionado. O processo de aprendizado supervisionado ocorre através da apresentação de bases de dados etiquetadas, onde cada padrão

possui uma classe previamente associada a ele. Os dados de treinamento, previamente classificados, são usados para o treinamento e ajuste dos parâmetros do modelo a ser aprendido, permitindo que o sistema possa depois proceder o reconhecimento automático de padrões, baseado nos exemplos apresentados.

2.3.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) é uma poderosa técnica para o aprendizado supervisionado e classificação de padrões (VAPNIK, 1995). O SVM cria e ajusta um hiperplano ou um conjunto de hiperplanos em um alto ou infinito espaço dimensional, que pode ser usado para classificação de padrões.

O SVM efetua o reconhecimento de padrões entre duas classes construindo um plano ou hiperplano que maximiza a distância entre dois pontos próximos no conjunto dos pontos de treinamento (HEISELE, 2001), conforme a Figura 2.5. O SVM pode ser usado na resolução de problemas de classificação e de regressão.

Figura 2.5: Hiperplano gerado pelo SVM. Primeira imagem: maximização da distância entre os pontos das duas classes(verde e roxo). Segunda imagem: um hiperplano ou uma reta/delimitadora.

Fonte: Intel AI Academy

Existem ferramentas (*toolkits*) de aprendizado de máquina que disponibilizam algoritmos como o SVM, como por exemplo o Weka (WAIKATO UNIVERSITY) e as bibliotecas SVM-Light e o LibSVM (LIBSVM; SVM-LIGHT). Estas ferramentas permitem o aprendizado a partir de uma base de dados previamente classificada (aprendizado supervisionado), podendo depois usar o modelo aprendido para o reconhecimento de padrões.

2.3.3 Deep Learning

Para BENGIO, (2009), as decisões e o reconhecimento de padrões que antigamente eram tomadas de forma subjetiva, atualmente os computadores conseguem resolver com as técnicas de Aprendizado de Máquina (*Machine Learning, ML*). Computadores possuem facilidades na resolução de problemas com regras bem definidas como cálculos matemáticos, jogos de tabuleiro ou cartas, porém têm grandes dificuldades em atividades como reconhecimento de padrões em imagens, músicas, textos, etc. Por conta disso, as dificuldades enfrentadas sugerem que sistemas de Inteligência Artificial (IA) precisam da capacidade de adquirir seu próprio conhecimento, aprendendo a extrair padrões de dados brutos, como pixels em imagens e aprendendo a reconhecer padrões complexos.

Como as técnicas convencionais de ML são limitadas no processamento de dados brutos para o reconhecimento de padrões, elas dependem muito das técnicas de extração de *features* que transformam os dados brutos em uma representação eficaz, com o intuito de facilitar o trabalho do classificador de inferir os padrões nos dados de entrada. Posto isso, o extrator de *features* tem uma enorme influência no desempenho de algoritmos de ML. Outra dificuldade nessas aplicações é que muitos fatores influenciam na variação dos dados observados, por exemplo identificar o modelo de um carro com diferentes iluminações e posições ou reconhecer a face de uma pessoa ao

deixá-la de perfil. Desse modo, obter boas representações dos dados para reconhecimento de padrões complexos em imagens é algo bastante dispendioso (LECUN; BENGIO; HINTON, 2015).

Representation learning (RL) é um conjunto de métodos que permitem que as máquinas recebam dados brutos como entrada e, automaticamente, descubram as representações necessárias para detecção ou classificação (LECUN; BENGIO; HINTON, 2015). *Deep Learning* (DL), é uma subárea da ML, possui métodos de RL com vários níveis de representação, formado pela união de módulos simples, mas não lineares, que transformam cada representação de um nível em uma representação em nível superior, ligeiramente mais abstrato. Assim, através de vários algoritmos, foi-se criando modelos de abstrações de alto nível de múltiplas camadas de processamento, realizando várias transformações não lineares (BENGIO; COURVILLE; VINCENT, 2013), (DENG e YU, 2014), (SCHMIDHUBER, 2015), (LECUN; BENGIO; HINTON, 2015). Com isso, as antigas Redes Neurais Artificiais que eram “rasas” e compostas por apenas poucas camadas, passaram a ser representadas por Redes Profundas (Aprendizagem Profunda / Deep Learning), com múltiplos níveis e camadas de pré-processamento e posterior reconhecimento de padrões.

O *Deep Learning* permitiu que modelos computacionais, compostos de múltiplas camadas de processamento aprendam representações de dados com vários níveis de abstração. Esses métodos melhoraram drasticamente o estado-da-arte em reconhecimento de fala, reconhecimento de objetos e muitos outros domínios, como descoberta de novas drogas e pesquisa do genoma. O *Deep Learning* descobre estruturas complexas em grandes conjuntos de dados usando o algoritmo como o de *backpropagation* e as Redes Convolucionais Profundas (CNN - *Convolutional Neural Nets*) para indicarem como uma máquina deve alterar seus parâmetros internos, que são usados para calcular a representação em cada camada da representação na camada anterior. As CNNs trouxeram muitos avanços no processamento de imagens,

vídeos e áudios, enquanto que redes recorrentes permitiram tratar os dados sequenciais (como texto e áudio) de modo mais eficiente (BENGIO, 2015).

O *Deep Learning* em processamento de imagem vem na forma de uma matriz de valores de pixel, e os recursos aprendidos na primeira camada representam, normalmente, a presença ou ausência de atributos (*features*) como bordas em orientações e locais específicos na imagem. A segunda camada normalmente detecta arranjos particulares das bordas, independentemente de pequenas variações nas suas posições. A terceira camada pode montar combinações maiores que correspondam a partes de objetos familiares, e camadas subsequentes detectam objetos como combinações dessas partes. O aspecto principal do *Deep Learning* é que essas camadas de recursos não são projetadas por engenheiros humanos, elas são aprendidas com dados usando um procedimento de aprendizado para fins gerais (BENGIO, 2015).

Algumas estruturas de DL tais como *Deep Neural Networks* (DNN), *Recurrent Neural Networks* (RNN), *Convolutional Neural Networks* (CNN) e *Deep Belief Networks* (DBN), vêm sendo aplicadas em diversas áreas, obtendo resultados muito melhores do que os resultados indicados até então em competições de ML e Reconhecimento de Padrões de imagens (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Nesse projeto trabalhou-se com modelos do tipo das redes CNN (*Deep Convolutional Neural Nets*).

2.3.3.1 Convolutional Neural Network(CNN)

O nome “Convolução” provém de uma operação matemática muito usada no processamento de imagens e em filtros para detecção de bordas e outros atributos, as CNN’s são arquiteturas de Deep Learning (DL) que desenvolvem o aprendizado das matrizes de convolução. As matrizes de convolução permitem a extração automática de

padrões (*features*) que são importantes elementos para o reconhecimento de padrões complexos baseados em imagens.

As redes CNN's, de inspiração biológica, são capazes de aprender representações invariantes de escala, translação, rotação e transformações afins (LECUN; KAVUKCUOGLU; FARABET, 2010). Esse tipo de rede vem apresentando resultados promissores na identificação de objetos, promovendo uma melhora significativa do estado-da-arte (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). As redes CNNs vem sendo empregadas com muito sucesso no reconhecimento de faces.

A convolução é um tipo específico de operação linear e para ser considerada uma CNN é preciso que a rede neural utilize a convolução em pelo menos uma das suas camadas. De forma geral, a operação de convolução corresponde a duas funções com argumentos de valores reais e pode ser identificada pelo símbolo asterisco “ * ” , como aparece na seguinte expressão:

$$s(t) = (I * K)(t)$$

onde I corresponde ao dado de entrada, como por exemplo uma imagem, K ao kernel utilizado para convolver I no intervalo t . A saída $s(t)$ é referenciada muitas vezes como o *feature map*.

Em aplicações de ML, a entrada é geralmente uma matriz multidimensional dos dados (podendo ser uma imagem com os seus canais de cores RGB por exemplo) e o kernel é geralmente uma matriz multidimensional de parâmetros que são auto-reguláveis. Essas matrizes multidimensionais são conhecidas como tensores. Dessa forma, ao se aplicar convoluções em mais de um eixo, como em uma imagem e kernels bi-dimensionais I e K , respectivamente, a operação pode ser descrita pela seguinte expressão:

$$s[i, j] = (I * K)[i, j] = \sum_m \sum_n I[i + m, j + n] K[m, n]$$

onde i, j e m, n são as dimensões de I e K respectivamente.

Cada CNN representa um campo receptivo composto por combinações de pelo menos uma das seguintes camadas: *filter bank layer*, *non-linearity layer* e a *feature pooling layer* (LECUN; KAVUKCUOGLU; FARABET, 2010).

Cada camada é aplicada a toda a imagem de entrada (ou camada anterior) por meio de uma operação de convolução. O resultado da aplicação deste filtro é chamado de *feature map*. Cada *feature map* compartilha os mesmos parâmetros, como é ilustrado pela Figura 2.6. Essa estratégia garante que uma determinada feature será detectada pelas features map, independentemente da sua posição da imagem (ou mapa) de entrada.

Figura 2.6: três unidades escondidas que pertencem ao mesmo feature map e os pesos da mesma cor são compartilhados.

Fonte: Camadas de Redes Neurais

As entradas de dados de cada camada da CNN são um conjunto de *feature maps* e a saída de cada camada corresponde à aplicação do filtro de convolução. O filtro de convolução destaca algumas features e cada filtro é responsável por enfatizar

uma feature diferente, como a primeira camada, quando são destacadas as linhas e gradientes em diferentes orientações.

A camada de filtragem (*filter bank layer*) realiza a convolução dos filtros W_k^i , correspondente ao i -ésimo filtro da camada K de tamanho $I_1 \times I_2$ na imagem. Cada filtro detecta uma característica particular em todas as localizações na imagem de entrada. Na primeira camada a imagem de entrada I_{mn} , sendo “ m ” a altura e “ n ” o comprimento, é representada através de um conjunto de *feature maps*, um para cada canal de cor.

A camada não-linear, *non-linearity layer*, é responsável por aplicar uma função não-linear em cada um dos elementos que compõem as *feature maps*. HINTON et al. (2012b), ZEILER & FERGUS(2013) utilizam em seus trabalhos a função *rectified linear unit* (ReLU) descrita por

$$relu(z) = \max(0, z)$$

A camada de redução (*feature pooling layer*) ou como também é conhecida, sub-sampling, calcula a média ou o máximo de uma vizinhança pré-determinada em cada uma das *feature maps*, ou seja, retira certas amostras da rede para torná-la robusta a *overfitting*. O resultado é um outro mapa com resolução menor que é invariante a pequenas translações. ZEILER & FERGUS (2013) demonstram um novo método para redução com base na distribuição das ativações em cada região.

Ao término das etapas de convolução, os *feature maps* são convertidos em uma dimensão de 2 (dois) por 1 (um) e estas saídas intermediárias são utilizadas no treinamento de um classificador full-connected, com tamanho de saída correspondente à quantidade de categorias ou classes. O resultado do classificador, que são as camadas formadas por uma *Multilayer Perceptron* (MLP), passa pela função softmax, descrita por

$$\text{softmax}(y_i) = \frac{e^{\tilde{z}_i}}{\sum_{j \in Y} e^{\tilde{z}_j}}$$

garantindo, dessa forma, que a soma de todas as probabilidades de saída da MLP resulte em 1 (um), onde Y corresponde ao vetor de saída da MLP.

2.4 Reconhecimento Facial

Segundo JONES et al. (2007), os seres humanos podem detectar rapidamente outros seres humanos dentro do seu campo de visão e, na maioria dos casos, podem determinar os aspectos biológicos como sexo, idade aproximada, postura, andar, peso, altura e forma.

A capacidade humana de extrair informações visuais complexas do mundo tem sido amplamente documentada in JONES et al., 2007 e teorias para a compreensão de estímulos ambientais e contextuais estão sendo criadas. (JONES et al., 2007)

O reconhecimento facial usualmente tem as seguintes etapas: detecção de rostos, alinhamento (alinhar faces em diferentes posições), extração de características (distância entre os olhos, tamanho da orelha, espessura dos lábios, etc), baseado nessas características é feita a classificação. Estes atributos podem ser obtidos através de um processamento “tradicional” visando uma extração explícita, e de certo modo “manual”, de cada um deles, ou, podem ser detectados através de extratores automáticos de atributos (geração automática de *feature maps*).

2.4.1 Detecção de Rostos

Detectar rostos é, obviamente, o primeiro passo no processo de reconhecimento facial. Uma excelente implementação para essa abordagem com relação a custo computacional e resposta em tempo real é o algoritmo de OpenFace (OPENFACE). Esse algoritmo é uma implementação que utiliza a linguagem Python junto com framework Torch de reconhecimento facial com Redes Neurais Convolucionais e baseado no FaceNet (sistema de reconhecimento de faces da Google). O Torch é um framework com amplo suporte para algoritmos de aprendizagem de máquinas que colocam as GPUs (Graphic Processing Unit) em primeiro lugar. É fácil de usar e eficiente, graças a uma linguagem de script fácil e rápida, chamada LuaJIT e uma implementação C / CUDA subjacente. O OpenFace realiza o processamento em tempo real das imagens sendo que pessoas são adicionadas e removidas do sistema, o qual deve ser treinado com os novos dados a cada vez. Além destes, tem-se ainda os ambientes de desenvolvimento do Caffé (CAFFE), o Theano(THEANO), Cuda-convnet, cuDNN da Nvidia (NVIDIA) e o Tensorflow(TENSORFLOW)

2.4.2 Alinhamento

Faces em posições diferentes são vistas diferentemente pelo computador. Assim, é necessário alinhá-las para que tenham os olhos, narizes, bocas, etc, na mesma posição possibilitando, assim, a extração de características qualitativamente.

Neste trabalho serão utilizados os 68 pontos de referência (*landmarks*) desenvolvidos por (KAZEMI, V.; SULLIVAN, 2014). Esses pontos foram gerados baseados nas faces do *ibug dataset*. O algoritmo de (KAZEMI, V. e SULLIVAN, 2014) utiliza de uma composição (*ensemble*) de árvores de regressão (*regression trees*) para estimar a posição dos *landmarks* de um conjunto de intensidade de pixels esparsos.

Essa composição de árvores de regressão é aprendida baseada em *gradient boosting* minimizando a soma dos erros quadrados.

Figura 2.7 – Exemplo dos 68 pontos de referência (landmarks) gerados pela biblioteca Dlib para posterior alinhamento de faces.

Fonte: <<https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/>> acessado 31/07/2017

A árvore de regressão é um modelo de predição não linear. A árvore é uma expansão da regressão linear onde ao invés de ter um modelo global do problema (regressão linear) divide-se o problema em partes de menor complexidade, empregando modelo mais simples. Assim, cada um dos nós folha pode ser mais facilmente modelado e predito. No decorrer do trabalho será descrito com mais detalhes como é feito o cálculo dos *landmarks*.

2.4.3 Extração de Características

Diversas abordagens podem ser utilizadas para a extração de características das faces para a elaboração de um sistema de reconhecimento facial sendo que diversos métodos demandam diferentes características (TOLBA A.; EL-BAZ, 2006). Atualmente, o melhor classificador de faces faz uso das redes neurais profundas(Deep

Learning). Não é necessário que o desenvolvedor do sistema pense nas características a serem extraídas como distância entre os olhos, tamanho da orelha, espessura dos lábios, etc, pois é a rede neural que decide quais características melhor classificam os indivíduos. O sistema proposto neste trabalho, baseado no OpenFace, extrai 128 características da imagem e baseado nessas características é feita a classificação. Portanto, as características extraídas neste trabalho são desconhecidas.

Tendo as características, o sistema realiza a classificação com o uso de um classificador, onde pode ser adotado o SVM, por exemplo.

2.4.4 Treinamento

Durante o treinamento as 128 características são aprimoradas, comparando as características entre as diversas classes (pessoas) a serem reconhecidas. Supondo-se duas classes existentes, quando uma imagem é treinada pela rede, são geradas suas 128 características. Essas características são então comparadas as características das duas classes existentes e sabendo-se a classe da imagem treinada (treinamento supervisionado) aprimora-se com o ajuste dos pesos da rede, para que as características entre as imagens da mesma classe sejam mais semelhantes e as características das classes distintas sejam mais diferentes. O OpenFace, assim como FaceNet e outros sistemas de reconhecimento facial, utilizam SVM ou Redes Neurais para classificação a partir das características derivadas de uma Rede Neural Convolucional.

Existem também diversas bases de dados de faces disponíveis na Internet para o treinamento das redes neurais e para testes de performance do algoritmo de aprendizado e de reconhecimento facial. Neste trabalho foi feito o uso específico da base de dados LFW, adicionando novas faces à base, para testar o reconhecimento de

faces específicas de pessoas novas que se deseja adicionar ao sistema de reconhecimento.

A Tabela 2.1 apresenta várias bases de dados de rostos de pessoas para serem usadas em reconhecimento de faces, porém nem todas são úteis e adequadas para o trabalho em questão. Algumas destas bases são extremamente grandes e propiciam obter apenas uma baixa acurácia (pois visam reconhecer um grande número de faces diferentes), além do que necessitam de hardwares específicos (placas de vídeos potentes e dedicadas) para o seu adequado treinamento. Destacamos que neste trabalho visamos o uso de uma solução de relativo baixo custo, e portanto, o sistema deve poder ser utilizado considerando-se um equipamento de custo não muito elevado.

Tabela 2.1: datasets de reconhecimentos de face

Dataset	Classes	Total de Imagens	Estado
LFW	5.749	13.233	Público
WDFRef	2.995	99.773	Público(somente as características)
CelebFaces	10.177	202.599	Privado
SFC	4.030	4.400.000	Privado
CACD	2.000	163.446	Público(parcialmente)
CASIA-WebFace	10.575	494.414	Público

Fonte: <http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html> acessado em 20/01/2018

Todas as bases possuem milhares de pessoas para treinar e classificar, porém há classes que possuem poucas imagens e classes que possuem centenas de imagens onde as bases de dados não possuem classes balanceadas. Isso certamente faz com que a rede seja treinada de forma não muito adequada, levando a erros de classificação, pois o treino será mais bem elaborado para classes com muitas imagens na base de treino e mal condicionado para classes com poucas imagens.

2.5 Reconhecimento de Vida

Ao trabalhar com sistemas biométricos, seja reconhecimento facial, reconhecimento de íris, impressão digital, entre outros, intrinsecamente há uma grande possibilidade de falha de segurança no sistema – em que é necessário detectar se o usuário é uma pessoa ou não (tentativa de fraude). Sabe-se que é possível reproduzir uma digital, ou mesmo uma face, neste caso usando uma simples fotografia em seu lugar, e o sistema não pode se enganar a ponto de identificar como verdadeira a presença de uma pessoa se estiver sendo usada apenas uma foto dela. Neste trabalho, no uso de um sistema de reconhecimento facial haverá a necessidade de proteção contra ataques por fotos e vídeos. Assim, é necessário que o sistema seja capaz de detectar vida, ou seja, detectar que a pessoa está realmente presente, “ao vivo” (*proof of life*), junto ao sistema, quando esta pessoa estiver tentando acessar o ambiente protegido pelo sistema. Encontra-se na literatura diversas abordagens de detecção de vida (*liveness detection*) (BOULKENAFET Z.; KOMULAINEN, 2008), (BAO W.; LI, 2009), (TAN X.; LI, 2010), (WEN D.; HAN, 2015), entre outros trabalhos.

Análises de medidas de qualidade de novas técnicas *anti-spoofing* baseada em 1600 amostras de imagens de íris reais e falsas provaram que alguns métodos podem ter um potencial muito alto como um esquema de proteção eficaz contra ataques diretos (GALBALLY, 2012). No entanto, neste trabalho usamos uma câmera comum, e o reconhecimento facial será baseado na face da pessoa e não em leituras específicas da íris ou de digitais, por exemplo.

Os trabalhos de GHIANI (2012), apresentaram um novo conjunto de características, baseado na quantização de fase local (LPQ) de imagens de impressões digitais. O método LPQ é bem conhecido por ser insensível aos efeitos de desfoque, por isso acredita-se que ele poderia ser útil para detectar as diferenças entre uma

impressão digital viva e uma falsa, devido à perda de informações que podem ocorrer durante o processo de fabricação da réplica. Porém, novamente, este tipo de abordagem depende de um equipamento específico adicional para a leitura das impressões digitais.

Neste trabalho o reconhecimento de vida foi avaliado pela análise do espectrograma da voz dos usuários.

2.5.1 Espectrograma da Voz

Medidas acústicas da voz têm recebido atenção considerável por serem de natureza não invasiva e relativamente aplicáveis em pesquisas relacionadas à fala e em reconhecimento de fala.

O espectrograma de voz é a tradução de padrões sonoros em padrões visuais gráficos que refletem a frequência fundamental (f_0) e os harmônicos de uma voz, bem como as zonas de forte intensidade dos sons da fala, utilizando uma análise matemática denominada “Análise de Fourier”, que determina os componentes da onda sonora. Este método faz a dissecação da onda acústica em seus componentes básicos, ou seja, expõe as frequências e as respectivas amplitudes que a constitui (BAKEN, R.J.; ORLIKOFF, R.F., 2000). Desta forma, o espectrograma pode ser definido como um gráfico que mostra as intensidades relativas dos componentes das ondas sonoras, indicadas pelo escurecimento ou coloração de faixas de frequência no eixo vertical em função do tempo no eixo horizontal.

A entrada de áudio é um componente que usualmente está integrado junto às câmaras do tipo webCam, sendo um dispositivo de baixo custo e de fácil acesso. Na última fase do projeto foi aprimorada a acurácia do reconhecimento de um indivíduo

considerando-se o uso do reconhecimento de voz para o caso da análise de verificação da autenticidade da pessoa reconhecida, portanto é considerada uma complexidade de validação do tipo 1:1, ou seja, é testada uma pessoa contra os dados de apenas uma pessoa - ela mesma (1 contra 1) a fim de validar o seu reconhecimento inicial (a identificação da pessoa já foi feita previamente pelo reconhecimento facial).

2.6 Tensorflow

Segundo o *whitepaper* do TensorFlow, este framework é uma ferramenta para escrever e executar algoritmos de aprendizado de máquina. Os cálculos são feitos em um grafo de fluxo de dados, no qual os “nós” são as operações matemáticas e as arestas são os tensors (matrizes de dados multidimensionais) que são trocados entre os nós. Um usuário constrói o grafo e escreve os algoritmos que são executados em cada nó. O TensorFlow cuida da execução do código assincronamente nos diferentes dispositivos, núcleos e threads.

O TensorFlow é executado tanto na CPU quanto em GPUs do desktop, servidor ou dispositivos móveis. Pode utilizar o *container* do Docker para fazer o *deploy* na nuvem. A versão de código aberto executa em máquinas isoladas e não em clusters.

O TensorFlow tem uma completa API em Python e interface C++ para construir e executar os grafos. Possui ainda API cliente em C. A Google incentiva a comunidade para escrever as interfaces para outras linguagens, as mais prováveis são Lua, R, Java, Go e JavaScript.

O Tensorflow possui implementações que são previamente disponibilizadas de diferentes Redes Neurais Convolucionais (CNNs e DCNNs) e que podem ser reaproveitadas em novos projetos de aplicações.

2.7 Considerações Finais

Portanto, para o reconhecimento de faces são utilizadas as Redes Neurais Artificiais Multicamadas, pois cada camada aprende uma característica específica como reconhecimento de olhos, nariz, boca, pele, cabelo e proporções áureas da face humana. Para isso são necessárias algumas etapas de processamento que vão desde a aquisição da imagem, pré-processamento (remapeamento, redução de ruídos e aumento de contraste para assegurar que as informações relevantes sejam detectadas e destacadas), extração de características (detecção de bordas, de cantos ou de pontos), detecção e segmentação (seleção de regiões de interesse) e processamento de alto nível (estimativa de parâmetros e classificação de objetos por SVM).

Aliado a isso tudo, será aplicado um tipo específico de Aprendizado de Máquina, o aprendizado supervisionado, no qual será preparada, antecipadamente, uma base de treino com as classificações já estabelecidas para que o algoritmo de Rede Neural Artificial possa ser treinado. Após o treinamento da rede e o ajuste dos parâmetros o algoritmo passará por uma base de testes para verificar se a rede foi bem treinada ou não.

Após o reconhecimento de face, será aplicada a prova de vida, na qual será avaliado se o usuário é realmente quem ele diz ser (confirmação da imagem da face que foi reconhecida) pela análise do espectrograma da voz dos usuários. Nessa última etapa do projeto a acurácia será mais aprimorada, dando maior confiança para a identificação do indivíduo e aumentando a segurança do sistema.

Existem vários trabalhos nessa área de reconhecimento de face com o uso de Redes Neurais Artificiais, como será descrito no próximo capítulo, porém poucos trabalhos incluem métodos adicionais e associados de prova de vida.

CAPÍTULO 3

TRABALHOS RELACIONADOS

Existem alguns trabalhos que estão mais diretamente relacionados com a abordagem proposta neste projeto, seja pelo tema abordado de reconhecimento facial e/ou de voz, ou pelo tipo de técnicas e abordagens empregadas nos mesmos.

O trabalho de PAULA e SILVA, (2017), propõe e avalia um sistema embarcado e inteligente (hardware para abertura/trancamento de portas) de reconhecimento facial para o controle de acesso, sendo composto de um painel web para administração do sistema, um sistema de reconhecimento facial juntamente com um sistema de reconhecimento de vida. Utilizou-se, neste trabalho, de sistemas embarcados (autônomos) e microcontroladores (Raspberry Pi e Arduino), conectados com uma câmera de vídeo e servidores de dados via Wi-Fi, realizando desta forma um processamento autônomo para o reconhecimento facial baseado em técnicas de *Deep Learning*. A análise do projeto consistiu na avaliação da precisão do sistema (Matriz de Confusão do classificador), especialmente se tratando de falso positivos onde pessoas não autorizadas adentram o ambiente protegido. Além disso também são avaliados tempos de execução e precisão total do sistema. Os datasets avaliados foram o *Labeled Faces in the Wild dataset* (LFW) (HUANG et al., 2007) para reconhecimento facial e *CASIA Face Anti-Spoofing Database* (ZHANG et al., 2012) para detecção de

vida. O Reconhecimento de Vida desse trabalho foi através da análise de textura da imagem, vectorscope, fluxo óptico e aplicação de luz infravermelho, sendo portanto, baseado exclusivamente em imagens da face. Os resultados finais obtidos neste trabalho indicam a necessidade de um aprimoramento da solução embarcada (capacidade de processamento local limitada, com tempo de resposta bastante elevado) e busca por melhorias na acurácia e confiabilidade do reconhecimento das faces).

Na pesquisa realizada por MINATO, CORAL e GARCIA (2017), é abordada a implementação de um protótipo em um hardware integrado, do algoritmo Viola-Jones, fazendo o uso de uma Rede Neural Artificial do tipo *Multilayer Perceptron* (MLP) utilizando o algoritmo *Backpropagation* para aprimorar o desempenho da detecção de faces em condições difíceis. O protótipo abordou 4 etapas de desenvolvimento, sendo elas captura da imagem por uma Webcam, depois um pré-processamento da imagem obtida (transformações para normalizar a imagem), seguida pela submissão ao algoritmo de Viola-Jones em paralelo ao treinamento das *harr-cascades* e recolhimento dos resultados do conjunto de teste. Esse projeto utilizou o banco de faces do Dr. Libor Spacek. A aplicação desenvolvida nesta pesquisa apresentou uma boa taxa de acurácia considerando o tempo de treinamento limitado e a quantidade de amostras treinadas. Com o intuito de contornar os problemas de iluminação foi aplicada uma RNA(Rede Neural Artificial) em conjunto com o algoritmo de Viola-Jones, entretanto durante a fase de testes foi constatado que o *Raspberry Pi* não possuía poder computacional para o treinamento de uma RNA, entretanto em testes feitos com um hardware mais robusto o software desenvolvido executou com sucesso o treinamento da rede.

No trabalho de XAVIER SERRA, 2017, denominado de *Face Recognition Using Deep Learning*, foi criado um sistema de reconhecimento facial com o uso da CNN(Convolutional Neural Network) para extrair as características faciais. Para treinar

sua rede e obter características dela foi utilizado o DeepFace (framework de reconhecimento facial do Facebook). Esse trabalho utilizou imagens da base LFW (HUANG et al., 2007), Casia (Yi et al., 2014), CACD (Chen, Chen e Hsu, 2014) e FaceScrub (Ng e Winkler, 2014). O sistema, primeiramente, direciona as faces das pessoas para a frente e exclui o plano de fundo da imagem, para centralizar o rosto da pessoa na imagem e então compara as faces entre si. Os resultados da pesquisa trouxeram uma acurácia entre 90% a 95%, muito satisfatórios para a aplicação em questão.

Na área de espectrograma de voz, os estudos de GILLIER, R. (2011), denominado *O Disfarce da Voz em Fonética Forense*, compararam indivíduos com vozes normais com indivíduos que utilizaram disfarces na voz, para mostrar o quão robusto e preciso é o reconhecimento de vozes com disfarces. A análise dos dados relativos à voz normal serviu, por um lado, para verificar a eficácia dos parâmetros na distinção de falantes e, por outro, como ponto de comparação para observar a influência dos disfarces nos vários parâmetros acústicos. Os resultados obtidos demonstraram que (i) os parâmetros de identificação de falantes e das formas das palavras são eficientes para a discriminação de indivíduos; (ii) os disfarces não atuaram de forma idêntica; (iii) nem todos os disfarces são eficazes, pois alguns não alteram as frequências dos parâmetros; (iv) a robustez de cada parâmetro está diretamente relacionada com o tipo de disfarce; (v) o factor falante desempenha um papel importante na performance do disfarce; (vi) o efeito do disfarce não é homogêneo entre as várias vogais; (vii) mesmo no disfarce mais intrusivo, é possível recuperar marcas específicas de cada falante. Em suma, esse trabalho pretende (i) aprofundar os conhecimentos sobre a robustez dos parâmetros de identificação de falantes face ao disfarce; (ii) contribuir para o desenvolvimento dos estudos na área da Fonética Forense e da Acústica Forense.

Nos trabalhos de BARAVIEIRA, P. B. (2016), o uso da Inteligência Artificial baseado em Redes Neurais Artificiais na identificação, e graduação da rugosidade e da soprosidade da voz, apresentou excelente confiabilidade no Coeficiente de Correlação Intraclasse ($CCI > 0,80$), com resultados semelhantes à concordância interjuízes. Foram selecionadas 123 vozes, desde neutras até com presença em grau intenso de rugosidade e/ou soprosidade, do banco de dados da Clínica de Fonoaudiologia da Faculdade de Odontologia de Bauru (FOB/USP). Foi realizada avaliação perceptivo-auditiva pela escala visual analógica de 100 mm e pela escala numérica de quatro pontos; extração de características do sinal de voz por meio da Transformada Wavelet Packet e dos parâmetros acústicos: jitter, shimmer, amplitude da derivada e amplitude do pitch; e validação do classificador por meio da parametrização, treino, teste e avaliação das redes neurais artificiais. Dessa forma, a rede neural artificial revela-se como uma metodologia promissora de avaliação vocal, tendo sua maior vantagem a objetividade na avaliação.

Os trabalhos de PETRY, A. (1999), utilizaram técnicas de processamento digital de sinais para a identificação automática de pessoas pela voz. Com a aquisição do sinal vocal, seguida pela detecção automática dos limites das palavras, pelo pré-processamento, extração das características e classificação dos padrões vocais obteve-se uma performance de 90% de aceitação das pessoas com direito de acesso e 96% de rejeição de impostores. O que mostra a viabilidade de diversas aplicações de autenticação de pessoas pela voz.

O trabalho de GIACOMO, A.A; CELSO, C.C; MAURÍCIO J.A.B, 2009, visa o reconhecimento de padrões de voz dividido em duas partes, sendo elas: primeiro, a análise da voz gravada através do espectro dinâmico dado por uma ferramenta matemáticas mais robusta, a Transformada em Ondeleta (TO), onde é possível detectar as principais frequências; e segundo, ao término da primeira parte, foi caracterizado quais as frequências que são importantes em cada gravação, formando

com isso um padrão característico da voz de uma pessoa, podendo agora ser utilizado na comparação com outros padrões armazenados. Esse trabalho mostra a identificação de pessoas por meio da análise das ondas sonoras transformadas pela TO e não por meio da análise do espectrograma da voz, ou seja, uma forma diferente de análise da voz.

Outra área de aplicação da análise da voz é na área da robótica, na qual comandos de voz são transformados em ações para os robôs. O trabalho de SAMPAIO DE PAULA, G.R. (2011), exemplifica o caso, no qual comandos de voz dizendo “direita” e “esquerda” são treinados na rede neural (totalizando 50 arquivos de áudio de cada um dos dois comandos). Após a rede estar devidamente treinada e pré testada por uma rotina de testes aleatórios implementada no MatLab, os sinais de voz foram direcionados à comparação de padrões que ocorre na rede neural para que a rede dê uma classificação para o padrão. Neste caso “direita” ou “esquerda. Com este conjunto a rede neural mostrou um comportamento na classificação próximo do ideal, atingindo mais de 90% de acerto

Outro trabalho que se aproxima desse é o de AMARAL, M.A.; BARRIVIERA, R.; COTRIN TEIXEIRA E, (2004), que faz uso de um agente inteligente para capturar e processar comandos de voz com o intuito de realizar uma ação. Um agente inteligente é um dispositivo eletrônico que possui autonomia e executa tarefas, estando sempre vinculado a uma ação disparada automaticamente ou a uma ação externa manual a ser executada. O trabalho relata o desenvolvimento de um Software de Reconhecimento de Voz para Automação Residencial que utiliza arquitetura baseada em agentes inteligentes. O objetivo desta aplicação é auxiliar e proporcionar mais independência a pessoas portadoras de necessidades especiais, como pacientes que possuem um alto grau de lesões, com movimentos apenas acima do pescoço. O sistema desenvolvido é composto de dois programas, um para reconhecimento de voz e o outro para controle do processo de automação residencial. Os programas são interligados por meio de

uma interface de software, e todo o sistema possui um circuito acoplado à saída paralela do computador. Este circuito ativa ou desativa oito pontos que estão conectados a qualquer aparelho eletro-eletrônico, através de comando de voz.

Outra área de aplicação do reconhecimento de voz é na indústria, feito por ALVARENGA, R.J. (2012). O objetivo específico do trabalho foi conceber e desenvolver um sistema de reconhecimento de voz, capaz de identificar comandos de voz, independentemente do locutor, com o uso dos comandos “direita”, “esquerda”, “frente”, “atrás”, “pare”, e “desligue”. A finalidade precípua do sistema é controlar movimentos de robôs, com aplicações na indústria e no auxílio de deficientes físicos. Utilizou-se a abordagem da tomada de decisão por meio de uma rede neural treinada com as características distintivas do sinal de fala de 16 locutores. A rede neural empregada como classificador foi um Perceptron Multicamadas (MLP), treinado pelo algoritmo *Backpropagation*. Várias experimentações foram realizadas para a escolha de limiares, valores práticos, *features* e configurações da rede neural. Os resultados foram considerados muito bons, alcançando uma taxa de acertos de 89,16%, sob as condições de pior caso da amostragem dos comandos.

O reconhecimento de voz também é aplicado em jogos, como foi feito por TORRES, E.S.; SILVA, L.J.; SANTOS R.M.; MACEDO H.T.; MATOS L.N. 2016. Uma das principais dificuldades no desenvolvimento de aplicações com reconhecimento de fala é a distorção de áudio causado pelo ruído do ambiente. Portanto, não é comum encontrar jogos com reconhecimento automático de fala (ASR), especialmente para plataformas móveis. Este artigo descreve o desenvolvimento de dois jogos com interface 2D, que utilizam um sistema de reconhecimento on-line de fala. A arquitetura híbrida de reconhecimento usa uma Rede Neural Convolucional (CNN) com Modelo Escondido de Markov (HMM) para reduzir o problema do ruído ambiente. O primeiro jogo, intitulado de "Cálculo de Aventura" é um RPG no estilo de turnos e é focado em crianças de 7 a 10 anos, cujos desafios são propostas matemáticas. O outro, intitulado

"Break Aracajú", é baseado no clássico Breaker Ball. Foi comparada a acurácia de reconhecimento em variados ambientes com ruídos aditivos. Resultados de experimentos com voluntários mostram que o ruído exerce pouca influência sobre o reconhecimento. O reconhecimento com o modelo CNN-HMM se mostrou factível para uso em jogos em dispositivos móveis.

3.1 Considerações Finais

Cabe destacar que os trabalhos citados acima em relação ao reconhecimento de voz, em sua maioria buscam justamente “generalizar” o reconhecimento de palavras e de comandos vocais, ignorando as diferenças entre a voz dos diferentes falantes. Este é o caso do Google Voice Recognition (Cloud Speech-to-Text – conhecido pelo comando “Ok Google”), que permite traduzir fala em textos, e que objetiva reconhecer a fala independentemente do usuário que está falando. No caso deste trabalho, desejamos reconhecer uma frase específica (senha) falada por um usuário, e o reconhecedor deve diferenciar a voz das diferentes pessoas (tanto em relação às características de voz, como o timbre), como também, deve identificar e diferenciar quando uma mesma frase é dita por pessoas diferentes. Portanto, apesar de encontrarmos na literatura muitos trabalhos avançados sobre reconhecimento de fala, poucos se propõem a reconhecer o falante e a fala ao mesmo tempo. Em termos de sistemas de autenticação e de controle de acesso/segurança, o reconhecimento preciso de quem está falando é essencial para o sucesso do sistema.

Por outro lado, no que se refere ao reconhecimento facial, muitos trabalhos buscam reconhecer um grande número de pessoas diferentes, onde é natural que tais sistemas “aceitem” um certo nível de erros na forma de Falsos Positivos. Em um sistema de autenticação e controle de acesso/segurança os Falsos Positivos são bastante perigosos pois podem dar acesso indevido a pessoas não autorizadas. Além disto, neste trabalho é importante que os usuários cadastrados e autorizados possam

ser reconhecidos com alto grau de confiabilidade, mesmo quando são apresentadas novas faces de pessoas desconhecidas (que devem ser rejeitadas). Sistemas genéricos de reconhecimento facial, treinados com grandes bases de dados de faces, não oferecem garantias de como estes irão se comportar em relação a apresentação de faces desconhecidas. Finalmente, poucos sistemas realizam uma etapa de “prova de vida”, que obrigue ao usuário a se identificar pessoalmente, pois uma forma relativamente simples de enganar um reconhecedor de faces é apresentando uma fotografia da pessoa autorizada. O reconhecimento de uma frase específica, falada por um usuário em particular, faz com que o sistema seja mais robusto a tentativas de enganá-lo.

CAPÍTULO 4

METODOLOGIA

4.1 Contextualização

Este projeto está voltado para um mercado de segurança de acesso em residências ou locais de acessos restritos e controlados (como escritórios, laboratórios, galpões e outros). A demanda por soluções de segurança é cada vez mais exigente em termos de robustez, com aplicações inteligentes e que exigem mais fatores de segurança para identificar as pessoas. Fatores que podem incluir o reconhecimento de face, de digital, de íris ou de voz.

As aplicações no mercado possuem diferentes complexidades, no qual as aplicações mais comuns são dos seguintes tipos:

Tipo I - Reconhecimento com uma relação N:N, ou seja, deseja-se poder identificar (reconhecer) “N” diferentes usuários, considerando que qualquer um destes “N” usuários pode se apresentar ao sistema e o sistema irá automaticamente identificar e reconhecer qual usuário está se apresentando. Este tipo de sistema usualmente considera uma base de treino de tamanho bastante grande ($N \gg 100000$) para treinar, com imagens não padronizadas (p.ex. Imagens de câmeras de segurança em aeroportos), e então usar uma base de teste grande com $N > 100000$ com todas as

classes a serem reconhecidas. Isto geralmente converge para resultados de menor acurácia (<50%) e com um alto índice de falsos positivos, pois tem que reconhecer várias pessoas entre grupo bem amplo de pessoas;

Tipo II - Reconhecimento com uma relação de 1:N, na qual $N < 20$, ou seja, deseja-se poder identificar (autenticar/verificar) se um usuário específico é realmente quem diz ser, considerando um conjunto de “N” diferentes usuários cadastrados no sistema. Neste tipo de reconhecimento, a base de treino não ultrapassa 20 classes (pessoas) e possui um número limitado de imagens de treino para cada classe (usualmente 100 imagens por classe). Essa condição é empregada na aplicação em questão, voltada para criar um produto comercialmente viável.

Tipo III - Reconhecimento com uma relação de 1:1, ou seja, deseja-se poder identificar (autenticar/verificar) se o usuário é realmente quem diz ser, validando a identificação. O sistema reconhecedor é especializado em reconhecer apenas uma única classe, um único usuário, em que a saída é binária (p.ex. Sim, este é o usuário esperado; não, este não é o usuário esperado). Neste tipo de sistemas, uma face é capturada, usando um número limitado de imagens para a base de treino. Contudo, o sistema vai reconhecer apenas uma única face e não uma entre várias. Essa relação, deve garantir uma autenticação com alta acurácia, quase 100%, e aplicações deste tipo não devem tolerar erros do tipo falsos positivos (exigem taxas extremamente baixas de falsos positivos). Essa relação é adotada em aplicações como o “*unlock*” de tela do iPhone X e também é empregada na aplicação desse projeto.

Em resumo, podemos considerar como aplicações exemplo do Tipo I, um sistema que procura por uma lista de criminosos e terroristas em imagens de um circuito interno de TV (CCTF) de um aeroporto; como exemplo do Tipo II, um sistema que identifica um grupo restrito de usuários autorizados para acessar a uma instalação de acesso controlado; e como Tipo III, um sistema de biometria individualizado, que

verifica se um determinado usuário é quem diz ser e libera o acesso exclusivo para este usuário, como no caso do *unlock* de um *smartphone*.

O projeto foi criado para uma aplicação do cotidiano, que consiste em criar um sistema de segurança para controle de abertura e fechamento de portas de residências e laboratórios. Esse sistema realiza o reconhecimento de face e de voz da pessoa (por meio do espectrograma de voz) através de Deep Learning e Aprendizado de Máquina, utilizando a ferramenta conhecida do mercado e da academia, o Tensorflow.

4.2 Esquema do Projeto

O projeto foi dividido em 4 Redes Neurais (1, 2, 3 e 4), nas quais foram treinadas de forma específica e organizadas de forma sequencial para a aplicação, visto na Figura 4.1 abaixo.

Figura 4.1 - Esquema das Fases e Etapas do Sistema Proposto

Fonte: Autoria própria

A aplicação realiza a captura da imagem da face e do áudio do usuário, então cada uma delas é direcionada para sua respectiva rede: a imagem da face para a RN-2 e o arquivo do áudio para a RN-3. Se a resposta da RN-2 for igual à RN-3, então a aplicação segue para uma das redes da RN-4.

As fases do processo de reconhecimento e validação do usuário (Fig. 4.1) serão descritas resumidamente a seguir. Em todas elas foi utilizada a ferramenta Tensorflow, como framework de base das Redes Neurais.

- RN-1: A primeira rede neural foi treinada com o dataset LFW modificado, contendo de 2 a 5 classes escolhidas, com no mínimo 80 imagens por classe, pois é usualmente o valor considerado adequado de imagens requeridas no Tensorflow para se treinar uma rede neural. Nessa etapa, várias arquiteturas foram testadas e comparadas pelo resultado das acurácias. A arquitetura com a melhor acurácia foi selecionada para ser aplicada na etapa seguinte (RN-2). Nessa fase trabalhou-se na complexidade 1:N, com N entre 2 e 5.
- RN-2: A segunda rede neural foi treinada com uma base própria de imagens dos usuários, contendo imagens capturadas pelo próprio autor deste trabalho. Esta base conta com 5 classes para aplicar a rede neural com a arquitetura escolhida na RN-1. É nessa fase que se inicia o filtro de autenticação, no qual o sistema vai identificar um possível candidato a ser autenticado, dentre os presentes na base de treino. Nessa fase trabalhou-se na complexidade 1:N, com N entre 2 e 5.
- RN-3: A terceira rede neural foi treinada com uma base própria de espectrogramas de voz dos usuários criada pelo autor deste trabalho, com 5 classes para aplicar a rede neural com a arquitetura escolhida na RN-1. É mais uma etapa de autenticação do usuário, usando o espectrograma de voz obtido na captura do áudio pela aplicação. Nessa fase trabalhou-se na complexidade 1:N, com N entre 2 e 5.
- RN-4: A quarta rede neural foi treinada com 2 classes: uma tendo os espectrogramas de um único usuário e outra contendo os espectrogramas dos

demais usuários, misturadas em uma única classe. A RN-4 só é acionada quando os resultados das redes RN-2 e RN-3 forem iguais (para dar maior confiabilidade nos resultados), caso contrário a aplicação requer nova captura de imagem e áudio do usuário. Essa etapa tem complexidade 1:1 e determina o limiar de aceitação, ou seja, um valor que determina se a aplicação permitirá ou não a entrada do indivíduo no local desejado.

4.3 RN-1 - Escolha da melhor arquitetura para o reconhecimento de face e de vida

A primeira RN, RN-1, usou de 2 a 5 classes, cada uma contendo em torno de 100 imagens para treino e 20 para teste, obtidas da base de dados LFW. Foram usadas várias arquiteturas como a Inception (site TENSORFLOW) e 16 variações da arquitetura Mobilenet (site TENSORFLOW) (que varia de acordo com o peso do parâmetro e o tamanho das imagens inseridas). Na arquitetura Mobilenet são adotadas 4 variações do parâmetro de peso ('1.0', '0.75', '0.50', ou '0.25') e mais 4 variações do parâmetro de tamanho da imagem ('224', '192', '160', ou '128'). A combinação desses parâmetros forneceu 12 diferentes pares de parâmetros válidos (Mobilenet 1.0_224, Mobilenet 0.75_224, Mobilenet 0.5_224, assim por diante até o fim das combinações). Porém 4 variações desta rede não funcionaram, pois o Tensorflow não era compatível (Mobilenet_0.5_224, Mobilenet_0.5_192, Mobilenet_0.5_160 e Mobilenet_0.5_128).

A Tabela 4.1 mostra um modelo de como foram analisadas as arquiteturas, ou seja, todas as 9 arquiteturas válidas foram testadas separadamente para a base de 2 classes, depois para a base de 3 classes, então para 4 classes e por último para a base de 5 classes. Portanto, diversas redes neurais foram treinadas para se obter a melhor acurácia, em relação às bases.

O melhor resultado dessas arquiteturas foi utilizado na RN-2 (que será descrita mais adiante). Os resultados analisados em relação a RN-1 foram relativos ao grau de

acurácia (acerto) na classificação das faces, para cada arquitetura da rede e para os diferentes testes realizados com números distintos de usuários, variando de 2 a 5 usuários. Estes resultados foram sintetizados em uma tabela, no formato apresentado na Tabela 4.1, selecionando-se a configuração que obteve o melhor desempenho: a Mobilenet_1.0_128 para 5 classes.

Nº de classes na rede	Acurácia com arquitetura apenas o Inception	Acurácia com arquitetura mobilenet 1.0_224	Acurácia com arquitetura mobilenet 0.75_224	Acurácia com arquitetura mobilenet 0.5_224	Acurácia com arquitetura mobilenet 0.25_224	Acurácia com arquitetura mobilenet 1.0_192	...
2							
3							
4							
5							

Tabela 4.1: Modelo de formato de análise das arquiteturas Mobilenet.

A base de dados de faces LFW é composta por 5749 pessoas famosas, ou classes, e cada pessoa possui uma certa quantidade de imagens, algumas pessoas com muitas imagens e outras com poucas imagens. O Tensorflow realiza um trabalho eficiente com uma classe que tenha no mínimo 80 imagens. Na base LFW utilizada, poucas classes (pessoas) possuíam mais de 80 imagens, sendo elas Colin Powell, Donald Rumsfeld, George W. Bush, Tony Blair e Sama Rouhani. Sendo esse último o próprio autor do projeto, cuja face foi inserida propositalmente na base para verificar se a base aceita e reconhece corretamente classes novas. Assim foram usadas 100 imagens para cada classe da base de treino e 20 imagens para cada classe da base de teste.

A Tabela 4.2 mostra exemplos de como são as imagens usadas pela RN-1.

Tony Blair

George W. Bush

Donald Rumsfeld

Colin Powell

Sama Rouhani

Tabela 4.2: Exemplos de imagens utilizadas na RN-1.

A Tabela 4.1 mostra que após treinar a rede, aplicando-a na base de treino, a rede foi então testada com um novo conjunto diferente de imagens das classes na base de teste. Os resultados de cada elemento da base de teste foram colocados em uma matriz de confusão para serem analisados. Além disso, foram obtidos os chamados TOP 1 e TOP 2, ou seja, ao se analisar cada indivíduo da base de teste observou-se em qual posição ele ficou na classificação, estando entre os dois primeiros (TOP 2) ou em primeiro lugar (TOP 1) e então foi produzida a matriz de confusão para esses resultados

4.4 RN-2 - Reconhecimento de face com base própria com o uso da arquitetura vencedora da RN-1

A segunda RN, que chamaremos de RN-2, foi treinada com a própria base criada pelo autor, com imagens próprias e de familiares. O objetivo da RN-2 foi treinar uma rede com uma base voltada para a aplicação do projeto no mercado de segurança, pois, uma casa ou um laboratório que tenha o equipamento instalado precisa permitir a entrada de um conjunto de pessoas pré-definido (familiares no caso da casa e usuários no caso do laboratório). Para isso, o sistema precisa ter essas classes na sua base de dados e esta deverá ser bem treinada com essas classes. Essa rede tem 5 classes, cada uma delas contendo 100 imagens para cada classe da base de treino e 45 para cada classe da base de teste. A arquitetura utilizada na RN-2 foi aquela que obteve o melhor resultado de acurácia obtida na RN-1. A melhor arquitetura da RN-1 foi a Mobilenet com peso do parâmetro sendo 1.0 e o tamanho da imagem 128 (arquitetura Mobilenet_1.0_128).

Para ambas as redes (1 e 2) foram obtidos os chamados TOP 1 e TOP 2, ou seja, ao se analisar cada indivíduo da base de teste observou-se em qual posição ele ficou, na classificação da rede, estando entre os dois primeiros (TOP 2) ou em primeiro lugar (TOP 1).

4.4.1 Análise e Comparação das Imagens Utilizadas

Na RN-2 as imagens foram capturadas pelo mesmo equipamento, um celular da marca Lenovo, modelo K3 Note, com Sistema Operacional Android 6.0. São imagens da metade do tronco do indivíduo para cima, no formato jpg, de tamanho entre 2,5M a 4,0M, com altura e comprimento de 1920 pixels cada. Já as imagens da RN-1 (provenientes da base de faces LFW) eram imagens com rostos muito próximos e grandes, comparados às imagens da RN-2, também no formato jpg, com tamanho entre 11kb a 15kb, de resolução muito menor, com altura e comprimento de 250 pixels cada.

A Tabela 4.3 mostra exemplos de como são as imagens das RN-1 e RN-2.

Imagens da RN-1

Imagens da RN-2

Tabela 4.3: Modelo de imagens das RN-1 e RN-2.

4.5 Reconhecimento de Vida parte 1: RN-3

A terceira RN, que chamaremos de RN-3, foi treinada com os espectrogramas de voz da própria base criada pelo autor, ou seja, a própria voz e a dos mesmos familiares que também capturaram suas imagens para a RN-2. O objetivo da RN-3 foi treinar uma nova rede com os espectrogramas dos usuários, com o intuito de comparar os resultados das duas redes: RN-2 e RN-3, para saber se as duas redes identificaram o mesmo usuário, como mostra o esquema do projeto na Figura 4.1. Além disso, tentou-se simular, de forma mais fiel possível, o ambiente para a atuação da aplicação, cujos áudios são dos próprios frequentadores da residência ou do laboratório, que além de fazer o reconhecimento facial também fazem o reconhecimento de voz (reconhecimento de vida).

As gravações de voz foram realizadas, também, pelo mesmo equipamento no qual utilizou-se para capturar as imagens utilizadas para a RN-2: celular da marca Lenovo, modelo K3 Note, com Sistema Operacional Android 6.0. Os áudios estão no formato “.ogg”, cujo tamanho possui uma média de 100KB. É importante destacar, que a captura também poderia ser realizada por outros dispositivos, como por exemplo, uma câmera Webcam com microfone embutido

Foram capturados 150 áudios de cada indivíduo, falando a frase “A senha secreta é tigre”, sendo um total de 5 classes/pessoas, ou seja, 750 áudios. Todos os áudios foram submetidos a vários cálculos matemáticos, aplicando a chamada Transformada de Fourier - FFT (um módulo do próprio Tensorflow possui esta função) e então divididos em base de treino (100 imagens por classe) e base de teste (45 imagens por classe). A arquitetura utilizada na RN-3 foi aquela que obteve o melhor resultado de acurácia obtida na RN-1. A melhor arquitetura da RN-1 foi a Mobilenet com peso do parâmetro sendo 1.0 e o tamanho da imagem 128 (arquitetura Mobilenet_1.0_128).

Para que a rede fosse treinada de forma completa, simulando barulhos do dia a dia, foram utilizados áudios com diversas variações de voz, como por exemplo, a mesma pessoa com a voz nasal (para simular um resfriado), com a voz rouca (para simular falha na voz) e em forma de cochicho (para simular ambientes com barulho restrito). Foram realizados ruídos propositais no fundo do áudio (para simular barulhos da rua ou obras próximas à casa), ruídos que eram barulhos de plástico e de batidas secas em alguma tábua. Os áudios foram gravados em intervalos de silêncio antes e depois da gravação. Essas variações tendem a simular o dia a dia da aplicação, com ruídos no áudio e condições de saúde do usuário.

Para todas as redes (RN-1, RN-2 e RN-3) foram obtidos os chamados TOP 1 e TOP 2, ou seja, ao se analisar cada indivíduo da base de teste observou-se em qual posição ele ficou na classificação da rede, estando entre os dois primeiros (TOP 2) ou em primeiro lugar (TOP 1).

A Figura 4.2 mostra um modelo de como ficaram as imagens dos espectrogramas de voz da frase “A senha secreta é tigre.”

Figura 4.2: espectrogramas de voz da frase.

4.5.1 Reconhecimento de Vida parte 2: RN-4

É a quarta RN da aplicação, na qual só é acionada se as pessoas escolhidas nas RN-2 e RN-3 forem iguais, caso contrário, a aplicação será direcionada para uma nova captura da imagem da face e do áudio do usuário. No início da aplicação, após a captura da imagem da face e do áudio do usuário, cada um dos arquivos é direcionado para sua respectiva rede: a imagem da face para a RN-2 e o arquivo do áudio para a RN-3.

Essa última rede foi treinada com os espectrogramas de voz da própria base criada pelo autor, ou seja, é a mesma base da RN-3, porém, possui apenas duas classes: a do usuário selecionado nas redes RN-2 e RN-3 (que chamaremos de “classe usuário”) e outra classe com todos os outros usuários juntos (que chamaremos de

“classe outros”). Portanto, foram treinadas 5 redes (RN-4.1, RN-4.2, RN-4.3, RN-4.4 e RN-4.5), uma para cada pessoa, para se ter um grau ainda maior de acurácia dos resultados. A base de treino teve 50 imagens para a classe outros e 50 imagens para a classe usuário. O objetivo é aumentar a segurança do sistema e evitar os falsos positivos (acesso indevido).

Essa formação da rede foi escolhida para promover uma melhor autenticação, com complexidade 1:1, o que permitiu obter resultados com um grau de confiança e credibilidade muito maiores que as outras redes. Essa complexidade, associada ao reconhecimento de vida, garante um grande avanço da aplicação, pois está presente após duas outras grandes autenticações, de reconhecimento facial na RN-2 e espectrograma de voz na RN-3. Porém, essas duas redes possuem complexidade 1:N, o que não garante uma autenticação cuja acurácia seja sempre próxima de 100%, ou seja, não garante um elevado grau de confiança no resultado da aplicação.

Assim como na RN-3, os resultados da RN-4 foram colocados em uma matriz de confusão e analisados para uma verificação final se o indivíduo teria o acesso liberado ou não. Com isso foi possível fazer uma análise da complexidade de 1:1, com elevado grau de confiança. Com os resultados foi possível obter a acurácia para a classe do indivíduo e assim obter o melhor valor do limiar de aceitação para cada indivíduo em cada Rede Neural.

Com os resultados foi possível obter a melhor acurácia para cada classe e consequentemente obter o melhor limiar para a aplicação. Esse limiar é de extrema importância, pois é a partir desse valor que a aplicação irá permitir ou não a entrada da pessoa no local restrito.

4.6 Medidas dos Limiares

De acordo com a análise dos resultados obtidos pelas RNs, pode-se extrair a informação de que o limiar de acurácia, definitivo para o acesso do usuário à residência ou laboratório, é mais bem definido separadamente por cada usuário em cada RN e não um limiar único da rede para todos os usuários. O valor do limiar foi definido através de uma média ponderada das faixas de acurácia, no qual o peso é a quantidade de valores dos resultados das acurácias da base de teste.

Como demonstrativo de cálculo do limiar, tomemos como exemplo hipotético a Pessoa X para a RN-2, RN-3 e RN-4, cujos resultados das acurácias estão presentes na Tabela 4.4.

$$\text{Limiar RN-2 Pessoa X} = ((5*30)+(9*55)+(4*65)+(6*75)+(3*85)+(18*95))/45 = 73,77$$

$$\text{Limiar RN-3 Pessoa X} = ((2*30)+(8*55)+(7*65)+(12*75)+(10*85)+(12*95))/45 = 85,66$$

$$\text{Limiar RN-4 Pessoa X} = ((10*30)+(6*55)+(8*65)+(1*75)+(8*85)+(12*95))/45 = 67,66$$

Pessoa X	RN-2	RN-3	RN-4
< 50	5	2	10
entre 50% a 60%	9	8	6
entre 60% a 70%	4	7	8
entre 70% a 80%	6	12	1
entre 80% a 90%	3	10	8
entre 90% a 100%	18	12	12
LIMIAR	73,77	85,66	67,66

Tabela 4.4: Tabela hipotética dos limiares da Pessoa X para cada RN

Nota-se que para o cálculo pegou-se o ponto médio de cada faixa de valor, ou seja, os pontos 55, 65, 75, 85 e 95. Porém, para os valores abaixo de 50, pegou-se o valor 30, para que o sistema tenha uma confiança maior, impedindo o acesso de pessoas com baixa acurácia. No final, divide-se o valor por 45 (soma dos pesos da média ponderada), que é a soma da quantidade de valores de acurácia da Pessoa X na RN-2, ou seja, é a quantidade de imagens na base de teste.

Esse tipo de análise é importante para avaliarmos a performance de cada Rede Neural para cada usuário, podendo melhorar o treinamento de uma rede para certo tipo

de usuário, permitindo assim aumentar os níveis de confiança dos resultados e da autenticação da aplicação.

CAPÍTULO 5

RESULTADOS E DISCUSSÃO

Observou-se que quanto maior o número de classes, na base de treino, menores são os resultados de acurácia. Na RN-1, para 5 classes, o resultado maior de acurácia foi de 96%, obtida por apenas uma arquitetura, a `mobilenet_1.0_128`, adotou-se então essa arquitetura “vencedora” como a única arquitetura a ser utilizada nas RN-2, RN-3 e RN-4.

Já na RN-2, o uso da base própria voltada para a aplicação, fez com que os resultados fossem bem melhores. O Tensorflow classificou melhor as imagens de teste, demonstrando que a rede foi bem treinada e que o sistema é bem eficiente quando voltado para a aplicação. Os melhores resultados obtidos na RN-2 podem ser justificados pela resolução e tamanho das imagens, o que possibilitou melhor treinamento pela rede neural do Tensorflow.

5.1 Resultados da RN-1 - Escolha da melhor arquitetura para o reconhecimento de face e de vida

As acurácias obtidas na RN-1 foram muito interessantes, pois para 2 classes quase todas as arquiteturas tiveram 100% de acurácia (com exceção da `Mobilenet_0.25_160` que obteve 96%). Isso pode indicar overfitting, pois pode haver erros de medição ou mesmo fatores aleatórios.

Os gráficos de 5.1 a 5.4 comparam os resultados das arquiteturas com 2, 3, 4 e 5 classes.

O Gráfico 5.1 mostra um panorama dos resultados das acurácias das arquiteturas analisadas com 2 classes. Observa-se que todas obtiveram o valor máximo (100%), com exceção da Mobilenet_0.25_160, que obteve um valor próximo do máximo, 96%.

Gráfico 5.1: RNs aplicadas para classificar 2 classes. Cada coluna se refere às arquiteturas da rede..

A rede treinada com 3 classes teve, de um modo geral, os valores das acurácias variando de 87% a 97%. As melhores arquiteturas que obtiveram 97% de acurácia, foram a mobilenet_1.0_224, mobilenet_0.75_192, Mobilenet_0.25_192 e a Mobilenet_0.75_128. As piores arquiteturas foram a Mobilenet_0.75_160 e a Mobilenet_0.25_160 com 87,4% de acurácia.

No Gráfico 5.2 observa-se que poucas arquiteturas obtiveram bons resultados (97%).

Gráfico 5.2: RNs aplicadas para classificar 3 classes. Cada coluna se refere às arquiteturas da rede.

Para 4 classes, os resultados foram piores, porém ainda elevados, variando de 89% a 94.7%. As melhores arquiteturas, representadas pelo Gráfico 5.3, que obtiveram 97.4% de acurácia, foram a Mobilenet_1.0_224, Mobilenet_0.25_224 e a Mobilenet_1.0_192. As piores arquiteturas foram a Inception com 89% de acurácia, a Mobilenet_0.75_160 e a mobilenet_0.25_160, ambas com a acurácia de 89,5%.

O Gráfico 5.3 mostra muito bem a comparação da acurácia das arquiteturas para a rede classificar 4 classes.

Gráfico 5.3: RNs aplicadas para classificar 4 classes. Cada coluna se refere às arquiteturas da rede.

Para 5 classes na base de treino, os valores das acurácias foram um pouco melhores do que para 4 classes, pois, os resultados variaram de 88% a 96%. Somente uma arquitetura obteve o melhor resultado, tendo 96% de acurácia, sendo ela a `mobilenet_1.0_128`. Já as piores arquiteturas foram a Inception e a `mobilenet_0.25_128`, ambas com 84% de acurácia.

Observou-se que quanto maior o número de classes, na base de treino, menores são os resultados de acurácia. Para 5 classes, o resultado maior de acurácia foi de 96%, obtida por apenas uma arquitetura, a `mobilenet_1.0_128`, adotou-se então essa arquitetura como a “vencedora”, como a única arquitetura a ser utilizada nas RN-2, RN-3 e RN-4.

O Gráfico 5.4 mostra, de forma clara a comparação da acurácia das arquiteturas para uma rede classificar 5 classes.

Gráfico 5.4: RNs aplicadas para classificar 5 classes. Cada coluna se refere às arquiteturas da rede.

Com os resultados da base de teste, foram criadas duas matrizes de confusão, sendo o alvo(usuário inserido na rede) o mais bem classificado (denominado TOP 1) e o alvo sendo o segundo mais bem classificado (denominado TOP 2). Essas matrizes são mostradas nas Tabelas 5.1 e 5.2.

A Tabela 5.1 indica a quantidade e a porcentagem das 20 imagens de teste(de cada usuário) que passaram pela rede para a classificação feita pelo Tensorflow. Os resultados mostraram que para Colin Powell 80% das tentativas deram positivo e consequentemente 20% deram falso negativo. Para Donald Rumsfeld obteve-se 55% positivo e 45% de falso negativo. Para George W. Bush, os resultados mostraram 75% de positivo e 25% de falso negativo. Para Sama Rouhani, o sistema apresentou 100% de positivo. Para Tony Blair, observou-se 75% de positivo e 25% de falso negativo. Os resultados mostraram que o índice de falso negativo ainda é alto. Considerando o sistema voltado para uma aplicação prática do dia-a-dia, os resultados são ruins, pois,

o sistema tem alto índice de falso negativo, ou seja, de impedimento de acesso dos próprios membros ao local desejado..

Porém, é interessante notar que a classe inserida (a do próprio autor), foi a que obteve os melhores resultados. Isso é possível devido à melhor resolução das imagens do autor, em comparação com as imagens da base LFW, as do autor foram imagens de tamanho entre 2,5M a 4,0M, com altura e comprimento de 1920 pixels cada. Já as imagens da base LFW com tamanho entre 11kb a 15kb, de resolução muito menor, com altura e comprimento de 250 pixels cada.

TOP 1						
	Colin Powell	Donald Rumsfeld	George W. Bush	Sama Rouhani	Tony Blair	%
Colin Powell	16	-	2	-	2	80
Donald Rumsfeld	4	11	4	-	1	55
George W. Bush	2	-	15	-	3	75
Sama Rouhani	-	-	-	20	-	100
Tony Blair	1	2	2	-	15	75

Tabela 5.1: matriz de confusão da base de teste para TOP 1 da RN-1. São 20 imagens da base de teste LFW, classificada em 4 classes mais a imagem do próprio autor

A Tabela 5.2 mostra que os resultados são melhores para o TOP 2, em comparação com os resultados do TOP 1, ou seja, o sistema apresenta melhores resultados de classificação quando o alvo está entre os dois primeiros da lista de classificação. Porém, os resultados ainda não são satisfatórios para uma aplicação prática voltada ao mercado.

TOP 2						
	Colin Powell	Donald Rumsfeld	George W. Bush	Sama Rouhani	Tony Blair	%
Colin Powell	20	-	-	-	-	100
Donald Rumsfeld	2	17	-	-	1	85
George W. Bush	1	-	18	-	1	90
Sama Rouhani	-	-	-	20	-	100
Tony Blair	-	2	1	-	17	85

Tabela 5.2: matriz de confusão da base de teste para TOP 2 da RN-1. São 20 imagens da base de teste LFW, classificadas em 4 classes mais a imagem do próprio autor

Os resultados da Tabela 5.2 mostram que no caso do Colin Powell 100% das tentativas deram positivo. Para Donald Rumsfeld foram 85% positivo e 15% de falso negativo. Para George W. Bush foram 90% de positivo e 10% de falso negativo. Para Sama Rouhani foram 100% de positivo. Para Tony Blair foram 85% positivo e 15% de falso negativo. Os resultados mostraram que o índice de falso negativo dos resultados é muito alto. Entretanto, os resultados para TOP 2 já são bem melhores, se comparados ao TOP 1. Em se tratando de uma aplicação prática do dia-a-dia, os resultados continuam insatisfatórios, pois, o sistema continua tendo alto índice de impedimento de acesso dos próprios membros da casa ou do laboratório.

O Gráfico 5.5 mostra de forma clara, os resultados das acurácias obtidas nos TOP 1 e TOP 2 da Rede Neural, utilizando a base LFW com 5 pessoas, com a arquitetura mobilinet_1.0_128. Pode-se observar que todas as classes no TOP 2 (tom

escuro) obtiveram bons resultados, se comparados aos resultados do TOP 1 (tom claro), chegando a 100% em dois casos (Colin Powell e Sama Rouhani).

Gráfico 5.5: gráfico de comparação dos TOP 1 e TOP 2 da RN-1 (da base LFW com arquitetura mobilinet_1.0_128)

5.2 RN-2 - Reconhecimento de face utilizando a própria base e a arquitetura mobilenet_1.0_128

A base própria envolveu familiares como esposa, mãe e amigos, cujos nomes foram preservados e substituídos por pessoas A, B, C, D e E. Portanto, essa base contém 5 classes. Em cada classe, existem 100 imagens para a base de treino e 45 imagens para a base de teste.

A Tabela 5.3 mostra os resultados da matriz de confusão para TOP 1, ou seja, apresenta quais indivíduos foram classificados em primeiro lugar. Os resultados indicam que 84% das imagens de teste demonstraram positivo para a pessoa A e 26% de falso negativo. Já para a pessoa B, os resultados foram 100% positivo. Para a

peessoa C, foram 95% positivos e 5% de falso negativos. Para a pessoa D, 97% positivos e 3% de falso negativos. Para a pessoa E, 100% positivos.

Com uma base mais realista e voltada à aplicação, os resultados foram muito melhores em comparação aos resultados apresentados na RN-1. Assim, o sistema não permite quase nenhuma pessoa indesejada entrar no local ou impedir a entrada das pessoas certas..

TOP 1	A	B	C	D	E	%
A	38	-	7	-	-	84,4
B	-	45	-	-	-	100
C	2	-	43	-	-	95,5
D	1	-	-	44	-	97,7
E	-	-	-	-	45	100

Tabela 5.3: matriz de confusão da base de teste para TOP 1 da RN-2. São 45 imagens de teste da base própria com arquitetura mobilenet_1.0_128. TOP 1 representa que o indivíduo ficou em primeiro lugar na classificação.

A Tabela 5.4 apresenta a matriz de confusão do TOP 2, ou seja, dos dois melhores na classificação do Tensorflow. Os resultados mostram que em todas as imagens de teste o indivíduo estava entre os dois primeiros classificados. Isso para todas as classes, ou seja, das 45 imagens de teste, em todas elas os indivíduos foram classificados entre os dois primeiros.

TOP 2	A	B	C	D	E	%
A	45	-	-	-	-	100
B	-	45	-	-	-	100
C	-	-	45	-	-	100
D	-	-	-	45	-	100
E	-	-	-	-	45	100

Tabela 5.4: matriz de confusão da base de teste para TOP 1 da RN-2. São 45 imagens de teste da base própria com arquitetura mobilenet_1.0_128. TOP 2 representa que o indivíduo ficou entre os dois primeiros na classificação.

O Gráfico 5.6 mostra, de forma mais clara, os resultados das acurácias obtidas nos TOP 1 e TOP 2 da RN-2, utilizando a base própria, de 5 classes, com a arquitetura Mobilenet_1.0_128. Pode-se observar que é elevado o nível de acurácia dos resultados no TOP 1 (tom claro), acima de 80%, e que todas as classes no TOP 2 (tom escuro) obtiveram resultados ainda melhores atingindo excelentes níveis de 100%.

Gráfico 5.6: gráfico de comparação dos TOP 1 e TOP 2 da RN-2. Base própria e com arquitetura Mobilenet_1.0_128.

De modo geral, o uso da base própria voltada para a aplicação, fez com que os resultados fossem bem melhores. O Tensorflow classificou melhor as imagens de teste, demonstrando que a rede foi bem treinada e que o sistema é bem eficiente quando voltado para a aplicação.

O motivo de se ter bons resultados foi a qualidade das imagens aplicadas na RN-2 (base própria voltada à aplicação). As imagens eram do tamanho de 2,5M a

4,0M, de altura 1920 pixels e comprimento também de 1920 pixels, sendo fotos tiradas da metade do corpo para cima no formato jpg e com a face voltada para frente e para o lado. Já as imagens da RN-1 (provenientes da base LFW) eram de tamanho entre 11Kb a 13Kb, de comprimento 250 pixels e altura 250 pixels, no formato jpg, com rostos mais próximos e grandes, se comparados às imagens da Análise B.

5.3 RN-3 - Reconhecimento de vida com complexidade 1:N

O Reconhecimento de Vida foi feito através da análise dos espectrogramas dos áudios das pessoas, aplicando sobre os arquivos de áudio uma sequência de cálculos matemáticos denominada Transformada de Fourier. A base própria envolveu os mesmos familiares da RN-2, porém foram capturados seus áudios da voz, ou seja, foi capturado diversas vezes a seguinte frase: “A senha secreta é tigre.”. Portanto, essa base contém 5 classes. Em cada classe, existem 100 imagens para a base de treino e 50 imagens para a base de teste.

Os resultados são mostrados na Tabela 5.5, através de uma matriz de confusão para TOP 1, ou seja, apresenta quais indivíduos foram classificados em primeiro lugar. Os resultados indicam que para a pessoa A, 92% dos espectrogramas da base de teste tiveram resultados positivos e 8% resultaram em falso negativo. Para a pessoa B, foram 70% de respostas positivas e 30% de falso negativo. Para a pessoa C, foi obtido 88% de positivo e 12% de falso negativo. Para a pessoa D, foram 72% positivos e 28% de falso negativo. E para a pessoa E, foram obtidos 92% de respostas positivas e 8% de falso negativo.

Essa relação é de 1:N, com $N=5$, ou seja, o sistema terá grandes chances de acerto após a primeira verificação da imagem da pessoa. Os resultados apresentados foram considerados bons, pois, serão comparados com o resultado final da RN-2, de Reconhecimento de Face.

TOP 1	A	B	C	D	E	%
A	46	-	3	1	-	92
B	-	35	9	2	4	70
C	1	-	44	2	3	88
D	-	-	4	36	10	72
E	-	-	3	1	46	92

Tabela 5.5: matriz de confusão da base de teste para TOP 1 da RN-3. São 50 imagens de teste da base de espectrograma com arquitetura mobilenet_1.0_128. TOP 1 representa que o indivíduo ficou entre os dois primeiros na classificação.

A Tabela 5.6 apresenta a matriz de confusão do TOP 2, ou seja, os dois melhores na classificação da rede. Os resultados mostram que o grau de acerto é muito maior para todas as classes, ou seja, dos 50 espectrogramas de teste, a grande maioria deu resultado entre os dois primeiros.

TOP 2	A	B	C	D	E	%
A	50	-	-	-	-	100
B	-	40	5	-	5	80
C	-	-	48	1	1	96
D	-	-	2	47	1	94
E	-	-	1	-	49	98

Tabela 5.6: matriz de confusão da base de teste para TOP 2 da RN-3. São 50 imagens de teste da base de espectrograma para arquitetura mobilenet_1.0_128 e com resultados em TOP 2.

A Gráfico 5.7 mostra, de forma clara, os resultados das acurácias obtidas nos TOP 1 e TOP 2 da RN-3, aplicada sobre os espectrogramas da base própria. Pode-se observar que é elevado o nível de acurácia dos resultados no TOP 1 (tom claro), acima de 70%, e que todas as classes no TOP 2 (tom escuro) obtiveram resultados ainda melhores, atingindo níveis acima de 80% e chegando até a 100% para a pessoa A.

Isso mostra que a rede foi bem treinada e que ela posiciona o alvo entre os dois primeiros colocados na grande maioria dos casos.

Gráfico 5.7: gráfico de comparação dos TOP 1 e TOP 2 da RN-3

5.3.1 RN-4 - Reconhecimento de vida com complexidade 1:1

As próximas tabelas mostram os resultados do RN-4, com complexidade 1:1, o que aumenta o grau de confiança da aplicação para permitir somente a entrada das pessoas que estão incluídas na rede e impedir as pessoas indesejadas. Para isso, foram criadas 5 novas redes, uma para cada indivíduo, com o intuito de aumentar a relação de confiança para se atingir a complexidade de 1:1, aumentando assim o grau de autenticação.

A RN-4 treinada, possui 2 classes: uma contendo espectrogramas do respectivo indivíduo e outra contendo os espectrogramas de todos os outros indivíduos. Com isso se garante a complexidade 1:1 do projeto. Lembrando que a RN-4 só será acionada na aplicação se as redes RN-2 e RN-3 obtiverem resultados diferentes para o indivíduo alvo.

A Tabela 5.7 mostra os resultados obtidos para o indivíduo A (RN-4.1, de acordo com o esquema do projeto no capítulo 4.2 da Metodologia), que não foram muito favoráveis, sendo 71% de positivos e 29% de falso negativos. Porém, se for analisar que essa etapa é a última verificação, após uma sequência de 2 verificações(RN-2 e RN-3), pode ser considerado como um bom resultado.

	Indivíduo A	Outros	%
Indivíduo A	32	-	71,1
Outros	-	13	28,8

Tabela 5.7: resultados da RN-4.1 para o indivíduo A.

A Tabela 5.8 mostra os resultados para o indivíduo B(RN-4.2 do esquema do projeto), que apresentaram excelentes índices, quase 98% positivo e 2% de falso negativo, ou seja, quase 100% de confiança nos resultados.

	Indivíduo B	Outros	%
Indivíduo B	44	-	97,7
Outros	-	1	2,2

Tabela 5.8: resultados da RN-4.2 para o indivíduo B.

A Tabela 5.9 mostra os resultados para o indivíduo C, que também apresentaram excelentes índices, tendo 93% de positivo e 7% de falso negativo, ou seja, quase 100% de confiança nos resultados.

	Indivíduo C	Outros	%
Indivíduo C	42	-	93,3
Outros	-	3	6,6

Tabela 5.9: resultados da RN-4.3 para o indivíduo C

A Tabela 5.10 mostra os resultados para o indivíduo D. Neste caso, também obteve excelentes resultados, 95% positivo e 5% de falso negativos (com quase 100% de confiança nos resultados).

	Indivíduo D	Outros	%
Indivíduo D	43	-	95,5
Outros	-	2	4,4

Tabela 5.10: resultados da RN-4.4 para o indivíduo D

A Tabela 5.11 mostra os resultados para o indivíduo E, com excelente índice de 100% de positivos, demonstrando alta confiança nos resultados.

	Indivíduo E	Outros	%
Indivíduo E	45	-	100
Outros	-	0	0

Tabela 5.11: resultados da RN-4.5 para o indivíduo E

A Gráfico 5.8 mostra, de forma mais clara, o resultado das cinco redes neurais, aplicadas separadamente para cada indivíduo. O Gráfico 5.8 mostra que todas as classes foram muito bem classificadas, com altos índices de acurácia, acima de 93% para quase todos os indivíduos, com exceção da pessoa A, que foi classificada corretamente em 71% dos casos.

Estes resultados reforçam a importância da complexidade de 1:1, o que garante o alto grau de autenticação da aplicação.

RN4: Reconhecimento de vida com complexidade 1:1 Comparação indivíduo X outros

Gráfico 5.8: resultado das cinco redes neurais de reconhecimento de vida com complexidade 1:1

5.4 Análise dos Limiares das Redes Neurais

Os resultados dos limiares mostrados nas tabelas abaixo indicam que as colunas representam as Redes Neurais e as linhas uma faixa de valores dos resultados das acurácias das imagens da base de teste. As células representam a quantidade de valores das acurácias, de cada indivíduo da base de teste avaliada pela rede, dentro das faixas de valores. A primeira linha apresenta valores abaixo dos 50% de acurácia, já a segunda linha valores entre 50% a 60%, a terceira valores entre 60% e 70%, a quarta valores entre 70% e 80%, a penúltima entre 80% e 90% e última linha indica valores entre 90% e 100% de acurácia.

5.4.1 Limiares da Pessoa A para cada RN

PESSOA A	RN-2	RN-3	RN-4
< 50	8	4	14
entre 50% a 60%	6	5	2
entre 60% a 70%	8	5	5
entre 70% a 80%	2	14	4
entre 80% a 90%	3	18	4
entre 90% a 100%	18	4	16
LIMIAR	71,22222222	81,77777778	67

Tabela 5.12: limiar da Pessoa A para cada RN

A Tabela 5.12 nos mostra que a Pessoa A obteve um limiar de 71,2% para a RN-2, 81,7% para a RN-3 e 67% para a RN-4. Com esses resultados podemos concluir que deve-se melhorar o grau de confiança da Pessoa A para a RN-4 e RN-2, com inserção de melhores imagens da face e áudios na rede, respectivamente. Para a RN-3, o grau de confiança para a Pessoa A é relativamente alto, não necessitando re-treinar a rede.

5.4.2 Limiares da Pessoa B para cada RN

PESSOA B	RN-2	RN-3	RN-4
< 50	0	25	1
entre 50% a 60%	0	1	0
entre 60% a 70%	1	7	1
entre 70% a 80%	1	6	1
entre 80% a 90%	1	9	0
entre 90% a 100%	42	2	42
LIMIAR	93,66666667	59,22222222	92,44444444

Tabela 5.13: limiar da Pessoa B para cada RN

A Tabela 5.13 nos mostra que a Pessoa B obteve um limiar de 93,6% para a RN-2, 59,2% para a RN-3 e 92,4% para a RN-4. Com esses resultados podemos

concluir que deve-se melhorar o grau de confiança da Pessoa B para a RN-3, com inserção de melhores áudios na rede, aumentando assim o grau de confiança na aplicação. Para as RN-2 e RN-4, o grau de confiança para a Pessoa B é bem alto, não necessitando re-treinar a rede.

5.4.3 Limiares da Pessoa C para cada RN

PESSOA C	RN-2	RN-3	RN-4
< 50	4	22	3
entre 50% a 60%	4	5	5
entre 60% a 70%	6	7	3
entre 70% a 80%	2	9	1
entre 80% a 90%	6	5	5
entre 90% a 100%	24	1	28
LIMIAR	81,55555556	57,44444444	82,66666667

Tabela 5.14: limiar da Pessoa C para cada RN

A Tabela 5.14 nos mostra que a Pessoa C obteve um limiar de 81,5% para a RN-2, 57,4% para a RN-3 e 82,6% para a RN-4. Com esses resultados podemos concluir que deve-se melhorar o grau de confiança da Pessoa C para a RN-3, com inserção de melhores áudios na rede, aumentando assim o grau de confiança na aplicação. Para as RN-2 e RN-4, o limiar dos resultados para a Pessoa B é bem elevada, não necessitando re-treinar a rede.

5.4.4 Limiares da Pessoa D para cada RN

PESSOA D	RN-2	RN-3	RN-4
< 50	2	16	2
entre 50% a 60%	2	6	2
entre 60% a 70%	1	5	2
entre 70% a 80%	0	5	0
entre 80% a 90%	7	11	1
entre 90% a 100%	33	7	37
LIMIAR	88,11111111	69,11111111	86,66666667

Tabela 5.15: limiar da Pessoa D para cada RN

A Tabela 5.15 nos mostra que a Pessoa D obteve um limiar de 88,1% para a RN-2, 69,1% para a RN-3 e 86,6% para a RN-4. Com esses resultados podemos concluir que deve-se melhorar o grau de confiança da Pessoa D para a RN-3, com inserção de melhores áudios na rede, aumentando assim o grau de confiança na aplicação. Para as RN-2 e RN-4, o limiar dos resultados para a Pessoa D é bem elevada, não necessitando re-treinar a rede.

5.4.5 Limiares da Pessoa E para cada RN

PESSOA E	RN-2	RN-3	RN-4
< 50	0	8	0
entre 50% a 60%	0	4	0
entre 60% a 70%	0	4	0
entre 70% a 80%	0	11	1
entre 80% a 90%	2	9	1
entre 90% a 100%	43	14	42
LIMIAR	94,55555556	80,88888889	92,22222222

Tabela 5.15: limiar da Pessoa E para cada RN

A Tabela 5.15 nos mostra que a Pessoa E obteve um limiar de 94,5% para a RN-2, 80,8% para a RN-3 e 92,2% para a RN-4. Com esses resultados podemos concluir que a Pessoa E possui elevados índices de aceitação da aplicação, demonstrando alto grau de confiança, por isso não precisa re-treinar a rede.

CAPÍTULO 6

CONCLUSÕES E TRABALHOS FUTUROS

O trabalho em questão apresentou bons resultados tanto para o reconhecimento facial quanto para o reconhecimento de vida. É importante ressaltar que a sequência de etapas descritas na metodologia (começando pela escolha da melhor arquitetura, passando por reconhecimento facial de complexidade 1:N e terminando com o reconhecimento de vida com complexidade 1:1) foi de extrema importância para aumentar a acurácia dos resultados, demonstrando que o sistema tem alto grau de confiabilidade.

Em termos de trabalhos futuros, um aprimoramento importante, para que a aplicação se torne um produto, seria na área de reconhecimento de vida, no qual pode-se vir a ser utilizado vídeos e câmeras térmicas ou sistemas com percepção 3D (câmeras estéreo, por exemplo). Entretanto, esses avanços demandam maior custo ao projeto, processamento de imagem e esforço de implementação. O problema do processamento de imagens em um sistema dedicado (sistema embarcado) e de baixo custo pode ser resolvido adaptando-se um aparelho do tipo smart TV Box Android. Estes aparelhos são de baixo custo, possuem uma boa capacidade de processamento de imagem e são produzidos em larga escala para transformar TVs normais em Smart TVs. Neste aparelho, pode-se instalar o Tensorflow (versão mobile) para poder utilizar a rede neural com muito mais rapidez, porém o treinamento seria feito em máquinas de

maior poder de processamento, possivelmente usando computação na nuvem, e apenas transferindo a rede neural já treinada para uso junto ao Smart Box

Com a adaptação do sistema em uma configuração como está descrita acima (Smart + Box + Computação na Nuvem) será mais rápido treinar uma base com mais pessoas, para o reconhecimento facial, podendo o produto se estender para estabelecimentos com maior fluxo de pessoas como pequenas empresas e eventos.

Outro aprimoramento seria a adaptação de um hardware para a abertura e fechamento da porta, além de uma interface web para poder cadastrar e gerenciar o controle de acesso de pessoas na residência ou no laboratório. Além disso, criar uma interface web com vários níveis de permissão e diversos tipos de funcionalidades para dar maior robustez ao sistema.

CAPÍTULO 7

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAMSON, N.; BRAVERMAN, D.; SEBESTYEN, G. **Pattern recognition and machine learning**. Information Theory, IEEE Transactions on, vol.9, ed. 4, p.257-261, 1963.

ALLIED MARKET RESEARCH. Disponível em:
<<https://www.alliedmarketresearch.com/facial-recognition-market>> Acesso em:
10/10/2017.

ALVARENGA, R.J. **Reconhecimento de Comandos de Voz por Redes Neurais**. Dissertação de Mestrado, 2012

AMARAL, M.A.; BARRIVIERA, R.; COTRIN TEIXEIRA E. **Reconhecimento de Voz para Automação Residencial baseado em Agentes Inteligentes**, 2004.

AMARO, R.; DIEGUEZ, A. Imagem do site ice interactive. 2010.

ARBIB, M.A. **The handbook of brain theory and neural networks**. p. 1290, 2002.

BAKEN, R.J.; ORLIKOFF, R.F. **Clinical measurement of speech and voice**, 2a ed. San Diego: Singular Thomson Learning; p.225-284, 2000

BAO W.; LI, H.L.N.J.W. **A liveness detection method for face recognition based on optical flow field**. v. 11, p. 233–236, 2009.

BENGIO, Y. **Deep Learning architectures for AI. Foundations and trends**, 2009.

BARAVIEIRA, P. B. **Aplicação de uma rede neural artificial para a avaliação da rugosidade e soprovidade vocal**. Tese (Doutorado)–Programa de Pós-Graduação Interunidades em Bioengenharia EESC/FMRP/IQSC, Universidade de São Paulo, 2016.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. **Representation learning: A review and new perspectives**. Pattern Analysis and Machine Intelligence, IEEE Transactions on, IEEE, v. 35, n. 8, p. 1798–1828, 2013.

BIANCHINI, A.R. **Arquitetura de redes neurais para o reconhecimento facial baseado no neocognitron**. p. 142, 2001.

BIOMETRIA: RECONHECIMENTO FACIAL. **Sistemas de Informação Industriais e Consultoria**, 2017. Disponível em
<<http://www.sinfic.pt/SinficWeb/displayconteudo.do2?numero=44661>> Acesso em: 09/10/2017.

BIOMETRICS. Disponível em <<http://biometrics.mainguet.org/alive/references.htm>> Acesso em: 17/10/2018.

BOULKENAFET Z.; KOMULAINEN, J. H. A. **Face spoofing detection using colour texture analysis**. v. 11, p. 1818–1830, 2008.

CAFFE. Disponível em <<http://caffe.berkeleyvision.org>>. Acesso em 19/09/2017.

CAMADAS DE REDES NEURAIS. Disponível em
<http://www.inf.ufpr.br/aghochuli/caffe/CNN_PPT.pdf> Acesso em 29/09/2018.

CROWLEY, J.L.; CHRISTENSEN, H.I. **Vision as Process**. Springer-Verlag, 1995.

Chen, Bor-Chun, Chu-Song Chen, and Winston H. Hsu (2014). “**Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval**”. In: Proceedings of the European Conference on Computer Vision (ECCV).

DAMASCENO SILVA, GUILHERME. **Desenvolvimento de um Sistema de Visão Computacional para o Estudo do Comportamento de Animais Experimentais**.

Disponível em:

<<http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/BUOS-8CUFBY>>. Acesso em 15/06/2016.

DENG, L.; YU, D. **Deep learning: methods and applications**. **Foundations and Trends in Signal Processing**, Now Publishers Inc., v. 7, n. 3–4, p. 197–387, 2014.

FAUGERAS, O. **Three-Dimensional Computer Vision, A Geometric Viewpoint**. MIT Press, 1993.

FAUSETT, L. **Fundamentals of neural networks architectures, algorithms, and applications**. 1995.

FERNANDES, A.M.R. **Inteligência Artificial: noções gerais**. Florianópolis: **Visual Books**, 2005.

FUKUSHIMA, K. Neocognitron: **A model for visual pattern recognition**. p. 613–616, 1995.

GALBALLY, J.; ORTIZ-LOPES, J.; FIERREZ, J.; ORTEGA-GARCIA, J. **Iris liveness detection based on quality related features**. **International Conference on Biometrics(ICB)**, 2012.

GIACOMO, A.A; CELSO, C.C; MAURÍCIO J.A.B. **Análise de Padrões de Voz através da Transformada em Ondeletas**. XIII Encontro Latino Americano de Iniciação Científica e IX Encontro Latino Americano de Pós-Graduação, 2009.

GILLIER, R. **O disfarce da voz em fonética forense**. Sistema Integrado de Bibliotecas, 2011.

GLOBAL IDENTITY MANAGEMENT. Disponível em
<<https://findbiometrics.com/year-review-liveness-detection-401180/>> Acesso em:
17/10/2018.

GOUVEIA WELLINGTON DA ROCHA; PAIVA, M. S. V. d. **Detecção de faces humanas em imagens coloridas utilizando redes neurais artificiais**, 2009.

GUIANI, L.; MARCIALIS, G.L.; ROLI, F. **Fingerprint Liveness Detection by Local Phase Quantization**. 21st International Conference on Pattern Recognition (ICPR), 2012.

HAYKIN, S. **Redes neurais: Princípio e prática**. p. 896, 1999.

HARTLEY, R.; ZISSERMAN, A. **Multiple View Geometry in computer vision**. Cambridge University Press, 2003.

HEISELE, B.; HO, P.; POGGIO, P. **Face recognition with support vector machines: Global versus component - based approach**, in Eighth IEEE International Conference on Computer Vision (ICCV), vol. 2. IEEE, p. 688 – 694, 2001.

HINTON, G.; SRIVASTAVA, N.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R.R. **Improving neural networks by preventing co-adaptation of feature detectors**. arXiv preprint arXiv:1207.0580, 2012b.

HOLLIEN, H.; Majewski, W.; Hollien, P. **Speaker identification by long-term spectra under normal, stress and disguised conditions**. Journal. Acoust. Soc. Am. (Suppl. 55),820, 1974.

HUA, W.; CUIQIN, M.; LIJUAN, Z. **A Brief Review of Machine Learning and Its Application**. Information Engineering and Computer Science. ICIECS. International Conference on, vol., no., p. 1-4, 2009.

HUANG, G.B.; RAMESH, M.; BERG, T.; LEARNED-MILLER, E. **Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments**. Tech. rep. 07-49. University of Massachusetts, Amherst, 2007.

INTEL AI ACADEMY. Disponível em
<<https://software.intel.com/en-us/articles/improving-svm-performance-with-intel-daal>>
Acesso em 10/08/2018.

JONES T. D.; LAWSON S. W.; BENYON D.; ARMITAGE A. **Comparison of Human and Machine Recognition of Everyday Human Actions**, V.G. Duffy (Ed.): **Digital Human Modeling**, LNCS 4561, 2007, 120–129p, HCII 2007.

KAZEMI V.; SULLIVAN, J. **One millisecond face alignment with an ensemble of regression trees**. p. 1867–1874, 2014.

KERLOW, I.V. **The Art of 3D Computer Animation and Effects**. 3.ed. Hoboken: John Wiley & Sons, Inc., 2004.

KIRBY, M; SIROVICH, L. **Low-dimensional procedure for characterization of human faces**. Journal of the Optical Society of America, vol 4, p.519, 1987

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. **Imagenet classification with deep convolutional neural networks**. In: **Advances in neural information processing systems**. [S.l.: s.n.], p. 1097–1105, 2012.

KRUPPA, H.; CASTRILLON-SANTANA, M.; SCHIELE, B. **Fast and robust face finding via local context**. In **Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)**, pages 157–164, 2003.

LECUN, Y.; KAVUKCUOGLU, K.; FARABET, C. **Convolutional networks and applications in vision**. In: IEEE. Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. [S.l.], p. 253–256, 2010.

LFW. Disponível em <<http://vis-www.cs.umass.edu/lfw/>>. Acesso em 17/04/2017.

LUCENA JÚNIOR, Alberto. **Arte da Animação: Técnicas e Estética Através da História**. 2.ed. São Paulo: Editora Senac São Paulo, 2005.

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep learning**. *Nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LORENZI, F.; SILVEIRA, S. R. **Desenvolvimento de Sistemas de Informação Inteligentes**. Porto Alegre: UniRitter, 2011.

LUCENA JÚNIOR, Alberto. **Arte da Animação: Técnicas e Estética Através da História**. 2.ed. São Paulo: Editora Senac São Paulo, 2005.

LUGER, G. F.; STUBBLEFIELD, W. A. **Artificial Intelligence: Structures and Strategies for Complex Problem Solving**, 6th Edition, 2009.

LIBSVM. Disponível em <<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>>. Acessado em 07/09/2018

PONTES PAL, Vieira VP, Gonçalves MIR, Pontes AAL. **Características das vozes roucas, ásperas e normais: análise acústica espectrográfica comparativa**. *Rev Bras Otorrinolaringol*. 2002;68(2):182-8.

MATHERON, G. **Random sets and integrated geometry**. Wiley, 1975.

MEHROTRA KISHAN; MOHAN, C. K. R. S. **Elements of artificial neural networks massachusetts**. p. 344, 2000.

MINATO, D.V.; CORAL, S.; MATOS GARCIA, M.C. **Análise de Características em Redes Neurais Artificiais Aplicadas em Visão Computacional com Ênfase em**

Reconhecimento Facial. Congresso Sul Brasileiro de Computação (SULCOMP), 2017.

MORRIS, T. **Computer Vision and Image Processing.** Palgrave Macmillan, 2004.

NVIDIA. Disponível em <<https://developer.nvidia.com/cudnn>> Acesso em: 18/11/2017

Ng, Hongwei and Stefan Winkler (2014). “**A data-driven approach to cleaning large face datasets**”. In: ICIP.

OPENFACE. Disponível em <<https://openface-api.readthedocs.io/en/latest/>> Acesso em 18/09/2018

PETRY, A; ZANUZ, A; BARONE, D.A.C. **Utilização de Técnicas de Processamento Digital de Sinais para a Identificação Automática de Pessoas pela Voz.** Simpósio sobre Segurança em Informática, São José dos Campos, SP, 1999.

RECONHECIMENTO FACIAL E TECNOLOGIA PARA VIGILÂNCIA NACIONAL.

Disponível em

<<http://socialismocriativo.com.br/reconhecimento-facial-e-tecnologia-para-vigilancia-nacional/>> Acesso em: 18/10/2018.

RECONHECIMENTO FACIAL EM AEROPORTOS. Disponível em

<<http://g1.globo.com/jornal-nacional/noticia/2016/07/reconhecimento-facial-passa-ser-usado-em-14-aeroportos.html>>. Acesso em: 17/10/2018

SAMPAIO DE PAULA, G. R. **Reconhecimento de fala por algoritmo de rede neural embarcado em DSP.** Biblioteca Pontifícia Universidade Católica do Rio de Janeiro, 2011..

SCHMIDHUBER, J. **Deep learning in neural networks: An overview. Neural Networks, Elsevier**, v. 61, p. 85–117, 2015.

SIROVICH, L.; KIRBY, M. **Low dimensional procedure for the characterization of human faces**. Journal of the Optical Society of America 4, no. 3, pp. 519–524, 1987.

SILVA, I. N.; SPATTI, D. H.; Flauzino, R. A. **Redes Neurais artificiais: para engenharias e ciências aplicadas**. 1 ed. São Paulo: Artliber, 2010.

SIROVICH, I; KIRBY, M. **Low-dimensional procedure for the characterization of human faces**. Journal of Optical Society of American A, vol 4, pag 519, 1987.

SCHUTZER, D. **Artificial intelligence: an applications-oriented approach**. New York: Van Nostrand Reinhold Company, 1987.

SPRINGER Disponível em

<https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-73003-5_76>

Acesso em: 18/10/2018.

SVM-LIGHT. Disponível em <<http://svmlight.joachims.org/>>. Acesso em 13/08/2018

TAN X.; LI, Y.L.J.J.L. **Face liveness detection from a single image with sparse low rank bilinear discriminative model**. p. 504–517, 2010.

TENSORFLOW. Disponível em <<https://www.tensorflow.org/>>. Acesso em 25/09/2017.

TOLBA A.; EL-BAZ, A. E.H.A. **Face recognition: A literature review. International Journal of Signal Processing**. v. 2, p. 88–103, 2006.

TORRES, E.S.; SILVA, L.J.; SANTOS R.M.; MACEDO H.T.; MATOS L.N. **Redes Neurais Convolucionais no Reconhecimento de Fala em Português para jogos em plataformas móveis**. SBC – Proceedings of SBGames 2016 | ISSN: 2179-2259, 2016.

THEANO. Disponível em <<http://deeplearning.net/software/theano/>> Acesso em 30/11/2018.

VAPNIK, V. N. **An Overview of Statistical Learning Theory**, IEEE Transactions on Neural Networks, vol 10 n° 5, 1999.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. Springer, 1995.

VELHO, L. **Computação Visual e Novas Mídias. Perspectivas e Grandes Desafios da Computação no Brasil**. IMPA, 2016.

WAIKATO UNIVERSITY. Disponível em <<https://www.cs.waikato.ac.nz/ml/weka/>>. Acessado em 07/09/2018.

WEN D.; HAN, H.J.A.K. **Face spoof detection with image distortion analysis**. v. 10, n. 4, p. 746–761, 2015.

ZEILER, M. D.; FERGUS, R. **Stochastic pooling for regularization of deep convolutional neural networks**. arXiv preprint arXiv:1301.3557, 2013.