UNIVERSIDADE DE TAUBATÉ

Denis Magalhães de Almeida Eiras

ARQUITETURA DE SISTEMA DE RECONHECIMENTO FACIAL DE BAIXO CUSTO PARA CONTROLE DE ACESSO EM AMBIENTES NÃO CONTROLADOS

Taubaté – SP
2019

Denis Magalhães de Almeida Eiras

ARQUITETURA DE SISTEMA DE RECONHECIMENTO FACIAL DE BAIXO CUSTO PARA CONTROLE DE ACESSO EM AMBIENTES NÃO CONTROLADOS

Dissertação apresentada para obtenção do Título de Mestre pelo Curso Engenharia Mecânica do Departamento de Engenharia Mecânica da Universidade de Taubaté
Área de Concentração: Automação
Orientador: Dr. Luís Fernando de Almeida

Taubaté – SP
2019

SIBi – Sistema Integrado de Bibliotecas / UNITAU

E35a Eiras, Denis Magalhães de Almeida
Arquitetura de sistema de reconhecimento facial de baixo custo para controle de acesso em ambientes não controlados / Denis Magalhães de Almeida Eiras. -- 2019.
124 f. : il.

Dissertação (Mestrado) – Universidade de Taubaté, Departamento de Engenharia Mecânica e Elétrica, 2019.
Orientação: Prof. Dr. Luís Fernando de Almeida, Departamento de Informática.

1. Controle de acesso. 2. Reconhecimento facial. 3. Redes neurais convolucionais profundas. I. Título. II. Mestrado em Engenharia Mecânica.

CDD – 629.8

Ficha catalográfica elaborada por Shirlei Righeti – CRB-8/6995

DENIS MAGALHÃES DE ALMEIDA EIRAS

ARQUITETURA DE SISTEMA DE RECONHECIMENTO FACIAL DE BAIXO CUSTO PARA CONTROLE DE ACESSO EM AMBIENTES NÃO CONTROLADOS

Dissertação apresentada para obtenção do Título de Mestre pelo Curso Engenharia Mecânica do Departamento de Engenharia Mecânica da Universidade de Taubaté,
Área de Concentração: Automação
Orientador: Dr. Luís Fernando de Almeida

Data:_____

Resultado_____

BANCA EXAMINADORA

Prof. Dr. Luis Fernando de Almeida UNIVERSIDADE DE TAUBATÉ

Assinatura_____

Prof. Dr. Álvaro Manoel de Souza Soares UNIVERSIDADE DE TAUBATÉ

Assinatura_____

Prof. Dr. Eugênio Esper de Almeida INPE

Assinatura_____

AGRADECIMENTOS

À minha esposa pela compreensão e incentivo.
Ao Prof. Dr. Luís Fernando de Almeida pela orientação do trabalho e ensinamentos.
Aos Professores da banca Dr. Álvaro Manoel de Souza Soares e Dr. Eugênio Esper de Almeida pelas colaborações.
Aos professores do curso Dr. Giorgio Eugênio Oscare Giacaglia, Dr. Francisco Carlos Parquet Bizarria, Dr. Álvaro Manoel De Souza Soares, Dr. Luís Fernando De Almeida, Dr. Eduardo Hidenori Enari, Dr. José Walter Parquet Bizarria pelos ensinamentos.
Ao funcionário André Pasquali, Prof. Ms. Daniel Merli Lamosa e aos alunos do Departamento de Informática pela colaboração na captura de imagens para os experimentos.
Ao fotógrafo Rafael Coelho pela colaboração nos conceitos fotográficos e de iluminação.

RESUMO

Nos últimos anos, sistemas de Reconhecimento Facial têm sido cada vez mais utilizados como forma de biometria confiável, devido à capacidade de identificar indivíduos sem a necessidade de colaboração espontânea, sendo utilizado na prevenção de ataques terroristas, controle de acesso, sistemas de segurança, de vigilância e outros. O presente trabalho propõe um protótipo de um sistema de controle de acesso, de uma pessoa por vez, em ambientes internos com pouco controle de iluminação e sem controle de pose das pessoas, construído a partir de um modelo criado para esse propósito, que define uma de arquitetura de *software*, projetada utilizando plataformas de código livre, uma arquitetura de *hardware* de baixo custo e definições espaciais e de iluminação de um ambiente. As arquiteturas foram projetadas a partir de uma pesquisa exploratória que identifica, de forma predominantemente qualitativa, técnicas rápidas e eficazes de Identificação de Humanos e de Reconhecimento Facial. A partir de 2012, diversos trabalhos envolvendo Reconhecimento Facial têm utilizado Redes Neurais com sucesso, superando outras técnicas e superando constantemente o estado da arte. A arquitetura de *software* propõe uma combinação de técnicas rápidas de Identificação de Humanos, Detecção Facial e de Reconhecimento Facial, o qual utiliza uma Rede Neural Convolucional Profunda e uma técnica inovadora de controle de luminosidade em imagens digitais, ao passo que, a arquitetura de *hardware* proposta, utiliza componentes de baixo custo. Utilizando essas arquiteturas, o atual trabalho cria um protótipo capaz de realizar o Reconhecimento Facial atingindo taxas de reconhecimento acima de 90%.

Palavras-chave: Reconhecimento facial. Redes Neurais Convolucionais Profundas. Controle de acesso.

ABSTRACT

In recent years, Facial Recognition systems have been increasingly used as a reliable form of biometrics, due to the ability to identify individuals without the need for spontaneous collaboration, and have been used in the prevention of terrorist attacks, access control, security systems, surveillance and others. The present work proposes an access control system prototype, one person at a time, which runs indoors with little control of lighting and without control of the person's pose, built from a model created for this purpose, that defines a software architecture, designed using open source code platforms, a low cost hardware architecture and spatial and lighting settings for an environment. The architectures were designed on an exploratory research that identifies, in a predominantly qualitative way, fast and effective techniques of Human Identification and Facial Recognition. From 2012, several works involving Facial Recognition have successfully used Neural Networks, surpassing other techniques and constantly surpassing the state of the art. The software architecture proposes a combination of fast Human Identification, Facial Detection and Facial Recognition techniques, which uses a Deep Convolution Neural Network and an innovative technique for controlling brightness in digital images, as the hardware architecture uses low cost components. By using these architectures, the current work creates a prototype which performs Facial Recognition reaching recognition rates above 90%.

Keywords: Facial recognition. Deep Neural Convolutional Networks. Access control.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CFT	Circuito Fechado de TV, *Closed-Circuit Television*.
DF	Detecção Facial, *Face Detection*.
DH	Detecção de Humanos, *Human Detection*.
DW	*Depth Wise,* Convolução em profundidade.
FA	Função de Ativação.
FP	Falsos Positivos
HGO	Histograma de Gradientes Orientados.
IA	Inteligência Artificial.
ILSVRC	*ImageNet Large Scale Visual Recognition Challenge*.
LBP	*Local Binary Pattern,* Padrão Binário Local.
LDA	*Linear Discriminant Analisys,* Análise Discriminante Linear.
MAC	*Multiply-Accumulates*. Número de operações de multiplicação e adição fundidas.
PCA	*Principal Component Analisys,* Análise de Componentes Principais.
PW	*Point Wise,* Convolução Pontual.
RF	Reconhecimento de Faces, *Face Recognition*.
RI	Região de Interesse, *Region of Interest*.
RNA	Redes Neurais *Artificiais, Artificial Neural Networks*.
RNC	Redes Neurais Convolucionais, *Convolutional Neural Networks.*
RNCP	Redes Neurais Convolucionais Profundas, *Deep Convolutional Neural Networks*.
RNP	Redes Neurais Profundas, *Deep Neural Networks*.
RP	Reconhecimento de Padrões, *Pattern Recognition*.
SVM	*Support Vector Machines*.
TC	Totalmente Conectada, *Fully Connected*.
TRF	Taxa de Reconhecimento de Faces, *Recognition Rate*.

SUMÁRIO

1 INTRODUÇÃO

Na última década, o avanço do computacional e de técnicas cada vez mais precisas de Reconhecimento Facial (RF), que envolvem, principalmente, as áreas de Visão Computacional, Reconhecimento de Padrões (RP) e Inteligência Artificial (IA), têm gerado uma grande aumento na quantidade de pesquisas e desenvolvimento de sistemas de RF. Esses sistemas estão sendo cada vez mais utilizados para a prevenção de ataques terroristas, procura de criminosos ou sujeitos ilegais nas fronteiras dos países e aeroportos, controle de acesso, sistemas de segurança e de vigilância. O RF vem sendo utilizado como uma forma de biometria confiável não intrusiva, que não requer necessariamente a cooperação dos indivíduos.

Sistemas de RF requerem a captura de imagens através de câmeras digitais e o armazenamento das imagens contendo faces em um banco de dados. Cada imagem a ser consultada contém uma face, que é utilizada para a Identificação de Face de uma pessoa, dentre diversas pessoas possíveis cadastradas no banco de dados, ou é utilizada para a Verificação de Face, onde imagens de uma mesma pessoa são utilizadas para verificar que, uma imagem a ser comparada, pertence a essa pessoa.

Sistemas de controle de acesso, de segurança e vigilância, utilizam técnicas de Identificação de Faces para identificar um sujeito que circula pelo ambiente projetado. Sistemas que precisam detectar uma determinada pessoa, como os sistemas de segurança utilizados em aeroportos, que procuram identificar criminosos, utilizam técnicas de Verificação de Faces.

Um sistema de RF utiliza a biometria facial para identificar unicamente um indivíduo. A biometria é um termo aplicado às diversas formas às quais uma pessoa pode ser identificada através de aspectos do seu corpo. As biometrias mais comuns, encontradas em sistemas de identificação, são a impressão digital e a íris, mas outras características humanas foram estudadas nos últimos anos, como a geometria de dedo ou de palma, voz, assinatura e face (ABATE et al., 2007).

A utilização da biometria adequada requer estudo para que seja apropriada ao tipo de sistema que se deseja implementar. O reconhecimento da íris é extremamente preciso, mas é caro de implementar e não é muito aceito pelas pessoas. As impressões digitais são confiáveis e não intrusivas, mas não são adequadas para indivíduos não colaborativos. Por outro lado, o RF tem uma maior

confiabilidade e aceitação social, não requerendo a colaboração pontual dos indivíduos, uma vez que as faces podem ser capturadas automaticamente através da utilização de câmeras de segurança digitais.

Em linhas gerais, um sistema de RF pode ser dividido em três etapas: Detecção de Faces (DF), extração de características e RF. Essas etapas geram problemas diferentes, que têm sido estudados separadamente.

A primeira etapa, DF, consiste em enquadrar uma face dentro de uma imagem, que pode ou não conter uma face. A segunda etapa, extração de características, tem como objetivo determinar quais características podem ser utilizadas para identificar unicamente uma face. A terceira etapa, a de RF, utiliza o conjunto de características extraídas para identificar unicamente uma face, como sendo de uma pessoa, dentro de um banco de dados de faces previamente cadastrado, utilizando algoritmos classificadores.

Embora o RF venha evoluindo constantemente, a tarefa não é trivial e muitos problemas precisam ser resolvidos, até que se alcance um sistema de reconhecimento visual próximo ao do humano. Assim como no reconhecimento humano, as dificuldades aparecem nas diferenças entre imagens de uma mesma pessoa, devido às variações de luminosidade, ângulo de visão, pose da cabeça, oclusão e expressões faciais. O cenário fica ainda mais complexo quando essas variações acontecem entre faces de diversas pessoas.

A maioria dos sistemas de RF funciona bem quando são utilizadas imagens estáticas, capturadas em condições controladas de iluminação e ambientação. No entanto, a utilização de imagens capturadas em uma câmera de um sistema de segurança, torna o problema muito mais complexo, devido à qualidade inferior das imagens, causada pela movimentação e variação da iluminação.

A qualidade da imagem é comumente influenciada pela escolha inadequada da câmera e suas configurações, que devem ser adequadas ao ambiente proposto, evitando efeitos indesejados, como o *motion blur*, que é o desfoque devido a um maior tempo de exposição da luz durante a captura de um objeto em movimento, ou as variações de imagem na região da face e do fundo da face, causadas pela quantidade de luz e posicionamento das fontes de luz, como lâmpadas ou luz natural do Sol. Outros problemas comuns são as variações de pose da face, expressões faciais ou a utilização de acessórios, como óculos e chapéus, que causam oclusão facial.

Algoritmos de DF, tais como o algoritmo de Viola e Jones (2001), que utiliza características Haar para representar a face, e técnicas de representação de características, tais como *PCA* (*Principal Component Analysis*), vêm sido utilizados há mais de uma década e estão disponíveis em bibliotecas de *software* livre como o OpenCV (*Open Source Computer Vision Library*) (OPENCV, 2018).

Na etapa de reconhecimento das faces, algoritmos classificadores como K-Nearest Neighbors, Random Forest e K-Star são frequentemente utilizados. No entanto, recentemente, as Redes Neurais Artificiais (RNA) do tipo Redes Neurais Convolucionais Profundas (RNCP) têm se destacado dentre as RNA, por serem flexíveis com relação às variações nas imagens, incluindo variações de intensidade e de localização, sendo assim, melhores adaptadas a sistemas de segurança em ambientes não controlados. As RNCP têm superado outros métodos de classificação de objetos em geral e podem ser encontradas em bibliotecas de código fonte livres, como o TensorFlow (ABADI et al., 2016).

Em síntese, a grande quantidade de *software* livre disponível, câmeras de segurança de boa resolução, computadores com uma ótima capacidade de processamento e placas gráficas com alto poder de processamento, disponíveis a um custo cada vez mais acessível, possibilitam a criação de que projetos de sistemas de segurança de baixo custo, quando arquiteturas ideais de *hardware* e *software* são definidas para esse propósito.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Este trabalho propõe criar um protótipo de sistema de controle de acesso, a partir da modelagem de uma arquitetura de *software*, aqui definida, composta de uma seleção de técnicas de RF rápidas e de uma técnica inovadora de controle de iluminação, que deve ser projetada sob de uma arquitetura de *hardware*, definida neste trabalho, composta de câmeras de vídeo, infraestrutura de rede local e computadores de baixo custo. O protótipo desenvolvido deve ser capaz de reconhecer a identidade das pessoas, que transitam em um ambiente com possível variação de luminosidade, sendo uma pessoa por vez, sem requerer a colaboração voluntária dos indivíduos.

1.1.2 Objetivos específicos

De uma forma específica, este trabalho propõe:

a) identificar, na literatura atual, técnicas rápidas de Detecção de Humanos (DH), DF e de RF que, quando utilizados em conjunto, compõem uma arquitetura de *software* rápida e adequada para plataformas de *hardware* de baixo custo;
b) propor um modelo de arquitetura de *hardware* de baixo custo, compatível com a arquitetura de *software* definida no item a);
c) definir parâmetros de ambientação do espaço utilizado pelo sistema proposto, tais como níveis de luminosidade, posicionamento das câmeras e região espacial adequada para a captura de faces;
d) criar um protótipo de sistema de controle de acesso, utilizando a arquitetura de *software* do item a) e a arquitetura de *hardware* proposta no item b), projetado para ser rápido e extrair o máximo de performance da plataforma de *hardware*, sem sobrecarregá-la;
e) implementar uma funcionalidade no protótipo que melhore a qualidade do RF, através de algum tipo de controle de iluminação;
f) validar a funcionalidade do protótipo em um experimento, que deve registrar faces de indivíduos distintos, em um mesmo ambiente, utilizando alguma variação de iluminação, de forma que a taxa de reconhecimento (TR) durante a passagem de um indivíduo registre uma confiabilidade mínima, aqui definida como 80%;
g) gerar resultados de comparativos das configurações que produziram melhores resultados no experimento do item f);
h) gerar resultados utilizando bancos de dados de faces populares;
i) gerar resultados de comparativos entre o custo do protótipo e de outros sistemas de controle de acesso com biometria facial.

1.2 DELIMITAÇÃO DO PROBLEMA

A pesquisa e a implementação do protótipo aborda a DF e RF de apenas um indivíduo por imagem, podendo ocorrer oclusão de objetos, tais como óculos ou chapéus.

O protótipo do sistema de controle de acesso é projetado apenas para ambientes internos, com possível entrada de luz externa. O protótipo do sistema de controle de acesso pode reconhecer no máximo 1.000 pessoas distintas, devido à arquitetura de RNA utilizada.

A pesquisa de DH, DF e de RF abrange as principais técnicas utilizadas na literatura, mas somente utiliza as técnicas que dispõe de código livre para a implementação do protótipo.

Este trabalho se limita a estudar um sistema de controle de acesso que utiliza biometria facial sem abordar as implicações legais de captura de imagens dos sujeitos.

1.3 JUSTIFICATIVA

Até a presente data, não se encontra na literatura trabalhos que tenham criado um modelo de sistema de controle de acesso de baixo custo, definido por uma arquitetura de *hardware* de baixo custo, técnicas de RF rápidas e flexíveis, disponíveis em bibliotecas de *software* gratuitas, e parâmetros de ambientação, implementados em um protótipo capaz de realizar o RF, sem controle de pose e com pouco controle de iluminação.

1.4 METODOLOGIA

A partir de uma pesquisa exploratória, foi utilizada uma abordagem predominantemente qualitativa para identificar as mais rápidas e melhores técnicas relacionadas ao RF, isto é, DH, DF e RF, considerando as bibliotecas de *software* e plataformas disponíveis de código fonte livres, que implementem os métodos eleitos.

Durante a pesquisa, foi considerado o baixo custo para o projeto, ou seja, a partir da definição de uma arquitetura de *hardware,* composta por computador, câmeras de segurança e equipamentos de rede Ethernet, o trabalho busca eleger uma combinação ideal de técnicas, que definem uma arquitetura de *software* projetada para extrair o máximo de performance sem sobrecarregar a plataforma de *hardware*, através da avaliação dos resultados que indicaram as melhores taxas de reconhecimento, robustez e rapidez durante o RF.

A pesquisa foi realizada utilizando artigos em revistas e anais, para levantar as mais recentes e utilizadas técnicas de RF. Livros, definições de padrões, teses e dissertações, foram essencialmente utilizados para o embasamento teórico do funcionamento das técnicas selecionadas, da arquitetura de redes de computadores e dos conceitos de fotografia e de ambientação. Também foram consultados sites da internet, que oferecem as bibliotecas de *software* livre que implementam as técnicas de RF pesquisadas.

Após a realização da pesquisa exploratória, foi gerada uma seleção das mais utilizadas e reconhecidas técnicas que envolvem o RF. A partir das técnicas selecionadas, foram utilizadas as bibliotecas de *software* que implementam as técnicas, de forma pontual em um sistema, criado para avaliar essas técnicas em um banco de dados popular, que contém faces de indivíduos em diferentes condições de iluminação e ambientação. Utilizando as técnicas que obtiveram os melhores resultados nos bancos de dados, considerando variações de iluminação, diferentes poses e oclusão, iniciou-se o desenvolvimento do sistema de controle de acesso.

Durante o desenvolvimento, foi utilizado um computador com poder de processamento e memória superiores ao custo desejado e câmeras de segurança de diferentes qualidades, para se obter as melhores qualidades possíveis de imagens, dentro das condições de processamento imposta pela arquitetura do sistema. De acordo com as mínimas definições de rede Ehternet local para implementar o protótipo, foram comprados os equipamentos da rede, tais como cabos, plugues, adaptadores, fontes e *switch* para a montagem do protótipo.

Concluída a montagem do protótipo, foi escolhido um ambiente compatível com as definições de ambientação, para realizar os testes do sistema de controle de acesso. A partir dos resultados do experimento, parâmetros do sistema foram ajustados, novas técnicas foram implementadas e algumas correções foram feitas. Em seguida, um segundo experimento, com 50 pessoas e uma grande quantidade de imagens, foi realizado para a geração dos resultados.

1.5 ESTRUTURA DO TRABALHO

Este trabalho está organizado em 5 capítulos: o presente capítulo apresenta a contextualização do problema e os objetivos e justificativas da pesquisa; o Capítulo 2 apresenta o referencial teórico, que mostra as recentes pesquisas sobre

DH, DF, RF, características de *hardware* e características de ambientação, necessários para definição das arquiteturas e a implementação do protótipo de controle de acesso; o Capítulo 3 apresenta uma introdução às RNA e exibe a definição de RNCP utilizada no protótipo; os testes realizados e os respectivos resultados são apresentados no Capítulo 4 e as conclusões no Capítulo 5.

2 RECONHECIMENTO DE FACES

Neste capítulo é apresentada uma visão geral sobre a problemática do RF (seção 2.1) e, em seguida, são apresentadas as etapas necessárias para a identificação das faces. Tal processo é constituído de três etapas: a primeira etapa a ser considerada é a etapa de DH (seção 2.2), a partir da qual a segunda etapa, DF (seção 2.3), é disparada. A terceira etapa, Identificação de Faces (seção 2.4), atua sobre as faces detectadas na etapa anterior. Finalmente, são apresentadas as principais características de sistemas de RF, encontradas nos mais recentes trabalhos (seção 2.5).

2.1 VISÃO GERAL DO PROBLEMA

A identificação de pessoas em um sistema de segurança, através da identificação de faces, é uma tarefa que abrange diversas disciplinas, como Processamento de Imagem, RP e Visão Computacional. Dentre os principais métodos de reconhecimento biométrico, podem ser citados o reconhecimento de impressão digital, reconhecimento de íris e o reconhecimento de face. O RF possui algumas vantagens sobre os demais:

a) não intrusivo: não necessita obrigatoriamente da cooperação do indivíduo que será reconhecido;
b) *hardware* simples: pode-se utilizar uma câmera simples, que não requer equipamentos específicos, como no reconhecimento de impressão digital ou íris;
c) baixo custo: quando utilizada uma infraestrutura simples, *hardware* simples e *software* livre.

Existem diversas técnicas para o RF, mas, em geral, pode-se dividir os métodos de RF em três módulos: DF, extração de características e classificação das faces em diferentes indivíduos. Para o presente trabalho, no qual será estudado o RF em vídeos, também foi pesquisada a DH, visando melhorar o desempenho.

Como a região espacial que contém a face é relativamente pequena, quando comparada ao tamanho da imagem inteira, o presente trabalho estuda as melhores formas de identificação de humanos na imagem, para que o sistema de RF atue somente na região da imagem que contém humanos. Essa abordagem tem como

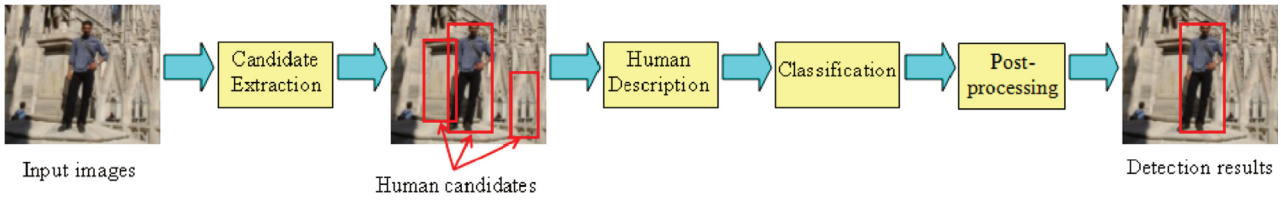
objetivo diminuir os esforços computacionais do RF, visto que, quanto maior a área em que o sistema de RF atuará, maior será o esforço computacional.

2.2 DETECÇÃO DE HUMANOS

Nguyen e Ogunbona (2016) definem o problema de DH como o processo de localizar pessoas automaticamente em uma imagem ou sequência de vídeos. Yang et al. (2011) mostram que, escolher as características adequadas ao tratar o reconhecimento de objetos é uma tarefa crítica. A unicidade é fundamental para que objetos possam sem distinguidos uns dos outros em um espaço de características, tais como: Gradiente, Cor, Textura, Espaço-tempo ou até mesmo a fusão de duas ou mais características.

Em geral, o processo de DH em imagens e vídeos pode ser feito através dos seguintes passos: extrair as regiões candidatas que, possivelmente, contenham humanos, descrever as regiões extraídas, classificar e verificar as regiões como humanos ou não humanos e pós-processar os resultados. Esse processo é ilustrado na Figura 2.1.

Figura 2.1 – Um *framework* genérico para detecção de humanos.



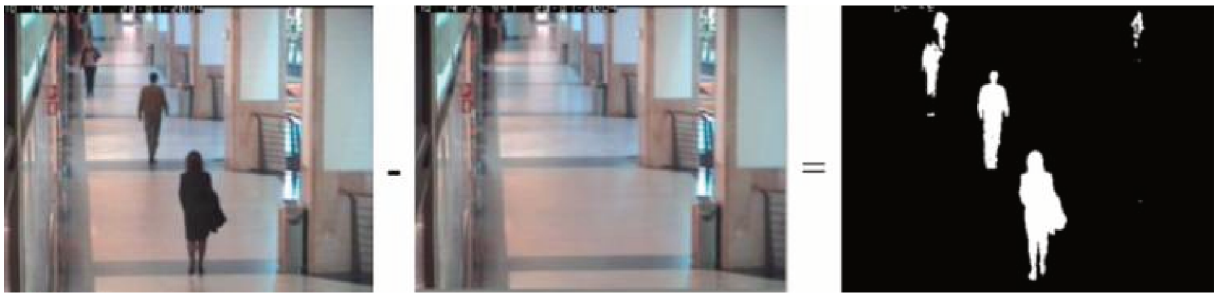
Fonte: (NGUYEN; LI; OGUNBONA, 2016)

A extração das regiões candidatas refere-se ao recorte de uma região da imagem, podendo ser armazenada em uma matriz tridimensional, onde as duas primeiras dimensões indicam a localização dos *pixels* e a terceira dimensão indica a cor ou tom de cinza do *pixel*.

Quando a entrada do sistema de detecção é uma sequência de vídeo, uma técnica bastante utilizada, chamada de *Background Subtraction*, ou Subtração de Fundo, pode ser usada para se enquadrar os objetos humanos candidatos. A utilização da técnica de subtração de fundo caracteriza-se pela sua eficiência e simplicidade (CHEN; WANG; SU, 2014).

Observa-se na Figura 2.2 que, objetos que se movem no vídeo são segregados do fundo através do cálculo da diferença entre a imagem corrente e um fundo utilizado como referência, ou seja, o valor da intensidade dos *pixels* da imagem de fundo é subtraído do valor dos respectivos *pixels* da imagem no plano de frente, que pode conter humanos. Em seguida, a imagem resultante é pós-processada utilizando uma técnica de binarização da imagem, para que seja identificada a região que contém humanos. No entanto, a utilização dessa técnica requer uma câmera fixa e um fundo de referência iniciado sem humanos. Alguns exemplos dessa técnica foram citados por Beleznai e Bischof (2009) e Zhao e Thorpe (2000).

Figura 2.2 – Métodos de extração de regiões candidatas baseado em Subtração de Fundo.



Fonte: (CHEN; WANG; SU, 2014).

A técnica de subtração de fundo pode identificar quaisquer objetos não-humanos que tenham se movimentado sobre a cena fixa. Por esse motivo, essa técnica não representa confiabilidade na identificação de humanos, pois podem surgir Falsos Positivos (FP), isto é, não humanos identificados. No entanto, pode ser muito útil na extração das regiões candidatas que são pós-processadas para se detectar as faces. Os FP podem ser gerados pela movimentação de objetos devido ao vento, passagem de animais ou devido à variação de luminosidade, fazendo com que possam surgir diferenças entre a imagem de fundo, capturada anteriormente, e a imagem sendo capturada no instante.

Embora a extração de regiões candidatas seja útil para aumentar a eficiência da detecção, através da limitação do espaço de pesquisa de objetos humanos, a criação de descritores humanos desempenha um fator chave na eficácia e robustez da DH, como pode ser visto na seção 2.2.1.

2.2.1 Descritores de objetos humanos

Em geral, um descritor humano é composto de características organizadas em uma estrutura, que deve ser capaz de detectar objetos humanos de vários ângulos e poses. As características podem capturar forma, aparência ou informação de movimento, e são capturadas em uma região que pode compreender desde um *pixel* até uma região da imagem. As regiões captadas podem ser distribuídas em uma estrutura densa, como uma grade regular (DALAL; TRIGGS; SCHMID, 2006), ou em uma estrutura esparsa, como em um conjunto de pontos (LEIBE; SEEMANN; SCHIELE, 2005).

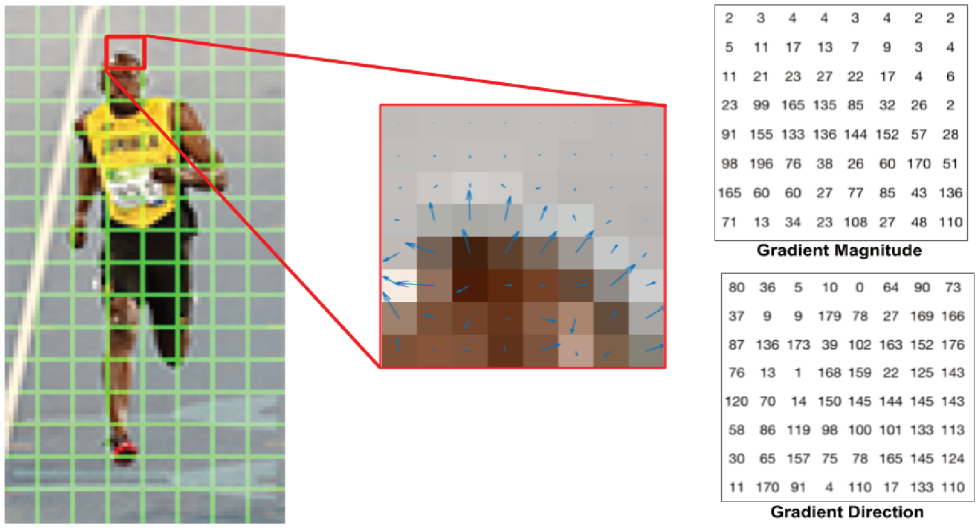
Para descrever a forma do objeto humano, podem ser usadas características baseadas em bordas, onde a localização, orientação e magnitude dos *pixels* de borda são levados em consideração. Exemplos dessa abordagem incluem os métodos descritos por diversos autores, destacando-se Gavrila (1999, 2000), Felzenszwalb, McAllester e Ramanan (2008), Ioffe e Forsyth (2001), Wu e Nevatia (2007), Lin et al. (2007) e Beleznai e Bischof (2009), dentre outros.

Em contraste com as características de bordas de *pixels*, também foram exploradas características de bordas de região, mais adaptáveis à deformação local da forma humana. Um exemplo bem conhecido desta abordagem é o Histograma de Gradientes Orientados (HGO), proposto por Dalal e Triggs (2005).

O HGO é calculado numa região retangular local, na qual cada *pixel* de borda é escolhido em um histograma binário correspondente à orientação da borda. Por exemplo, dado o recorte da Figura 2.3, foram criadas duas matrizes 8×8, uma que contém as direções dos gradientes, de 0 a 180 graus, onde a direção é considerada única para dois sentidos opostos, e outra matriz que contém as intensidades.

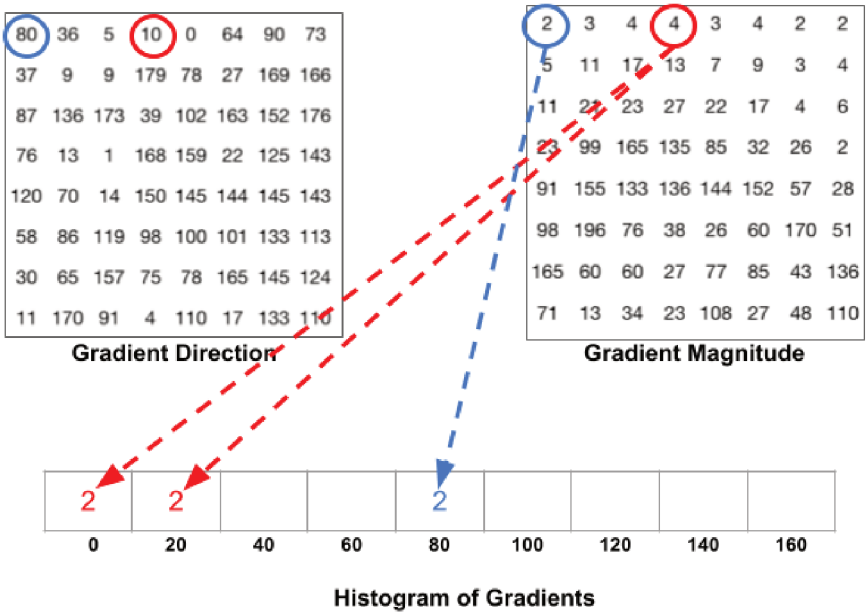
O histograma contém 9 caixas correspondentes aos ângulos 0, 20, 40... 160. A Figura 2.4 ilustra o processo, onde se observa a magnitude e direção do gradiente do mesmo recorte 8×8 da Figura 2.3. Uma posição é selecionada com base na direção e na magnitude. Por exemplo, na Figura 2.4, no *pixel* circulado em azul, existe um ângulo (direção) de 80 graus e magnitude de 2. Assim, a magnitude 2 entra na quarta posição. O gradiente no *pixel* circundado em vermelho tem um ângulo de 10 graus e magnitude de 4. Como 10 graus está entre 0 e 20, a posição do *pixel* divide-se uniformemente nas duas posições.

Figura 2.3 – Gradientes de magnitude e direção (os gradientes são representados como setas ao centro e representados como números à direita).



Fonte: (MALLICK, 2016).

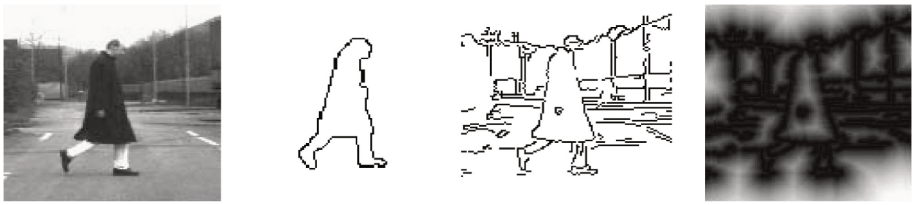
Figura 2.4 – Histograma de Gradientes Orientados na forma de vetor (gradientes de magnitude e direção representados como números, na parte de cima da figura, e como HGO, na parte de baixo).



Fonte: (MALLICK, 2016).

Outra importante abordagem, utilizando forma, é a técnica de *Template Matching* (correspondência de modelos). Essa técnica utiliza as seções de imagens para tentar identificar um modelo dentre diversos modelos pré-definidos, utilizando algum algoritmo detector de bordas. Esse processo é ilustrado na Figura 2.5.

Figura 2.5 – Um processo convencional de *Template Matching (*da esquerda para a direita: imagem original, *template* de bordas, mapa de borda gerado utilizando-se algum detector).

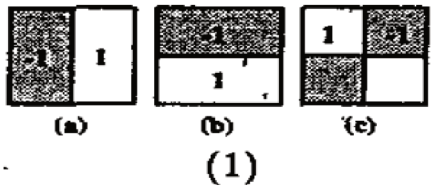


Fonte: (FELZENSZWALB; HUTTENLOCHER, 2004).

Características de aparência são utilizadas principalmente para capturar a cor e textura, e estas também são extraídas em regiões locais da imagem. Segundo Papageorgiou e Orem (1998), detectar objetos do mundo real, tais como faces e pessoas, é bastante complexo devido às variações significativas de cor e textura e das semelhanças entre o plano de fundo com os objetos. A técnica que utiliza exemplos, para construir determinadas características, foi aplicada em seu trabalho para detectar humanos e faces, as chamadas características Haar.

Três tipos de características Haar foram utilizadas, onde cada característica tem a forma de um quadrado. A primeira característica foi representada por dois retângulos, posicionados lado a lado horizontalmente, a segunda por dois retângulos, posicionados lado a lado verticalmente e a terceira com quatro quadrados, formando uma textura quadriculada, como pode ser visto na Figura 2.6 (PAPAGEORGIOU; OREN, 1998).

Figura 2.6. Três tipos de características Haar utilizadas.



(1)

Fonte: (PAPAGEORGIOU; OREN, 1998).

Essas características foram determinadas através da análise estatística das características encontradas na análise de 2429 imagens de faces, de tamanho 19x19 *pixels*. O valor de uma característica de dois retângulos é a diferença entre a soma dos *pixels* das duas regiões retangulares.

Outros métodos importantes também têm sido utilizados, como o *Local Binary Pattern* (*LBP),* que foi utilizado para descrever a aparência do corpo humano

nas pesquisas de Hussain et al. (2010), sendo bem conhecido pela sua robustez contra mudanças de iluminação, poder discriminativo e simplicidade computacional.

A cor também tem sido considerada uma característica importante. Por exemplo, na pesquisa de Ott e Everingham (2009), o recurso HGO foi calculado em uma parte da imagem, que foi segmentada a partir de uma imagem de entrada, através de uma técnica de distribuição de cores de fundo e de primeiro plano.

A disponibilidade de informação de movimento também pode ser explorada em descritores humanos, como em pedestres que realizam movimentos cíclicos ao andar (VIOLA; JONES; SNOW, 2005), ou através de fluxos ópticos (DALAL; TRIGGS; SCHMID, 2006).

2.2.2 Classificadores e algoritmos de aprendizagem

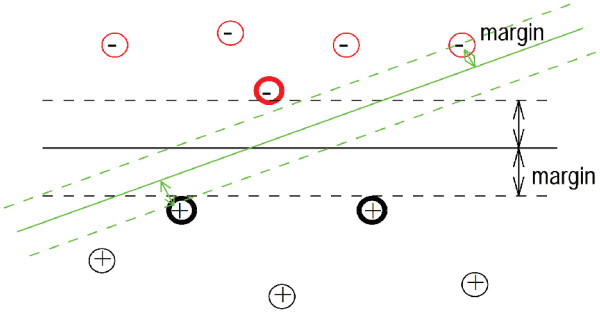
Uma vez que os descritores de humanos são extraídos das regiões candidatas, o passo de classificação deve ser executado para classificar as regiões como sendo humanas ou não humanas. As classificações podem seguir as abordagens generativas ou discriminativas.

Métodos generativos objetivam construir um modelo do objeto de interesse, como exemplo, nos modelos de forma, pode ser citado o método baseado em correspondência de padrões (*Template matching-based*) (GAVRILA, 2007; LIN et al., 2007), modelos estruturais (MIKOLAJCZYK; SCHMID; ZISSERMAN, 2004), que mostra como o objeto humano foi identificado, "*Implicit Shape Model*" (ISM), ou modelagem de forma implícita, proposta por Leibe et al. (2005).

Métodos discriminativos têm sido usados desde que as classificações de regiões candidatas foram consideradas um problema de classificação binária. *Support Vector Machines* (SVM) são frequentemente usados (TRIGGS, 2005; DALAL; TRIGGS; SCHMID, 2006) para classificar descritores humanos e não humanos, através da maximização da margem entre essas duas classes, como pode ser visto na Figura 2.7 (ZHANG, 2017).

Assim, o conjunto de treinamento positivo utilizado para treinar cada SVM linear foi determinado por agrupamento, no qual cada grupo representa um grupo de exemplos de objetos humanos em algumas poses.

Figura 2.7 – *Support Vector Machines* – SVM (o conjunto de exemplos positivos, abaixo, foi separado linearmente dos conjuntos de exemplos negativos, acima, através da maximização da margem entre as classes).



Fonte: (ZHANG, 2017).

2.3 DETECÇÃO DE FACES

Detecção automatizada de faces pode ser considerada como o pilar de aplicações envolvendo análise facial, não somente limitada às aplicações de RF e verificação, mas também pode ser aplicada no rastreamento de pessoas em sistemas de segurança, análise comportamental de faces, reconhecimento de atributos de faces (tais como identificação de sexo e idade e aplicações para avaliação de beleza), reconstrução facial, organização de faces em álbuns digitais, dentre outros.

A DF é a etapa inicial para todos modernos sistemas de interação baseados em visão computacional, de interação humano-computador e robô-computador (por exemplo, o robô comercial “Nao”, que vêm com um módulo de DF (ROBOTICS, 2014).

Ainda, podem ser citadas as mais modernas câmeras digitais que contém recursos de DF para realizar o foco automático, assim como o recurso de DF utilizado no Facebook para a marcação de pessoas (ZAFEIRIOU; ZHANG; ZHANG, 2015).

Os primeiros trabalhos de DF não obtiveram muito sucesso em ambientes não controlados, também chamados de *in-the-wild*, ou seja, ambientes sem nenhum tipo de controle de iluminação ou ambientação. A DF, nesse tipo de ambiente, é um problema que vem sendo estudado e que se aplica a sistemas aplicados ao mundo real. Tal eficácia foi conseguida pela primeira vez no trabalho de Viola e Jones

(VIOLA; JONES, 2001), o qual mostrou o primeiro algoritmo que tornava possível a DF em ambientes *in-the-wild* em sistemas de DF *real time.*

A última revisão abrangente de algoritmos de DF até a data atual, organizou algoritmos nas duas categorias principais a seguir (ZAFEIRIOU; ZHANG; ZHANG, 2015):

a) a família de algoritmos baseados em modelos rígidos, que incluem:
   - variações de *boosting*: o principal representante dessa família de algoritmos inclui o algoritmo de DF de Viola-Jones e suas variações (VIOLA; JONES, 2001, 2004).
   - algoritmos baseados em Redes Neurais de Convolução (RNC) e RNCP: recentemente, as RNCP mostraram desempenho excepcional na detecção de classes multi-objeto (GIRSHICK et al., 2014; REDMON et al., 2016) e atualmente suas arquiteturas de aprendizado têm sido investigadas para DF (HE et al., 2017; JALALI; MALLIPEDDI; LEE, 2017; SCHROFF; KALENICHENKO; PHILBIN, 2015; TRIANTAFYLLIDOU; NOUSI; TEFAS, 2018; YANG et al., 2017).
   - Métodos que aplicam estratégias inspiradas na recuperação de imagem (LI et al., 2014; SHEN et al., 2013) e transformação de massa generalizada (BALLARD, 1981; LEIBE; LEONARDIS; SCHIELE, 2008).

b) a família de algoritmos que aprende e aplica um modelo baseado em peças deformáveis para modelar uma potencial deformação entre as partes faciais (FELZENSZWALB et al., 2009; FELZENSZWALB; HUTTENLOCHER, 2005; FISCHLER; ELSCHLAGER, 1973). Esses métodos também podem combinar a DF e a localização das partes faciais (ZHU; RAMANAN, 2012).

Trabalhos recentes na DF, utilizando RNCP, como o trabalho de Triantafyllidou e Tefas (2018), têm superado constantemente o estado da arte na DF, inclusive o trabalho de Viola e Jones (2001). Em seu trabalho é apresentada uma nova RNCP leve para DF. No entanto, sua implementação requer tempo e conhecimentos específicos, diferentemente das técnicas de Viola e Jones (2001), que já estão estabilizadas em bibliotecas de código livre e são muito utilizadas pela comunidade de desenvolvedores e pesquisadores da área, como a biblioteca OpenCV (OPENCV, 2018).

2.3.1 Detecção de faces utilizando características Haar

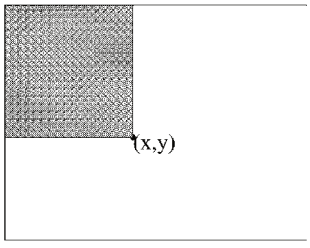
Um dos trabalhos mais citados utilizando características Haar é o trabalho de Viola e Jones (2001), no qual foi possível detectar uma face de 24x24 *pixels* em 0,067 segundos em uma imagem de 384×288 *pixels*, utilizando um processador Pentium III de 700MHz. Nessa direção, foi construído um sistema de DF que atingiu taxas altas de acerto com um baixo número de FP, equivalendo ou superando as taxas de outros trabalhos da época, porém sendo até 15 vezes mais rápido.

A primeira contribuição do trabalho é uma nova representação de imagem chamada de imagem integral. A segunda contribuição é um simples e eficiente classificador utilizando características Haar e o algoritmo Adaboost. A terceira contribuição é um método para combinar classificadores em uma estrutura de cascata.

A integral de imagem é um algoritmo, também conhecido como tabela de soma de áreas (CROW, 1984). Características de retângulos podem ser calculadas muito rapidamente, ao utilizar uma representação intermediária da imagem, chamada de integral de imagem.

A integral da imagem, na posição (*x*, *y*), contém a soma dos *pixels* acima e à esquerda de *x*, *y*, inclusive. A Figura 2.8 exibe uma integral de imagem, onde o ponto *x*, *y* contém a soma de todos os níveis de cinza da região cinza, desde a origem (0, 0) até o ponto (*x*, *y*) (P. VIOLA, 2001).

Figura 2.8 – Integral da imagem (O valor da integral de imagem é a soma dos *pixels* acima e à esquerda do ponto (*x*, *y*)).



Fonte: (VIOLA; JONES, 2001)

A equação 2.1 indica como calcular a integral de imagem em uma determinada coordenada:

$$ii(x,y) = \sum_{\acute{x} \leq x, \acute{y} \leq y} i(\acute{x}, \acute{y}) \quad (2.1)$$

onde *ii*(*x*, *y*) é a integral da imagem nas coordenadas do pixel (*x*, *y*) e i(*x'*, *y'*) é a imagem original.

A partir da integral da imagem é possível identificar padrões utilizando características do tipo Haar. Cada tipo de característica pode ajudar a reconhecer um determinado padrão, por exemplo, a característica central da Figura 2.9 permite identificar uma área na imagem onde há uma diferença de intensidade significativa entre a parte superior e a parte inferior de uma região, como, por exemplo, na região dos olhos, que é mais escura do que a região das bochechas.

Figura 2.9 – Melhores características Haar aplicadas à imagem.

Fonte: (P. VIOLA, 2001).

As informações das características Haar são, então, armazenadas, para que sejam utilizadas na DF.

2.3.2 O algoritmo Adaboost

O algoritmo AdaBoost para aceleração de treinamento, com arquitetura em cascata, foi utilizado pela primeira vez para DH no trabalho de Viola e Jones (2011). Dado um conjunto de características, deve-se treinar o sistema com imagens positivas (faces) e imagens negativas (tudo menos faces), utilizando um algoritmo de aprendizagem que utilize as características Haar durante o processo de aprendizado.

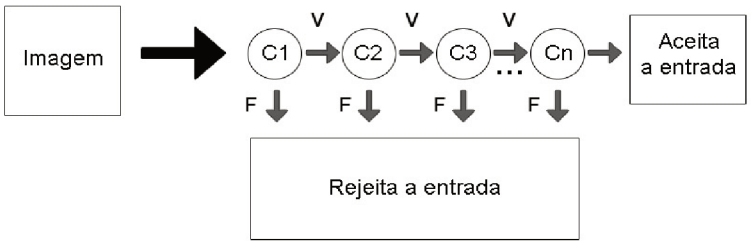
A denominação do algoritmo AdaBoost é derivado de *Adaptative Boosting*, sendo um método de aprendizado de máquina que utiliza a combinação de vários classificadores fracos para obter uma classificação forte. O *boosting* é utilizado tanto para selecionar um conjunto de características como para treinar o classificador.

Durante o treinamento, as características retangulares são localizadas. Em seguida, através da utilização da imagem integral, é verificado se as características podem ser aplicadas a uma região da imagem. Os contrastes naturais proporcionados pelas características da face são examinados, considerando suas relações de espaço.

A Figura 2.9 exibe os dois melhores atributos encontrados durante o treinamento. O processo de treinamento seleciona as características mais coincidentes nas mesmas localizações dentre todas as imagens treinadas com faces, de forma a criar classificadores mais fortes, utilizando as principais características.

Após o processo de treinamento, as faces são detectadas utilizando o algoritmo em cascata, onde os classificadores fortes em cascata processam as regiões da imagem em busca de um padrão. Cada estágio na cascata aplica um classificador mais específico e complexo do que o anterior, de modo que o algoritmo rejeite rapidamente regiões que sejam muito distintas da característica procurada e termine o processo de procura neste caso, evitando que os estágios posteriores sejam executados desnecessariamente. Isso faz com que muitos dos cenários e panos de fundo sejam descartados nos primeiros estágios e apenas faces e outros objetos semelhantes a faces sejam analisados mais exaustivamente, como pode ser observado na Figura 2.10 (VIOLA; JONES, 2001).

Figura 2.10 – Classificadores em cascata.



Fonte: (BRAGA, 2013).

O escaneamento do detector em múltiplas localizações é feito através do deslizamento de uma janela a cada determinado espaço de deslocamento na imagem. A escolha da escala é feita na etapa de detecção e afeta tanto a velocidade do detector quanto a sua acurácia. Quanto menor a escala, mais quadros são processados e uma melhor taxa de faces é atingida, porém uma menor velocidade é observada no detector.

Além da escala, outros parâmetros também podem ser ajustados utilizando a biblioteca OpenCV (OPENCV, 2018):

a) *scaleFactor* – especifica o quanto a imagem é reduzida a cada trecho da imagem escaneada pelo detector;
b) *minNeighbors* – parâmetro que especifica a quantidade mínima de vizinhos que cada face candidata deve conter para que seja considerada uma face;
c) *minSize* – tamanho mínimo da região que pode conter faces;
d) *maxSize* – tamanho máximo da região que pode conter faces.

Durante o treinamento realizado no trabalho de Viola e Jones (2001), foram utilizadas apenas imagens de faces com alguma variação de rotação e inclinação, o que fez com que o detector encontrasse somente imagens de faces inclinadas em até 15 graus e com até 45 graus de rotação. Por esse motivo, o detector não se apresenta confiável com rotações e inclinações maiores do que essas.

Faces com pouca luminosidade também causam falhas em algumas situações. O detector também se demonstra bastante falho ao testar imagens com faces em oclusão, como, por exemplo, com oclusões sobre a área dos olhos, provavelmente devido à utilização de poucas imagens de faces com oclusão durante o treinamento, mas o real motivo não foi citado no trabalho.

2.3.3 Métricas de avaliação

No processo de avaliação do resultado da classificação de imagens, como sendo faces ou não, duas medidas são importantes: a quantidade de objetos identificados incorretamente como face (FP) e a quantidade de faces que não foram identificadas (Falsos Negativos) (VIOLA; JONES, 2001, 2004). Estas medidas são adotadas neste trabalho.

2.3.4 Considerações

A seção apresentou os principais avanços da última década sobre a DF. Apesar do ótimo desempenho em detectar faces utilizando RNCP, como no artigo de Triantafyllidou et al. (2018), a simplicidade de utilização e a pronta disponibilidade de bibliotecas que implementam o algoritmo de Viola e Jones (2001), fazem com que esta técnica continue sendo amplamente utilizada para esta tarefa.

2.4 IDENTIFICAÇÃO DE FACES

A próxima etapa do reconhecimento de faces, a identificação de faces, consiste na utilização das imagens de faces, selecionadas pelo processo de DF, e na classificação destas imagens de faces, onde cada classe representa um indivíduo distinto, previamente cadastrado em um banco de dados.

Uma imagem contendo uma face contém uma grande quantidade de informações, isto é, quanto maior a sua dimensão, mais *pixels* são observados e, portanto, uma maior quantidade de informação precisa ser processada para se realizar a identificação de faces. Por esse motivo, comumente a redução de dimensionalidade é necessária para gerar características que representem as faces.

Em seguida, deve-se classificar as faces de forma que cada classe represente um indivíduo, utilizando uma métrica. Diversos algoritmos de classificação podem ser utilizados para a tarefa de classificação, como *K-Nearest Neighbors, Random Forest, MultiLayer* Perceptron e RNCP.

Para este trabalho, foram utilizadas as RNCP para classificação e, por esse motivo, elas foram detalhadas em uma seção à parte (seção 3).

Os sistemas de RF se enquadram em duas categorias: verificação e identificação. A verificação de face é uma correspondência de 1:1, que compara uma imagem de face com uma imagem modelo de rosto, cuja identidade está sendo reivindicada. Diferentemente, a identificação de face é um problema 1:N, que compara uma imagem de face de consulta com todas as imagens de faces em um banco de dados de rosto para determinar a identidade da face de consulta.

Em sistemas de segurança comerciais, normalmente se deseja identificar uma pessoa dentre todas as cadastradas, que pode ser uma imagem por pessoa ou várias imagens por pessoa. Umas das grandes dificuldades em se implementar

sistemas de segurança comerciais é a captura de várias faces de um mesmo indivíduo. Por exemplo, em um sistema de segurança de fronteira ou aeroportos, uma única imagem, a do passaporte, é utilizada para a comparação. Por esse motivo, estudos de identificação de faces utilizando uma imagem por pessoa têm sido pesquisados recentemente, e diversas técnicas, que utilizam uma imagem por pessoa, têm sido pesquisadas, tais como *PCA*, *LDA*, *LBP* e *SVM* (HU et al., 2015).

Apesar da grande quantidade de pesquisas de sistemas que utilizam uma face por pessoa, especialmente antes de 2010, como exibe a pesquisa exploratória de Tan et al. (2006), as RNCP têm sido utilizadas em trabalhos que são o estado da arte no RF *in-the-wild*, podendo requerer várias faces por pessoa. (SCHROFF; KALENICHENKO; PHILBIN, 2015). A utilização desta técnica pode ser aplicada em sistemas de segurança controlados, onde há um prévio registro de diversas faces, podendo ser capturadas através da utilização de vídeos.

Na seção 2.4.1, são apresentadas as principais dificuldades em reconhecer faces. Em seguida, a seção 2.4.2 aborda o RF utilizando uma imagem como referência, e a seção 2.4.3 aborda o RF com várias imagens como referência.

2.4.1 Dificuldades encontradas durante o reconhecimento de faces

Os recentes trabalhos realizados em bancos de dados populares contendo faces têm superado a marca de 90% de TRF. No entanto, o mesmo não pode ser dito em sistemas de segurança em ambientes não controlados, pois a imagem pode sofrer diversos tipos de variações durante sua aquisição. Segundo ABATE et al. (2007), há cinco fatores que podem afetar significativamente a performance do RF:

a) variações da iluminação, devido ao reflexo da luz sobre a pele em diferentes poses da cabeça;
b) variações de pose da cabeça, como rotação e inclinação da cabeça;
c) variações devido ao aumento de idade;
d) oclusão da face, devido a utilização de chapéus, óculos etc.

2.4.2 Identificação de face utilizando uma imagem

A utilização de apenas uma imagem contendo uma face, como modelo para a identificação de faces, tem sido pesquisada desde os primeiros trabalhos

realizados que investigam o RF. Para se ter uma ideia das dificuldades que podem ser encontradas, imagina-se colocar uma foto sobre outra, de uma mesma pessoa. Em primeiro lugar, as fotos podem não ser do mesmo tamanho, as faces não estão alinhadas, as faces não têm o mesmo tamanho, inclinação e rotação. Também, encontram-se divergências na saturação, brilho e contraste.

Por esses motivos, os trabalhos de RF, que utilizam uma imagem, aplicam um pré-processamento na imagem para melhorar a classificação. Após o pré-processamento da imagem, características são extraídas através da redução de dimensionalidade da imagem.

2.4.2.1 Pré processamento de Imagem

Com o objetivo de fazer com que duas faces em duas imagens diferentes sejam parecidas visualmente, requer-se um pré-processamento nas imagens. As técnicas utilizadas são o alinhamento geométrico das faces das imagens, ajuste de brilho, contraste, saturação ou luminância, de forma a normalizar as imagens para que o reconhecimento se torne o mais eficaz possível.

A luminância é padronizada pelo CIE (Comissão internacional de iluminação (CIE, 2019)) como intensidade física da luz na região do espectro visível, ponderada por uma curva de sensibilidade espectral que corresponde à sensação de brilho do sistema de visão humano. A recomendação ITU-R BT.709 define que luminância pode ser computada a partir das primárias RGB para sistemas de vídeo e computação gráfica modernos, sendo denominada Luminância Relativa, como pode se observar na equação 2.2 (LEE, 2018).

$$Y = (0{,}2125\,x\,R) + (0{,}7154\,x\,G) + (0{,}0721\,x\,B) \tag{2.2}$$

onde: Y é uma matriz que representa a Luminância Relativa, R é matriz bidimensional de tons vermelhos da imagem, G é a matriz bidimensional de tons verdes e B a matriz bidimensional de tons azuis.

Em diversos algoritmos, como *FischerFaces* (*LDA*) e *EigenFaces* (*PCA*), detalhados na seção 2.4.2.2, é necessário ajustar o posicionamento da face dentro

de uma imagem, por exemplo, fazendo com que os olhos de todas as imagens fiquem posicionados nas mesmas coordenadas.

Ainda, a conversão da imagem para escala de cinza, antes de realizar a extração de características, pode ser feita antes da aplicação de outras técnicas, como a equalização do histograma, um método muito utilizado para normalizar a distribuição da probabilidade de ocorrência de valores de intensidade na imagem (SHAFEY, 2017).

2.4.2.2 Principais técnicas de Identificação de faces utilizando uma imagem

As principais características que representam as faces, dentro de um grande volume de informações (*pixels)* em uma imagem contendo face, podem ser representadas de forma compacta, para que os sistemas de RF processem a menor quantidade de informações possível, tornando-se assim mais eficientes e rápidos. Recentemente, muitos métodos de redução da dimensionalidade do espaço de características foram desenvolvidos (ANGADI; KAGAWADE, 2017).

A técnica *Principal Component Analysis* (*PCA*) é uma técnica que pode ser aplicada holisticamente ou localmente, e é provavelmente a técnica mais utilizada em trabalhos de RF, como no trabalho de Grgic et al. (2011), onde a técnica foi utilizada para comparar a qualidade do RF em diversos bancos de dados populares. A técnica *PCA* atua na redução da dimensionalidade, maximizando a dispersão entre os elementos (BELHUMEUR; HESPANHA; KRIEGMAN, 1997).

PCA, quando aplicado ao RF, recebe o nome de *Eigenfaces*. Cada imagem facial é representada como um vetor de características (*feature vector*) denominado *Eigenfaces*, armazenados em vetores unidimensionais (TURK, 1991). Uma desvantagem desta abordagem é que a dispersão maximizada dos elementos não é apenas a dispersão entre as classes, mas também a dispersão dentro da mesma classe, o que é prejudicial para a classificação.

Por esses motivos, como se pode constatar no trabalho de Grgic et al. (2011), observa-se que a técnica *PCA* não se adapta a diferentes condições de iluminação, variações de pose, tamanho da face, variações de expressões faciais e a utilização de acessórios, tais como óculos e chapéus.

O método *Linear Discriminant Analysis* (*LDA*), também conhecido como *Fisher's Linear Discriminant*, ou *Fischerfaces,* quando aplicado ao RF, pode servir

para achar uma combinação linear de características que caracteriza ou separa duas ou mais classes de objetos ou eventos. O *LDA* reduz o espaço de forma que sejam selecionadas as características mais discriminantes entre as classes, superando as deficiências do método *PCA*. O método *Fisherfaces* realiza o reconhecimento através de uma projeção linear que combina *PCA* e *LDA*, mostrando-se bem-sucedido nesse sentido.

O método *Local Binary Pattern* (*LBP*) tem sido utilizado para extrair características faciais, atuando como um descritor de aparência local que é capaz de capturar detalhes de aparência e textura facial (AHONEN; HADID; PIETIKÄINEN, 2006). Devido à sua robustez às variações de iluminação, expressão facial, envelhecimento e outras mudanças, o *LBP* alcançou o desempenho de RF de última geração no cenário em que apenas uma amostra por pessoa é usada para treinamento (WAGNER et al., 2012).

Os filtros Gabor também representam uma poderosa ferramenta, tanto no processamento de imagens quanto na codificação de imagens, graças à sua capacidade de capturar recursos visuais importantes, como localização espacial, frequência espacial e seletividade de orientação. Na maioria dos casos, os filtros Gabor são usados para extrair as principais características das faces, como no trabalho de Lades et al. (1993).

Para o RF 3D, entende-se uma classe de métodos que trabalham em um conjunto de dados tridimensional, representando a forma da face e da cabeça como dados de intervalo ou malhas poligonais. A principal vantagem das abordagens baseadas em 3D é que o modelo 3D mantém todas as informações sobre a geometria da face (ABATE et al., 2007).

Como se pode observar, as técnicas que utilizam uma face como modelo único de representação, requerem pré-processamentos nas imagens, que necessitam de atenção aos resultados. Sistemas que utilizam imagens 3D são ainda mais complexos, pois requerem o mapeamento dos pontos tridimensionais, além da necessidade de equipamentos especiais.

2.4.2.3 Algoritmos Classificadores

Após a etapa de criação das características e redução da dimensionalidade, algoritmos classificadores são utilizados para classificar as faces entre as classes

disponíveis, isto é, entre as diferentes pessoas dentro do conjunto de faces utilizadas na classificação. Os principais algoritmos são descritos em seguida.

O Classificador *K-Nearest Neighbors* (K-NN) é um dos algoritmos de classificação mais utilizados na área de aprendizagem de máquina, principalmente quando aplicados na classificação de faces. Seu método se baseia na procura dos *k* vizinhos mais próximos do padrão de teste. A busca pela vizinhança é realizada utilizando uma medida de distância, como a medida Euclidiana, observada nos trabalhos de Hu et al. (2015), distância de Manhattan ou a Euclidiana normalizada (WITTEN; FRANK; HALL, 2011).

O Classificador *Random Forest* utiliza uma técnica de agregação de classificadores do tipo árvore, construídas de maneira aleatória. A classe de uma instância é determinada através da combinação de várias árvores de decisão, por meio de um mecanismo de votação. Cada árvore dá uma classificação, ou um voto para uma classe. A classificação final é dada pela classe que recebeu o maior número de votos entre todas as árvores da floresta (BREIMAN, 1999).

O Classificador *K-Star* é um classificador baseado em exemplos, isto é, baseia-se na classe das instâncias de formação semelhante, determinada por uma função de similaridade. Ele difere de outros por utilizar funções de distância baseadas na entropia e assume que os exemplos similares terão classes similares (WITTEN; FRANK; HALL, 2011).

2.4.3 Identificação de faces utilizando mais de uma imagem

Recentemente, técnicas de Aprendizado de Máquina (*Machine Learning – ML*), em especial as RNA, têm sido utilizadas para treinar a classificação de objetos, em classes genéricas como cachorros, carros, humanos etc, ou realizar a classificação dentro de uma classe específica, como é o caso do RF. Isso ocorreu principalmente devido a uma maior disponibilidade de bases de dados com milhões de imagens e o aumento na capacidade computacional (RUSSAKOVSKY et al., 2015).

Apesar da dificuldade em se utilizar mais de uma imagem de face por pessoa em sistemas de segurança com grande volume de dados, essa abordagem pode ser útil em sistemas de segurança que controlem algumas centenas ou alguns milhares de pessoas em instituições ou empresas. Nessas condições, um sistema

de segurança pode coletar faces constantemente através de vídeos de câmeras e utilizar algumas delas para atualizar a base de dados com uma frequência pré-determinada.

A abordagem em que se utiliza mais de uma imagem por pessoa será feita em uma seção à parte, seção 3 – Redes Neurais Convolucionais Profundas, um tipo de RNA utilizada no protótipo deste trabalho, para introduzir os principais conceitos de RNA e suas variações aplicadas ao RF.

2.4.4 Bancos de dados e protocolos de avaliação

Para garantir uma comparação justa de algoritmos, é necessário que todos os algoritmos recebam as mesmas imagens a serem utilizadas durante o treinamento e avaliação, e que exista um protocolo de avaliação. Os bancos de dados podem ser divididos em duas categorias: bancos de dados controlados e não controlados (*in-the-wild)*.

Os bancos de dados controlados são os bancos de dados nos quais a iluminação, pose e oclusão são gerenciados. A maioria deles também realiza um pré-processamento nas imagens para realizar o alinhamento da face na imagem, ou o alinhamento dos pontos faciais.

Os bancos de dados *in-the-wild* têm pouco ou nenhum controle sobre as faces capturadas. Esses tipos de bancos de dados normalmente são criados a partir de imagens ou vídeos retirados da internet.

Para realizar os mesmos tipos de testes em um banco de dados, normalmente os bancos de dados reportam seus resultados em termos de TRF (ou *Recognition-Rate*) ou em termos de quantidade de características coincidentes (*Cummulative Match Characteristics*). Alguns protocolos também dividem os dados em três conjuntos: treinamento, desenvolvimento e avaliação. Outros protocolos dividem as imagens em diversas categorias como variações na iluminação, oclusão ou pose, permitindo que algoritmos possam ser avaliados nessas condições.

2.4.4.1 Bancos de dados controlados

Os bancos de dados controlados agrupam as imagens para que seja possível testar a qualidade de algoritmos sob a mesma condição. Os agrupamentos

podem ser feitos por exemplos, rotação da face, oclusão, iluminação degradada, gêneros ou idade.

Por exemplo, o banco de dados FERET foi coletado em 15 sessões entre agosto de 1993 e julho de 1996. Três categorias principais foram utilizadas: Imagens de faces com óculos, imagens com mudanças na iluminação e imagens de faces com diferenças de tempo entre uma e outra seção, totalizando 14.126 imagens de 1199 indivíduos. O lapso de tempo foi importante porque permitiu aos pesquisadores estudar, pela primeira vez, mudanças na aparência de um sujeito que ocorrem ao longo dos anos (JONATHON PHILLIPS et al., 2000).

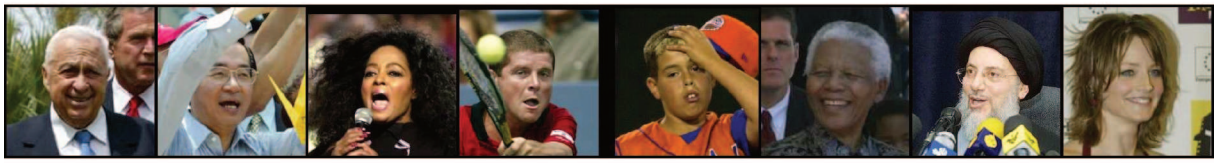
Outro bem conhecido banco de dados foi coletado pelos pesquisadores da Carnegie Mellon University, o CMU Multi-PIE, que contém 337 pessoas, capturadas em 15 pontos de vista e 19 condições de iluminação em quatro sessões de gravação, totalizando mais de 750.000 imagens. O banco de dados é um exemplo dos bancos de dados que não possui protocolos de avaliação (GROSS et al., 2010).

2.4.4.2 Bancos de dados não controlados (*in-the-wild)*

Com o objetivo de se testar o RF em condições não controladas, diversos bancos de dados têm sido criados em trabalhos que requereram essas variações na imagem.

O banco de dados *Labeled Faces in-the-wild* (LFW) (HUANG et al., 2007; HUANG; LEARNED-MILLER, 2014), fornece um grande conjunto de imagens faciais, relativamente sem restrições, isto é, rostos que mostram uma grande variação da vida cotidiana, incluindo variações de pose, iluminação, expressão, plano de fundo, raça, etnia, idade, sexo, roupas, penteados, qualidade da câmera, saturação de cor, oclusão e outros parâmetros, como visto na Figura 2.11. A distribuição de imagens por pessoa também é variável, o que o torna um bom candidato para testes em sistemas de segurança em ambientes não controlados.

Figura 2.11 – Imagens do banco de dados *Labeled Faces in-the-wild*



Fonte: (HUANG; LEARNED-MILLER, 2014)

No trabalho de Grgic et al. (2011), foi criado o banco SCface – *surveillance cameras face database,* onde foram armazenadas 4160 imagens estáticas de 130 indivíduos sob condições de iluminação descontrolada. Imagens de câmeras, com diferentes qualidades e resolução, imitam condições do mundo real e permitem testes robustos de algoritmos de RF, enfatizando diferentes cenários de aplicação de lei e vigilância.

2.5 SISTEMAS DE RECONHECIMENTO FACIAL

O RF é realizado por sistemas de RF, que capturam imagens em um ambiente e as processam, através de um *software* para esse propósito, podendo ser utilizados para videovigilância, controle de acesso ou outras finalidades. Os sistemas de RF são compostos por um *hardware,* capaz de capturar as imagens e armazená-las, através da utilização de câmeras IP, materiais de rede e computadores, e por um *software,* capaz de realizar o RF utilizando as imagens digitais armazenadas. Comumente, a captura é realizada em um Circuito Fechado de Televisão (seção 2.5.1).

2.5.1 Circuito Fechado de TV

Tradicionalmente, a vigilância por vídeo também é conhecida como Circuito Fechado de Televisão (CFT). Apesar do forte apelo para a área de segurança, os sistemas de CFT não estão restritos somente a esse tipo de aplicação, tendo sido utilizados também no controle de tráfego, de multidões e de pacientes, em laboratórios de pesquisa, treinamentos, controle de produção em indústrias, avaliações de desempenho profissional, gerenciamento de informações, dentre outras inúmeras aplicações (BEZERRA, 2012).

Depois de 2001, especialmente após o ataque de 11 de setembro, os programas de RF e outros avanços digitais se tornaram uma prioridade maior. Câmeras de vigilância, com acesso através da Internet, tornam-se cada vez mais comuns. Hoje em dia, os sistemas de videovigilância são muito mais avançados. Com a Internet e redes sem fio, a vigilância por vídeo pode ser usada e monitorada em qualquer lugar do mundo (ZHANG, 2017).

2.5.1.1 Características de câmeras de CFT

Câmeras de CFT têm como função capturar o sinal de vídeo ou imagens para a equipe de segurança localizada na central de controle. Os principais componentes das câmeras são detalhados a seguir.

As lentes em uma câmera têm por função direcionar a luz refletida do ambiente registrado para o sensor de imagem da câmera. Os tipos de lentes encontradas no mercado podem ser divididos em:

a) íris fixa: é o tipo mais simples de lente, possuindo somente ajuste de foco. Pode ser utilizada em ambientes em que a iluminação não se altera de forma excessiva. Não é recomendável para sistemas de CFT;
b) íris manual: permite ajuste da luz refletida, permitindo direcionar a quantidade ideal de iluminação a ser captada pelo sensor de imagem. Pode ser utilizada para ambientes com iluminações muito intensas ou pouco intensas;
c) auto íris: permite o ajuste automático da íris da lente de acordo com o nível de iluminação do ambiente. Possuem motores e sistemas de verificação que definem quando a íris deve ser aberta ou fechada;
d) vari focais: permitem o ajuste de sua distância focal (zoom manual). Podem apresentar íris manual ou automática. Recomendável para sistemas de CFT, pela flexibilidade de ajustes para o dimensionamento do projeto, no entanto são as lentes mais caras do mercado.

O sensor de imagem é o dispositivo eletrônico encontrado no interior da câmera, que contém elementos sensíveis às variações de iluminação, convertendo a imagem visual da cena observada pela lente da câmera em sinais elétricos. Atualmente, há dois tipos principais de sensores de imagem encontrados em câmeras de CFT: CMOS (*Complementary Metal Oxide Semiconductor*) e CCD (*Charged Coupled Device*). São comuns os formatos de área do CCD de 1/2'', 1/3'' e 1/4'' (polegadas na diagonal). Os CCDs de 1/4″ são os mais modernos atualmente.

A resolução da câmera pode ser definida como a clareza de detalhes em que uma imagem pode ser distinguida. As unidades de medida mais comuns para resolução espacial são os pares de linhas por unidade de medida, número total de *pixels* ou *pixels* por unidade de medida (BEZERRA, 2012).

Outra característica importante das câmeras refere-se à quantidade mínima de iluminação, que é medida em lux. Em condições de baixa iluminação no ambiente, quanto menor o parâmetro de lux da câmera, melhor será a imagem registrada. (BEZERRA, 2012).

O Controle de Ganho pode ser utilizado para ajustar a iluminação do ambiente, no entanto, a sua utilização pode comprometer a qualidade das imagens. O contrário pode ser dito da Compensação de Luz de Fundo, sendo que esta pode funcionar como um atenuador de iluminação do ambiente, melhorando a definição da imagem captada (SWGIT, 2014).

A utilização do modo de vídeo, preto e branco, infravermelho ou colorido, também pode variar de acordo com a técnica de RF utilizada. Por exemplo, as RNCP podem requerer vídeo colorido, mas outras técnicas, como a *PCA,* não necessitam de cores.

Câmeras utilizadas em sistemas de segurança internos não são tão suscetíveis a variações de umidade, poeira, sujeiras, fumaça, além de pequenas variações de temperatura. O nível de proteção IP66, regulamentado pela norma IEC (*International Electrotechnical Commission*) 60529, define os valores de proteção contra poeira e água. Os números (níveis) 6, indicam proteção contra poeira e proteção contra jatos potentes de água, respectivamente.

2.5.1.2 Ambiente de captura

A definição do espaço físico de captura do vídeo é uma tarefa importante para se projetar as regiões ótimas a serem capturadas pelo sistema de segurança. A prévia identificação das características das câmeras e da qualidade da imagem capturada, devem ser feitas de antemão, ao se projetar o ambiente.

A região de atuação de captura de imagens deve ser compatível com a resolução da câmera escolhida, dado que, a resolução da imagem da face capturada, deve ser boa o suficiente para que os sistemas de RF possam alcançar uma taxa de acerto satisfatória. Alguns autores citam que a resolução mínima de uma imagem para se reconhecer uma face é de 19x19 *pixels* (GRGIC; DELAC; GRGIC, 2011). De acordo com Bezerra (2012), o RF através do método utilizado em seu trabalho, *PCA Eigenfaces*, requer uma imagem, contendo somente a face, de uma largura mínima de 40 *pixels*.

De acordo com Trigo (2012), a profundidade da área de captura, com qualidade de imagem, é determinada pelos conceitos de distância hiperfocal e profundidade de campo. A distância hiperfocal é a distância mínima entre a lente e o objeto focalizado, para que se obtenha uma boa focalização do objeto. A distância hiperfocal pode ser calculada através da equação 2.3.

$$H = \frac{f^2}{A \times c} \tag{2.3}$$

onde: *H* – Distância Hiperfocal (mm); *A* – Abertura da lente; *c* – Círculo de confusão (mm).

A região de focalização nítida no campo dos objetos é denominada profundidade de campo e deve-se ao limitado poder de resolução do olho. Existe, portanto, uma região que se estende à frente e atrás do ponto focalizado em que as imagens são praticamente nítidas. As posições dos pontos próximo e afastado podem ser determinadas a partir das equações 2.4 e 2.5 da óptica geométrica:

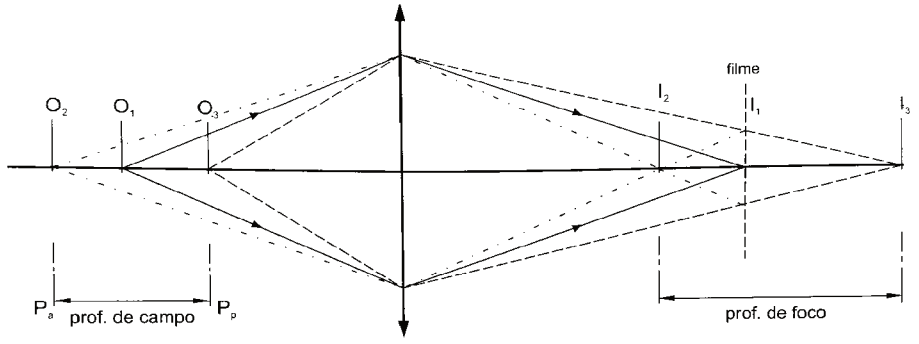
$$P_p = \frac{H \times d}{H + d} \tag{2.4}$$

$$P_a = \frac{H \times d}{H - d} \tag{2.5}$$

onde: d – Distância do objeto focalizado; H – Distância Hiperfocal; P_p – Distância mínima de foco entre o objeto e a câmera; P_a – Distância máxima de foco entre o objeto e a câmera

A Figura 2.12 ilustra as distâncias das equações (2.4) e (2.5). O Eixo Y representa a lente. Os objetos estão posicionados no eixo *x,* à esquerda, e o sensor digital está à direita, na posição I_1. Se o foco está no objeto O_1, a imagem é projetada em I_1, na mesma posição em *x* do sensor digital, portanto, o foco é perfeito. No entanto, existe uma região denominada profundidade de campo, onde a focalização é nítida, situada entre os objetos O_3 e O_2, que estão projetados respectivamente em I_3 e I_2.

Figura 2.12: Profundidade de campo e distância hiperfocal



Fonte: (TRIGO, 2012)

Pode-se observar nas equações (2.4) e (2.5) que, quando o objeto está focalizado no infinito, isto é, d = ∞ , P_p = H, ou seja, a profundidade de campo vai de H até o infinito (TRIGO, 2012). Câmeras de segurança sem controle de foco seguem esse padrão.

A abertura das lentes da câmera e o tamanho do sensor definem os ângulos de visão horizontais e verticais da câmera. O ângulo de visão horizontal determina a largura de captura do ambiente, enquanto que o ângulo de visão vertical determina a altura de captura.

Em geral, sistemas de reconhecimento que atingem taxas altas de reconhecimento são aqueles em que há um controle de iluminação. De acordo com Bezzerra (2012), o valor mínimo aceitável é de 110 lux, o máximo de 1080 lux e o valor ótimo de 500 lux, utilizando lâmpadas de mercúrio metálico branca. Os testes foram realizados utilizando a técnica *PCA*, na qual a variação de luminosidade tem grande influência na qualidade de reconhecimento. Outras técnicas, como a RNCP, podem trabalhar melhor com a variação de luminosidade.

2.5.1.3 Padrão de compressão de vídeo

O padrão de compressão de vídeo H.264 é baseado no MPEG-4 *Part* 10 ou AVC (*Advanced Video Coding*). A versão H.264H gera imagens com uma melhor qualidade do que o padrão H264, porém exige uma capacidade maior de armazenamento. O padrão H.264 é mantido pela *International Organization for Standarization* através da ISO/IEC 14496-10 (ISO, 2014).

A taxa de *bit*, ou *bitrate*, representa a quantidade de *bits* necessária para

codificar um segundo de vídeo, e esta deve ser calculada para o sistema de reconhecimento utilizado. Quanto menor a taxa de *bits*, maior será a distorção (TADEU; SEARA, 2007). A taxa de *bits* utilizada depende da resolução do vídeo e da quantidade de quadros (*frames)* por segundo (FPS).

2.5.1.4 Definições de proteção contra ataques ao equipamento

Os equipamentos expostos são as câmeras, os cabos UTP e adaptadores RJ45 e P4, que requerem algum tipo de proteção contra poeira e água, como o IP66. Os plugues dos cabos RJ45 e do conector de energia P4 da câmera devem ser protegidos com jaquetas de mesmo nível de proteção, regulamentado pela norma IEC 60529 (IEC, 2018).

2.5.1.5 Rede *Ethernet* local

O padrão de rede *Ethernet* 802.3 100BASE-TX é mantido pelo comitê de padrões IEE 802 LAN/MAN. O padrão 100BASE-T estende o MAC IEEE 802.3 a 100 Mb/s. A taxa de *bits* é mais rápida, os tempos de transmissão de pacotes são reduzidos e os atrasos de transmissão no cabo são menores – tudo em proporção à mudança na largura de banda (IEEE, 2018). O padrão 100BASE-TX é o padrão mais utilizado na rede *Fast Ethernet*, utilizando dois pares do cabo de par trançado categoria 5 ou 5e (cabo CAT5 contém 4 pares, sendo usados apenas 2 para transmissão e recepção de dados (IEEE, 2012).

Os cabos de categoria 5, incluindo CAT 5e (categoria 5e; o “e” vem de *enhanced*), foram definidos pela ANSI/TIA/EIA-568-A, sendo o requisito mínimo para redes 100BASE-TX, seguindo padrões de fabricação muito mais estritos, com suporte a frequências de até 100 MHz (TIA, 2018). Atualmente é a categoria mínima de cabo reconhecida para transmissão de dados em ANSI / TIA-568-C (OLIVIERO; BILL, 2014).

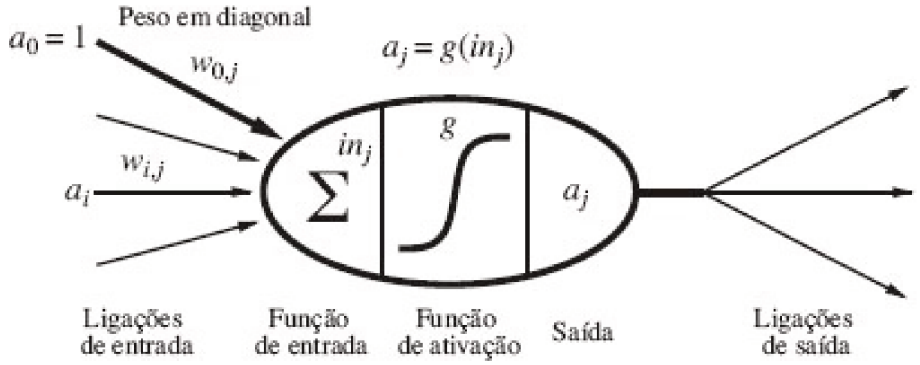
3. REDES NEURAIS CONVOLUCIONAIS PROFUNDAS

Este capítulo descreve o funcionamento das Redes Neurais Convolucionais Profundas, iniciando por uma breve introdução às Redes Neurais Artificiais (seção 3.1), que são estruturas fundamentais para o funcionamento das Redes Neurais Convolucionais Profundas (seção 3.2). Em seguida, a seção 3.3 detalha o funcionamento e a arquitetura de um tipo de Rede Neural Convolucional Profunda, denominada MobileNet, utilizada no protótipo deste trabalho.

3.1 REDES NEURAIS ARTIFICIAIS

Inspirado na hipótese de que a atividade mental consiste basicamente na atividade eletroquímica em redes de células cerebrais chamadas neurônios, alguns dos trabalhos mais antigos de IA tiveram o objetivo de criar as RNA. A Figura 3.1 apresenta um modelo matemático simples do neurônio desenvolvido por McCulloch e Pitts em 1943 (RUSSELL; NORVIG, 2013).

Figura 3.1 – Modelo matemático simples de um neurônio.



Fonte: (RUSSELL; NORVIG, 2013).

Quando uma combinação linear de suas entradas excede algum limiar (rígido ou suave), o neurônio é disparado. O disparo do neurônio, isto é, a ativação de saída da unidade, é dada pela equação 3.1:

$$a_j = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right) \tag{3.1}$$

onde: a_i é a ativação de saída da unidade i, $w_{i,j}$ é o peso sobre a ligação da unidade i com essa unidade e g é a função de ativação.

As RNA são compostas por nós ou unidades (ver Figura 3.1) conectadas por ligações direcionadas. Uma ligação da unidade i para a unidade j serve para propagar a ativação a_i de i para j. Cada ligação também tem um peso numérico $w_{i,j}$ associado, que determina a força e o sinal de conexão. Assim como em modelos de regressão linear, cada unidade tem uma entrada fictícia $a_0 = 1$ com peso associado $w_{0,j}$. Cada unidade j primeiro calcula uma soma ponderada de suas entradas, como se observa na equação 3.2:

$$\text{in}_j = \sum_{i=0}^{n} w_{i,j} a_i \tag{3.2}$$

Em seguida, uma função de ativação g é aplicada a essa soma para obter a saída, na equação 3.3:

$$a_j = g(\text{in}_j) = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right) \tag{3.3}$$

As funções de ativação permitem que pequenas mudanças nos pesos e *bias* (viés) causem apenas uma pequena alteração na saída. Esse é o fator crucial que permitirá que uma rede de neurônios artificiais aprenda. A função de ativação g pode conter um limiar rígido, caso em que a unidade é chamada de Perceptron, ou pode conter uma função logística, caso em que a unidade é chamada de perceptron sigmoide, onde uma função de ativação não linear é utilizada (RUSSELL; NORVIG, 2013).

A função de ativação faz a transformação não-linear nos dados de entrada, capacitando o aprendizado e execução de tarefas mais complexas, como traduções de idiomas (Processamento de Linguagem Natural) e classificações de imagens (Visão Computacional). As transformações lineares nunca seriam capazes de executar tais tarefas (DATA SCIENCE ACADEMY, 2018).

Dentre as principais funções de ativação, pode-se citar: a Função de Etapa Binária, que implementa uma lógica se/senão; a Função Linear, que representa uma

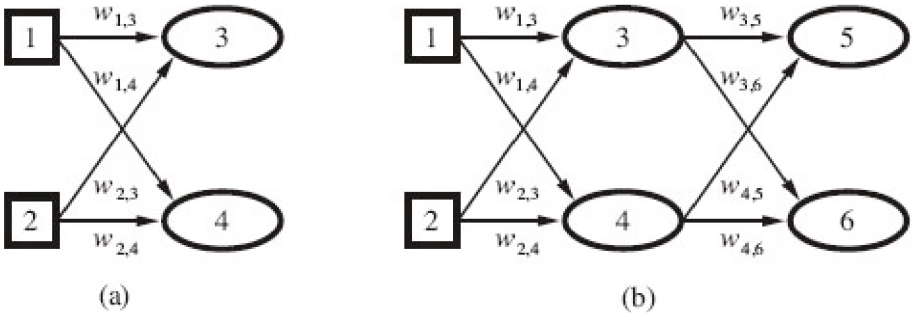
reta; a função Sigmoide, amplamente utilizada, que resulta em um valor de 0 a 1 em forma de “S”; a função Tanh, que é uma versão escalonada da função Sigmoide; a função ReLu, uma das mais utilizadas atualmente, que resulta em 0, caso um valor seja negativo, senão retorna o próprio valor; a função SoftMax, que é um tipo de Sigmoide, mas é utilizada para realizar classificação, uma vez que retorna uma probabilidade, ao dividir o resultado pela soma das saídas, sendo comumente utilizada na última camada de RNA de classificação, como na última camada das RNC (WITTEN; FRANK; HALL, 2011).

Dado o modelo matemático para “neurônios” individuais, a próxima tarefa é conectá-los para formar uma RNA, que pode ser feita com alimentação para frente, que tem conexões somente em uma direção, as chamadas redes *feedfoward*, e outra chamada de rede recorrente, que alimenta suas saídas de volta às suas próprias entradas (RUSSELL; NORVIG, 2013).

Nas RNA com alimentação para frente, normalmente estão dispostas em camadas, de tal forma que cada unidade recebe a entrada somente a partir de unidades na camada imediatamente anterior. As RNA podem conter camadas múltiplas, que têm uma ou mais camadas de unidades ocultas que não são conectadas às saídas da RNA (RUSSELL; NORVIG, 2013).

Uma RNA com todas as entradas conectadas diretamente com as saídas é chamada de rede neural de camada única ou rede P*erceptron*. A Figura 3.2 (a) mostra uma rede perceptron simples de duas entradas e duas saídas, e a Figura 3.2 (b) mostra uma RNA com duas entradas, uma camada oculta de duas unidades e uma unidade de saída, conhecido como rede *Multilayer Perceptron*, ou Perceptron Multicamadas, a qual possui pelo menos uma camada oculta e é composta por nós do tipo perceptron.

Figura 3.2 – Rede *Perceptron* (a) e Rede *MultiLayer Perceptron* (b)



Fonte: (RUSSELL; NORVIG, 2013).

Uma rede perceptron, como a da Figura 3.2 (a), não é capaz de aprender a realizar uma separação não linear. Adicionando mais camadas, a RNA pode realizar mais combinações de saídas que podem realizar uma separação não linear (RUSSELL; NORVIG, 2013).

O treinamento de uma RNA consiste na atualização dos pesos de seus neurônios em função da perda gerada na saída. Os pesos são sugeridos através de um algoritmo que altera os valores dos pesos, através de um parâmetro chamado de taxa de aprendizado, até que o valor da saída seja o suficiente para resolver um problema.

O erro nas camadas ocultas parece misterioso, porque os dados de treinamento não dizem qual o valor os nós ocultos devem ter. Felizmente, verifica-se que podemos retro propagar o erro da camada de saída para as camadas ocultas, em direção a camada de entrada, através do algoritmo *Backpropagation* (Retro propagação), de forma que, os pesos são ajustados até que se atinja um pré-determinado limiar de erro da RNA (RUSSELL; NORVIG, 2013).

Certamente, as RNA são capazes de tarefas muito mais complexas de aprendizagem, embora seja necessária certa quantidade de esforço para obter a estrutura de rede correta e alcançar a convergência para algo próximo ao ótimo global no espaço de pesos.

3.2 REDES NEURAIS CONVOLUCIONAIS

Redes Neurais Convolucionais (RNC) são RNA em que se aplica a operação de convolução em, pelo menos, uma de suas camadas. Esse tipo de rede foi desenvolvido para um conjunto particular de problemas de classificação e de regressão, em que cada amostra de uma base de dados segue uma topologia específica. Nessa topologia, considera-se a existência de uma relação entre valores de índice próximo em uma representação da amostra. Por exemplo, considerando-se uma imagem representada por uma matriz, espera-se que valores de índices próximos sejam altamente relacionados (MAZZA, 2017).

Nos últimos anos, técnicas de Aprendizado Profundo (*Deep Learning*), aplicadas através de Redes Neurais Profundas (RNP), têm revolucionado diversas áreas de aprendizado de máquina, em especial a visão computacional. Isso ocorreu,

principalmente, por dois motivos: a disponibilidade de bases de dados com milhões de imagens e o avanço do poderio computacional (RUSSAKOVSKY et al., 2015).

As RNP são, provavelmente, o modelo de rede *Deep Learning* mais conhecido e utilizado atualmente, com destaque para as Redes Residuais (ResNet) (HE et al., 2016) e Inception (SZEGEDY et al., 2014).

Recentemente, as RNCP têm mostrado seu potencial na DH (LUO et al., 2014; OUYANG; WANG, 2012, 2013; ZENG et al., 2014). Por exemplo, nos trabalhos de Ouyang e Wang, várias combinações de partes do corpo foram representadas por nós em uma rede neural profunda (OUYANG; WANG, 2012).

O RF tem apresentado resultados impressionantes, por exemplo, no trabalho de Schroff e Philbin (2015), um sistema denominado FaceNet foi implementado utilizando uma RNCP capaz de atingir um novo recorde de reconhecimento no banco de dados *Labeled Faces in-the-wild* (HUANG; LEARNED-MILLER, 2014), atingindo uma taxa de reconhecimento de 99,63%.

No banco de dados *YouTube Faces DB* (WOLF; HASSNER; MAOZ, 2011), atingiu-se uma taxa de reconhecimento de 95,12%, diminuindo a taxa de erro em 30% em comparação com as melhores publicações da época. Portanto, a RNCP é um tipo de RNP adaptável à classificação de objetos em imagens, podendo ser utilizada no RF com grande confiabilidade.

Uma grande vantagem das RNCP é a eliminação da necessidade do tratamento de imagens antes da etapa do reconhecimento. As RNCP têm sido utilizadas com maior recorrência desde 2012 e são as grandes responsáveis pelas primeiras colocações na competição de reconhecimento visual de grande escala ImageNet (*ImageNet Large Scale Visual Recognition Challenge* – ILSVRC), como é o caso da RNCP AlexNet, primeira colocada na competição de 2012, realizada anualmente desde 2010 (RUSSAKOVSKY et al., 2015).

A técnica utilizada pelas RNCP se adapta a variações de luminosidade, pose, rotação e expressão facial, o que a torna muito eficiente na utilização de sistemas de segurança em ambientes não controlados, diferentemente da técnica *PCA*, que encontra muitos problemas destes tipos, sendo necessários diversos processamentos de imagem antes da etapa de reconhecimento.

A maior desvantagem da utilização das RNCP é a maior quantidade de imagens para realizar o treinamento da RNA, antes que seja feito o reconhecimento.

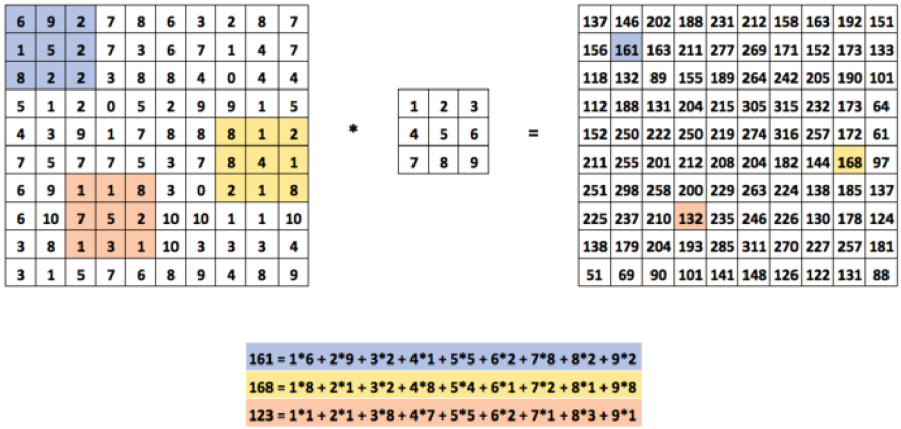
No entanto, um vídeo de poucos segundos é suficiente para capturar diversas imagens de faces de um indivíduo que passa por um ambiente uma única vez.

Nas próximas seções dentro desta, serão abordados alguns dos aspectos mais importantes para se entender o funcionamento de uma RNCP. Em seguida, será abordada a arquitetura da RNCP MobileNet, utilizada no protótipo desse trabalho.

3.2.1 Camada Convolucional

A convolução de matrizes em imagens, envolve a aplicação de um filtro (*kernel*) sobre uma imagem, gerando uma nova imagem como resultado. O *kernel* também é uma matriz, comumente de tamanho 3x3 ou 5x5, que "desliza" sobre toda a imagem, modificando-a. Para um dado deslocamento, os elementos do *kernel* são multiplicados pelos *pixels* da imagem, que, por fim, são somados e resultam em um elemento da imagem, como pode ser observado na Figura 3.3 (BARBOSA; BONADIO, 2018).

Figura 3.3 – Convolução de matrizes



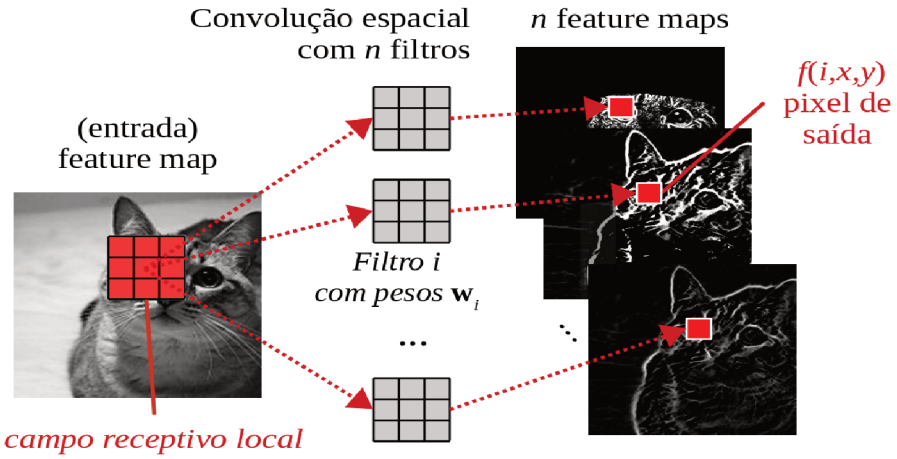
Fonte: (BARBOSA; BONADIO, 2018)

Na camada convolucional, cada neurônio é um filtro aplicado a uma imagem de entrada e cada filtro é uma matriz de pesos. Seja uma imagem RGB de tamanho 224×224×3 (o 3 indica os canais de cor R, G e B), definida como a entrada de uma camada convolucional. Cada neurônio dessa camada servirá como filtro e processará a imagem, realizando uma combinação linear dos *pixels* vizinhos, gerando um peso para cada elemento da imagem (PONTI; DA COSTA, 2018).

A título de exemplo, em uma camada tradicional totalmente conectada (TC), ou *Fully Connected*, teríamos, para cada neurônio do nosso exemplo, 224×224×3 = 150528 pesos, um para cada *pixel* de entrada. Para diminuir a dimensionalidade dessa camada, em vez de 150528 pesos, pode-se definir filtros de tamanho $k \times k \times d$, em que k é a dimensão espacial do filtro (a ser definida) e d a dimensão de profundidade (essa depende da entrada da camada). Por exemplo, se definirmos k = 5 para a primeira camada convolucional, então, têm-se filtros 5×5×3, sendo 3 o número de camadas de cores, diminuindo a dimensão para 5×5×3 = 75 pesos.

Cada região da imagem processada pelo filtro é chamada de campo receptivo local (*local receptive field*). Um valor de saída (*pixel*) é uma combinação dos *pixels* de entrada nesse campo receptivo local (Figura 3.4).

Figura 3.4 – Camada Convolucional



Fonte: (PONTI; DA COSTA, 2018).

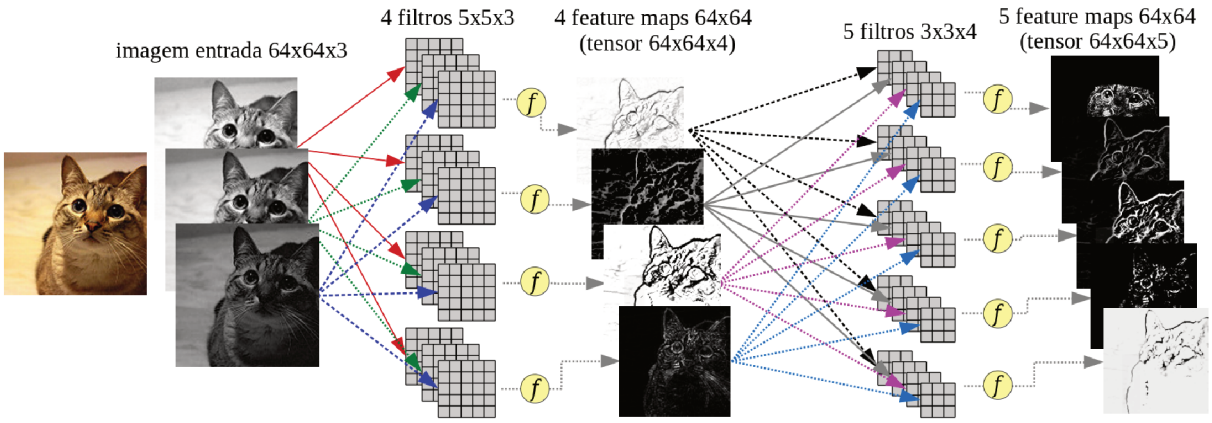
No entanto, todos os campos receptivos são filtrados com os mesmos pesos locais para todo *pixel*, tornando a camada convolucional diferente da camada TC. No exemplo com k = 5, teremos uma combinação linear de 25 *pixels* da vizinhança para gerar um único pixel de saída. Os tamanhos de filtros mais utilizados são 5×5×d, 3×3×d e 1×1×d. O termo tensor é comumente utilizado ao se trabalhar com matrizes multidimensionais (com profundidade d) (PONTI; DA COSTA, 2018). .

Considere uma RNCP de duas camadas convolucionais, onde a entrada é uma imagem RGB, de tamanho 64×64×3. A primeira camada possui 4 filtros de tamanho k_1 = 5, e a segunda 5 filtros de tamanho k_2 = 3. Considere ainda que a convolução é feita utilizando extensão da imagem com preenchimento por zeros

(*zero padding*), filtrando todos os *pixels* da imagem, mantendo seu tamanho. Dessa forma, teríamos a seguinte composição: $\hat{y} = f(x) = f_2(f_1(x_1; W_1; b_1); W_2; b_2))$, em que W_1 possui dimensão 4×5×5×3 (4 filtros de tamanho 5×5, entrada com profundidade 3), e portanto a saída da camada 1, $x_2 = f_1(x_1)$ terá tamanho: 64×64×4 (PONTI; DA COSTA, 2018), como se observa na Figura 3.5.

Denomina-se de *feature* maps, ou mapa de características, a saída de cada neurônio da camada convolucional

Figura 3.5 – Ilustração de duas camadas convolucionais (a primeira produz um tensor com 4 *feature maps*, a segunda um novo tensor com 5 *feature maps*. Os círculos após cada filtro denotam funções de ativação, como por exemplo a ReLU).



Fonte: (PONTI; DA COSTA, 2018).

Os resultados das convoluções de cada filtro g com cada um dos f*eature maps* correspondentes da saída da camada anterior são somados e, em seguida, uma função de ativação é disparada sobre esta soma das matrizes, o que define o *feature map* na saída deste filtro. Com isso, a união de todos os *feature maps* gerados, um para cada filtro g da camada convolucional, gera a saída dessa camada (MAZZA, 2017)..

A função de ativação (comumente a função ReLU) trunca para zero os *pixels* negativos. A Figura 3.5 ilustra esse processo, bem como o da camada 2, que recebe por entrada o tensor 64×64×4. A segunda camada possui 5 filtros 3×3×4 (já que a profundidade do tensor de entrada tem $d = 4$), e gera como saída $x_3 = f_2(x_2)$, um tensor de tamanho 64×64×5. (PONTI; DA COSTA, 2018).

Outro aspecto importante para se mencionar é o passo ou *stride*. A convolução convencional é feita com *stride* 1, ou seja, todos os *pixels* são filtrados e,

portanto, para uma imagem de entrada de tamanho 64×64, uma nova imagem de tamanho 64×64 é gerada. O uso de *strides* maiores que 1 é comum quando se deseja reduzir o tempo de execução, pulando *pixels* e, assim, gerando imagens menores. Ex: com *stride* = 2, tem-se como saída uma imagem de tamanho 32×32 (PONTI; DA COSTA, 2018).

3.2.2 *Feature maps* (mapas de características)

Cada representação gerada por um filtro da camada convolucional é conhecida como "mapa de características", do inglês *feature map* ou *activation map*. Os mapas gerados pelos diversos filtros da camada convolucional são empilhados, formando um tensor cuja profundidade é igual ao número de filtros. Esse tensor será oferecido como entrada para a próxima camada como mostrado na Figura 3.5.

Note que, como a primeira camada convolucional utiliza 4 filtros, gera 4 tensores (*feature maps)* 64×64×4. Os filtros da segunda camada terão que ter profundidade 4, pois os 4 *feature maps*, gerados na camada anterior, são as entradas da próxima camada. Se uma terceira camada convolucional fosse adicionada, 5 *feature maps* 64x64x5 (profundidade 5) seriam utilizados como entrada, logo, os filtros da próxima camada deverão ter profundidade 5 (PONTI; DA COSTA, 2018).

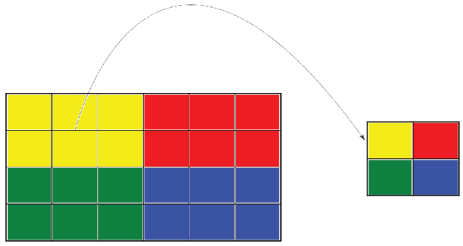
3.2.3 *Pooling*

A dimensão espacial dos mapas ao longo das camadas da RNCP é comumente reduzida através da técnica de *pooling,* sendo a operação de máximo *maxpooling* comumente empregada, no qual o valor do *pixel* de saída será o maior dentre todos os valores dos *pixels* de entrada. Essa operação tem dois propósitos: em primeiro lugar, reduzir o custo computacional, pois como a profundidade dos tensores *d* costuma aumentar ao longo das camadas, é conveniente reduzir a dimensão espacial dos mesmos. Em segundo, reduzindo o tamanho das imagens, obtém-se um tipo de composição de banco de filtros multirresolução que processa imagens em diferentes espaços e escala (PONTI; DA COSTA, 2018).

Um *pooling* usualmente não altera o número de *feature maps*, em vez disso, reduz-se o número de linhas e/ou colunas da entrada.

Por exemplo, um *pooling* de dimensões (2,3), em um *feature map* de dimensões 4x6, resulta num *feature map* com dimensões 2x2, reduzindo o número de atributos por um fator de 6. O procedimento é ilustrado na Figura 3.6, a qual, regiões de mesma cor do *feature map* 4x6, à esquerda, gera valores para o *feature map* 2x2, à direita (MAZZA, 2017).

Figura 3.6 – *Pooling* (2,3) em um feature map 4x6



Fonte: (MAZZA, 2017).

3.2.4 Arquiteturas de Redes Neurais Convolucionais e seus parâmetros

RNC tradicionais são uma combinação de blocos de camadas convolucionais (Conv) seguidas por funções de ativação (FA), eventualmente utilizando também *pooling* e, em seguida, uma série de camadas TC, também acompanhadas por funções de ativação, da seguinte forma:

$RNC \equiv P \times [C \times (Conv \rightarrow FA) \rightarrow Pool] \rightarrow F \times [TC \rightarrow FA].$

Uma RNC pode ser criada através da definição dos seguintes parâmetros: número de camadas convolucionais *C* (para cada camada, o número de filtros, seu tamanho e o tamanho do passo – *stride*), número de camadas de *pooling P* (sendo, nesse caso, necessário escolher também o tamanho da janela e o *stride* que definirão o fator de subamostragem) e o número de camadas totalmente conectadas *F* (e o número de neurônios contidos em cada uma dessas camadas).

O número de parâmetros em uma RNC está relacionado, basicamente, aos valores a serem aprendidos em todos os filtros nas camadas convolucionais, os pesos das camadas totalmente conectadas e os valores dos *bias*.

Como exemplo, considere uma arquitetura para analisar imagens RGB com dimensão 64×64×3, cujo objetivo é classificar essas imagens em 5 classes. Essa arquitetura será composta por três camadas convolucionais, duas *max pooling*, e duas camadas *TC*, da seguinte forma:

a) *Conv.L1*: 10 filtros 5×5×3, saída: tensor de dimensão 64×64×10;
b) *Max pooling* 1: subamostragem com fator 4 (janela de tamanho 2×2 e *stride* 2), saída: tensor de dimensão 16×16×10;
c) *Conv.L2*: 20 filtros 3×3×10, saída: tensor de dimensão 16×16×20;
d) *Conv.L3*: 40 filtros 1×1×20, saída: tensor de dimensão 16×16×40;
e) *Max pooling* 2: subamostragem com fator 4 (janela de tamanho 2×2 e *stride* 2), saída: tensor de dimensão 4×4×40;
f) *TC.L1*: 32 neurônios, saída: 32 valores;
g) *TC.L2* (saída da RNCP): 5 neurônios (um por classe), saída: 5 valores.

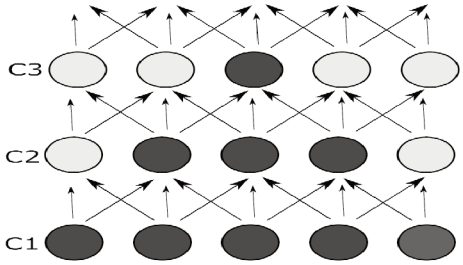
Diferentemente de uma rede neural tradicional, a operação de convolução age somente sobre uma região definida pelas dimensões da máscara de convolução. No caso de imagens, por exemplo, um filtro *g* com número de linhas e colunas iguais a 5, como no caso da *Conv.L1*, causa uma dependência do valor central somente nos *pixels* adjacentes do próprio *feature map* e em todas as regiões 5x5 equivalentes dos outros *feature maps*. Esse aspecto torna a convolução ideal para tratamento de sinais que podem apresentar grande dimensionalidade, como sinais de áudio e imagens (MAZZA, 2017).

Observa-se ainda a partir do exemplo que, a camada convolucional faz uso de um mesmo filtro *g* que iterará sobre toda a imagem para gerar o *feature map* na saída. A utilização dos mesmos parâmetros, para geração de mais de um elemento na saída, é denominada de Compartilhamento de Parâmetros. Ao se reutilizar o conjunto de parâmetros, uma menor quantidade de memória é utilizada pela camada convolucional, além de fazer com que a translação na imagem, de um determinado valor, também translade a saída da camada convolucional (MAZZA, 2017).

A convolução, como parte central do modelo matemático, permite o processamento de sinais de grande dimensionalidade, a partir de Compartilhamento de Parâmetros e Campos Receptivos, o que diminui consideravelmente o número de parâmetros do modelo. Isso ocorre, entretanto, sem que haja perda de generalidade, já que os Campos Receptivos de elementos em camadas mais profundas aumentam a cada camada, como pode ser visto na Figura 3.7.

A arquitetura fundamental de uma rede convolucional alterna camadas de *pooling* com camadas convolucionais para generalizar a classificação. A última camada da RNCP utiliza uma rede neural tradicional para realizar a classificação das

Figura 3.7: Campos Receptivos em três camadas.



Fonte: (MAZZA, 2017).

imagens, por exemplo, identificar qual pessoa, dentre 10 pessoas diferentes, pertence uma face, dentre 1000 faces utilizadas na camada de entrada da RNCP. Nesse caso, o número de neurônios da RNCP na última camada é o mesmo número de pessoas distintas, ou seja, 10 neurônios.

3.3 REDES NEURAIS CONVOLUCIONAIS PROFUNDAS MOBILENET

Desde que a rede neural AlexNet venceu a competição ImageNet em 2012 (RUSSAKOVSKY et al., 2015), as RNCP têm se tornado onipresentes nas competições de classificação de imagens de larga escala. No entanto, nem todas as RNCP se preocupam com a eficiência e velocidade da rede, geralmente demonstrando interesse em uma melhor eficácia, construindo redes complexas para atingir esse objetivo.

Em diversas aplicações, nas quais a velocidade é importante, tais como robótica, carros autodirigíeis, e realidade aumentada, existe a necessidade de se criar RNA mais rápidas, que possam ser utilizadas em tempo real e/ou em uma plataforma com poder de processamento computacional limitado.

Para suprir a necessidade de implementação de sistemas de tempo real, que necessitam utilizar RNC pequenas e rápidas, as redes MobileNet (HOWARD et al., 2017a) foram desenvolvidas com a promessa de serem adaptáveis as diversas restrições de tamanho e latência, encontradas em diferentes aplicações de diferentes plataformas, incluindo aplicações de dispositivos móveis.

3.3.1 Arquitetura MobileNet

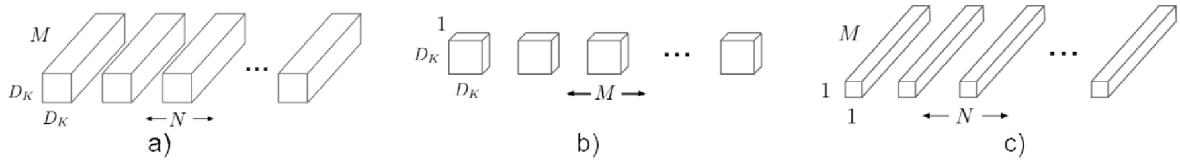
A arquitetura da RNCP MobileNet é baseada em Convoluções separáveis em

Profundidade (*Depth Wise – DW*), que é uma forma de convoluções que fatorizam uma convolução padrão em uma convolução em profundidade e em uma convolução 1×1, chamada de Convolução Pontual (*Point Wise – PW)*. A *DW* aplica um único filtro a cada canal de entrada. A *PW*, em seguida, aplica uma convolução 1×1 para combinar as saídas da *DW*.

Uma convolução padrão filtra e combina as entradas em um novo conjunto de saídas em uma única etapa. A *DW* divide-a em duas camadas, uma camada separada para filtragem e uma camada separada para combinação. Essa fatoração tem o efeito de reduzir drasticamente o tamanho do cálculo e do modelo.

A Figura 3.8 mostra como uma convolução padrão (a) é fatorada em uma *DW* (b) e em uma *PW* 1×1 em (c), onde *K* é o núcleo de convolução, D_K é a dimensão espacial do núcleo que se supõe ser quadrada, *M* é o número de canais de entrada e *N* é o número de canais de saída (HOWARD et al., 2017a).

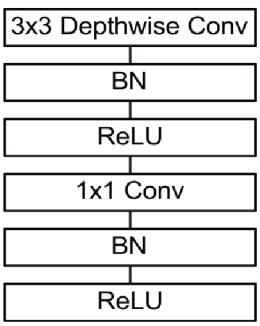
Figura 3.8: Convoluções Padrão, em Profundidade e Pontuais (da esquerda para a direita: (a) Filtros Convolucionais Padrão; (b) Filtros *DW*; (c) Filtros *PW* no contexto de *DW*).



Fonte: (HOWARD et al., 2017a).

A estrutura MobileNet é construída em *DW*, com exceção da primeira camada que é uma convolução completa. Através da definição da rede nestes simples termos, a topologia de rede pode ser facilmente explorada para se encontrar uma boa arquitetura. A arquitetura da camada MobileNet é ilustrada na Figura 3.9.

Figura 3.9: Camadas MobileNet (com *DW* e *PW* seguidas de *batchnorm* e ReLu).



Fonte: (HOWARD et al., 2017a).

Todas as camadas são seguidas de uma Normalização em *Batch* e de uma função de ativação ReLu não linearizada, com exceção da última camada totalmente conectada, que pode ser linear e que alimenta uma camada *softmax* para classificação. A Normalização em *Batch*, ou *batchnorm* (IOFFE; SZEGEDY, 2015), como alternativa, normaliza o vetor na saída da camada.

Um exemplo de arquitetura MobileNet pode ser vista na tabela Arquitetura MobileNet, no anexo A. Como se pode observar, a última camada da RNCP MobileNet é capaz de classificar 1000 objetos distintos. Para o RF isso implica que, no máximo 1000 indivíduos distintos podem ser reconhecidos utilizando essa arquitetura.

3.3.2 Multiplicador de Largura MobileNet: largura reduzida

Embora a arquitetura básica MobileNet seja pequena e com baixa latência, pode ser necessário diminuir o tamanho e aumentar a velocidade do modelo, dependendo da aplicação. Para atingir esse objetivo, o parâmetro chamado Multiplicador de Largura foi introduzido, que reduz uniformemente a largura em cada camada. Para cada camada e Multiplicador de Largura α, o número de canais de entrada M torna-se $\alpha \times M$, e o número de canais de saída N torna-se $\alpha \times N$.

Uma camada convolucional padrão toma como entrada um mapa de características $D_F \times D_F \times M$ e produz um mapa de características $D_F \times D_F \times N$, onde D_F é a largura e altura espaciais de um mapa de características de entrada quadrado, M é o número de canais de entrada (profundidade de entrada) e N é o número de canais de saída (profundidade de saída).

A camada convolucional padrão é parametrizada pelo núcleo de convolução K de tamanho $D_K \times D_K \times M \times N$, em que D_K é a dimensão espacial do núcleo que se supõe ser quadrada, M é o número de canais de entrada e N é o número de canais de saída, conforme definido anteriormente. Convoluções padrão têm o custo computacional: $D_K \times D_K \times M \times N \times D_F \times D_F$.

O custo computacional de uma Convolução separável em Profundidade com o Multiplicador de Largura α é: $D_K \times D_K \times \alpha M \times D_F \times D_F + \alpha M \times \alpha N \times D_F \times D_F$, que representa a soma da Convolução em Profundidade com a Convolução Pontual, onde $\alpha \in (0, 1)$ com configurações típicas de 1, 0.75, 0.5 e 0.25, onde $\alpha = 1$ é a linha de base MobileNet e $\alpha < 1$ são redes reduzidas.

O Multiplicador de Largura tem o efeito de reduzir o custo computacional e reduzir o número de parâmetros de forma quadrática por aproximadamente α^2. A precisão da rede diminui suavemente até que a arquitetura seja bastante diminuída até $\alpha = 0.25$, como se verifica na Tabela 3.1 (HOWARD et al., 2017b).

O número de operações de multiplicação e adição fundidas são representadas pela sigla *MAC (Multiply-Accumulates)*. A quantidade de *MACs* e de parâmetros estão apresentados em milhões de unidades, na Tabela 3.1.

Tabela 3.1 – Resultados MobileNet na ILSVRC.

Configuração	*MACs*	Parâmetros	Precisão Top-1	Precisão Top-5
MobileNet_v1_1.0_224	569	4.24	70.7	89.5
MobileNet_v1_1.0_192	418	4.24	69.3	88.9
MobileNet_v1_1.0_160	291	4.24	67.2	87.5
MobileNet_v1_1.0_128	186	4.24	64.1	85.3
MobileNet_v1_0.75_224	317	2.59	68.4	88.2
MobileNet_v1_0.75_192	233	2.59	67.4	87.3
MobileNet_v1_0.75_160	162	2.59	65.2	86.1
MobileNet_v1_0.75_128	104	2.59	61.8	83.6
MobileNet_v1_0.50_224	150	1.34	64.0	85.4
MobileNet_v1_0.50_192	110	1.34	62.1	84.0
MobileNet_v1_0.50_160	77	1.34	59.9	82.5
MobileNet_v1_0.50_128	49	1.34	56.2	79.6
MobileNet_v1_0.25_224	41	0.47	50.6	75.0
MobileNet_v1_0.25_192	34	0.47	49.0	73.6
MobileNet_v1_0.25_160	21	0.47	46.0	70.7
MobileNet_v1_0.25_128	14	0.47	41.3	66.2

Fonte: (HOWARD et al., 2017b).

3.3.3 Multiplicador de Resolução MobileNet: representação reduzida

O segundo parâmetro para reduzir o custo computacional da RNCP MobileNet é o Multiplicador de Resolução ρ. Quando aplicado à imagem de entrada, a representação interna de cada camada é subsequentemente reduzida pelo mesmo multiplicador. Na prática, o parâmetro é implicitamente utilizado através da alteração da resolução de entrada.

O custo computacional para as camadas principais da RNCP MobileNet, do tipo Convoluções *DW*, com Multiplicador de Largura α e Multiplicador de Resolução ρ é $D_K \times D_K \times \alpha M \times \rho D_F \times \rho D_F + \alpha M \times \alpha N \times \rho D_F \times \rho D_F$, onde $\rho \in (0, 1)$ é tipicamente configurado implicitamente, de modo que a resolução de entrada da rede seja 224,

192, 160 ou 128, $\rho = 1$ é a linha de base MobileNet e $\rho < 1$ são redes com custo computacional reduzido. Multiplicadores de Resolução reduzem o custo computacional em ρ^2, e a precisão reduz sutilmente, como pode ser observado na Tabela 3.1 (HOWARD et al., 2017b).

3.3.4 Resultados MobileNet

Existem 16 modelos MobileNet pré-treinados, utilizados na competição ILSVRC, nas tarefas de classificação de objetos, sendo possível escolher um dos modelos da MobileNet para ajustar o orçamento de latência e tamanho, como pode ser observado na Tabela 3.1. O tamanho da RNCP MobileNet na memória e no disco é proporcional ao número de parâmetros. A latência e o uso de energia da RNCP MobileNet são dimensionados com o número de *MACs*.

As precisões Top-1 e Top-5 foram medidas no conjunto de dados ILSVRC, utilizadas para os resultados da Tabela 3.1 (HOWARD et al., 2017b).

A Tabela 3.2 compara a arquitetura MobileNet completa com a arquitetura original GoogleNet (SZEGEDY et al., 2014) e VGG16 (SIMONYAN; ANDREW ZISSERMAN; ZISSERMAN, 2015). A MobileNet é quase tão precisa quanto a VGG16, enquanto que, é 32 vezes menor e têm custo computacional 27 vezes menor; é mais precisa que a arquitetura GoogleNet, menor e têm custo computacional 2,5 vezes menor.

Tabela 3.2 – Comparativo MobileNet com modelos populares.

Modelo	Precisão ILSVRC	*MACs*	Parâmetros
1.0 MobileNet-224	70,6%	569	4,2
GoogleNet	69,8%	1550	6,8
VGG 16	71,5%	15300	138

Fonte: (HOWARD et al., 2017a).

A Tabela 3.3 compara uma arquitetura MobileNet, reduzida com um Multiplicador de Largura $\alpha = 0,5$ e Multiplicador de Resolução 160, com outros modelos populares. A arquitetura reduzida é 3% melhor que a AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), sendo 45 vezes menor e consumindo 9,4 vezes menos tempo de processamento. Além disso, é 2,7% melhor

que a rede Squeeezenet (IANDOLA et al., 2016), sendo mais ou menos do mesmo tamanho e 22 vezes mais rápida.

Tabela 3.3 – Comparativo de uma arquitetura MobileNet menor com outros modelos populares.

Modelo	Precisão ILSVRC	*MACs*	Parâmetros
0.5 MobileNet-160	60,2%	76	1,32
Squeezenet	57,5%	1700	1,25
AlexNet	57,2%	720	60

Fonte: (HOWARD et al., 2017a).

Nas comparações realizadas com faces, destaca-se o modelo FaceNet (SCHROFF; KALENICHENKO; PHILBIN, 2015), que é um modelo de RF de última geração. Ele constrói partes faciais baseadas em *triplet loss*, que é uma técnica que utiliza uma imagem âncora para comparação, uma imagem positiva, que é da mesma pessoa da imagem âncora, e uma imagem negativa, que é de outra pessoa, para realizar o treinamento da RNCP MobileNet. Os resultados de modelos MobileNet muito reduzidas podem ser encontrados na Tabela 3.4.

Tabela 3.4 – Resultados MobileNet e FaceNet.

Modelo	Precisão 1e-4	*MACs*	Parâmetros
FaceNet	83%	1600	7,5
1.0 MobileNet-160	79,4%	286	4,9
1.0 MobileNet-128	78,3%	185	5,5
0,75 MobileNet-128	75,2%	166	3,4
0,75 MobileNet-128	72,5%	108	3,8

Fonte: (HOWARD et al., 2017a).

Pode-se observar que, a RNCP MobileNet é uma rede adaptável às mais diferentes necessidades, sendo rápida e enxuta, além de ser comparável ao estado da arte no RF, podendo ser aplicada em sistemas de RF rápidos e de baixo custo.

4. PROTÓTIPO PROPOSTO

Este capítulo apresenta o protótipo construído e os resultados obtidos com sua utilização. Na seção 4.1, é exibida uma visão geral do protótipo. Em seguida, são apresentadas as proposições da arquitetura de *hardware* (seção 4.2), das definições de parâmetros de ambientação (seção 4.3), da arquitetura de *software* de RF (seção 4.4) e a implementação das arquiteturas propostas no protótipo (seção 4.5). Na seção 4.6, são exibidos resultados da implementação das arquiteturas no protótipo e resultados da utilização do protótipo em um experimento, realizado em um ambiente compatível com as definições de ambientação. Finalmente, é realizado um comparativo entre o custo do protótipo e dois sistemas de RF, encontrados atualmente no mercado nacional.

4.1 VISÃO GERAL DO PROTÓTIPO

O protótipo do sistema de RF para controle de acesso é composto por uma arquitetura de *software* e uma arquitetura de *hardware* de baixo custo, de modo que a arquitetura de *software* é projetada para trabalhar sob a arquitetura de *hardware*.

A arquitetura de *hardware* é composta por duas câmeras do tipo IP, que transmitem dados via rede *Ethernet*, conectadas a um *switch*, que é conectado a um computador, onde as imagens da câmera são recebidas e processadas pelo sistema de RF, configurando assim um CFT.

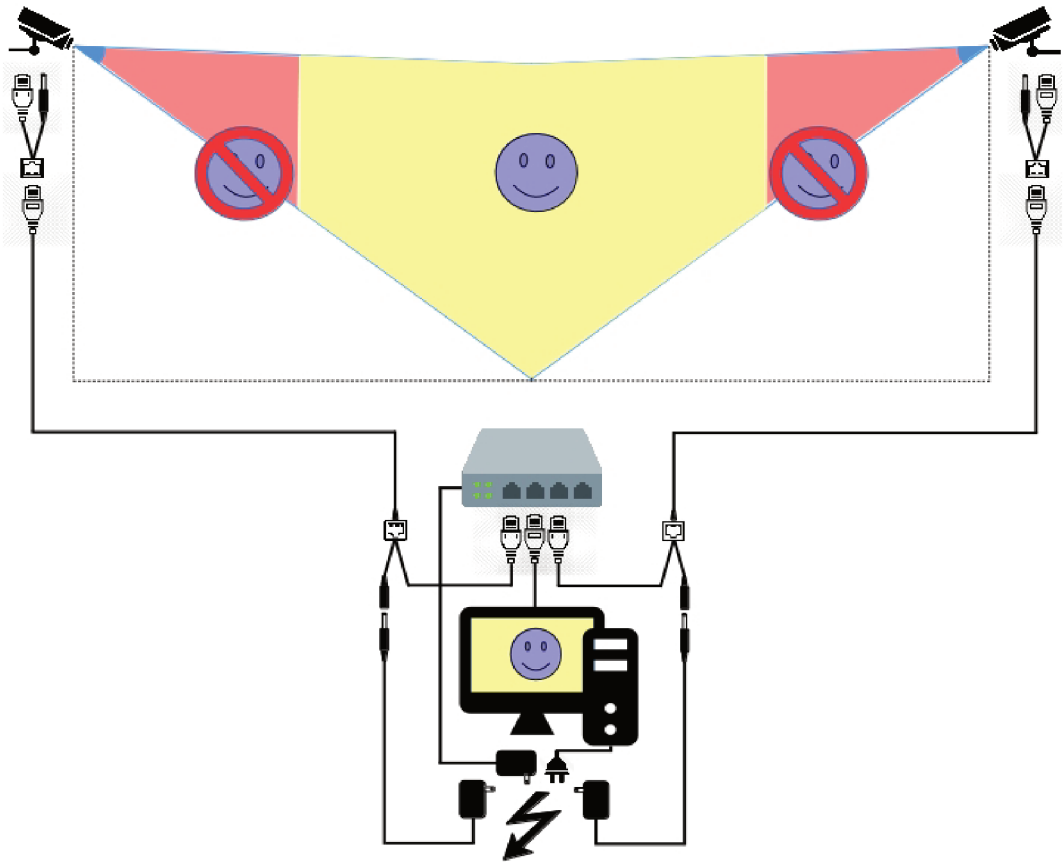
As duas câmeras utilizadas são da mesma marca e modelo, configuradas de forma idêntica, para que o sistema de RF também atue de forma idêntica nas imagens das duas câmeras. A Figura 4.1 exibe o esquemático do modelo.

O protótipo do sistema realiza a captura das imagens de vídeo das câmeras, posicionadas na entrada e saída do ambiente. Em seguida, as faces são recortadas das imagens e são armazenadas em um banco de dados. As pessoas são identificadas em um processo manual de registro de pessoas, ao final da captura. Ao passo que mais pessoas são registradas, um treinamento da RNA é executado para que o sistema seja capaz de identificar as próximas faces capturadas.

Ao passo que o sistema realiza a captura das faces, a base de dados de pessoas registradas é atualizada, quando as faces são identificadas. Por outro lado, se as faces não são identificadas, ou são identificadas com uma baixa

confiabilidade, é criado um registro de pessoa não identificada, o qual deve ser utilizado posteriormente para a identificação manual.

Figura 4.1 – Esquemático geral da arquitetura



Fonte: Autor

Após a identificação manual, o sistema deve ser treinado novamente, para que seja possível reconhecer as faces destas pessoas em uma nova passagem pelo ambiente.

4.2 MODELO DE ARQUITETURA DE *HARDWARE* DE BAIXO CUSTO

As características das lentes das câmeras, tais como distância focal, abertura e ângulo de visão, foram determinadas de acordo com o ambiente idealizado pelo protótipo no item 4.4. A resolução da câmera, escolhida para o modelo, compatível com o tamanho do ambiente proposto, foi fixada a 720p, pois o aumento da resolução da câmera implica um aumento considerável no custo da

câmera e do computador, dado que o custo do processamento de operações de detecção, reconhecimento e treinamento da RNA é linearmente proporcional à área da imagem. Nos testes realizados com a câmera do protótipo, foi fixada a taxa de 14 FPS, uma vez que foi verificada uma taxa média de detecção de 14 FPS.

O padrão de compressão de vídeo escolhido, H.264, e a taxa máxima de 2048kbps, foram definidos para capturar o vídeo sem perda de qualidade, considerada a resolução de 720p e a taxa de 14 FPS, permitindo a utilização da arquitetura com duas câmeras sem sobrecarregar o tráfego da rede *Ethernet* local. De fato, 2048Kbps multiplicados por duas câmeras, resulta em 4096Kbps, ou 4Mbps, sendo que o padrão de rede *Ethernet* utilizado, 802.3 100BASE-TX, suporta até 100Mbps.

A velocidade mínima de detecção e reconhecimento de uma face, nos experimentos realizados, utilizando um Multiplicador de Resolução da RNCP MobileNet 128 e Multiplicador de Largura 1.0, foi de 3 FPS e a distância máxima para captura de 4,88 m (definido no item 4.3). Um indivíduo caminhando em média a 4 Km/h, ou 1,1 m/s, leva 5,41 segundos para percorrer essa distância, o que representaria até 3 FPS x 5,41 segundos, ou aproximadamente 16 *frames* com RF em série durante a passagem.

Considerando-se somente a DF, a qual se atingiu 14 FPS, é possível capturar 14 frames x 5,41 segundos, ou aproximadamente 75 faces por passagem, de forma que o RF é feito após ou paralelamente à detecção. Para atingir esses resultados, foi utilizado um computador com processador i7, de sétima geração, com 4 processadores de *clock* 2.70 GHz. Nos testes do protótipo, a memória RAM não excedeu 4 GB em nenhum momento, utilizando o sistema operacional Ubuntu Linux 16.04.

Os equipamentos expostos são as câmeras, os cabos UTP e adaptadores RJ45 e P4, que requerem proteção contra poeira e água. Para a câmera, como foi citado no item 2.5.1.1, dado que o ambiente é interno e pode ter janelas, foi determinada proteção IP66. Os plugues dos cabos RJ45 e do conector de energia P4 da câmera são protegidos com jaquetas de mesmo nível de proteção IP66, regulamentado pela norma IEC 60529.

O *switch* utilizado deve conter pelo menos três entradas para conectar as duas câmeras ao computador. Com o objetivo de reduzir os custos das câmeras e do *switch*, não foi escolhida a alimentação *PoE* (*Power Over Internet*) para alimentar

esses equipamentos, o que tornaria a solução muito mais cara. Por esse motivo, a alimentação é feita através dos cabos e adaptadores.

Um cabo UTP é conectado entre a câmera e o *switch*, sendo um cabo para cada câmera. A alimentação de cada câmera é feita através do respectivo cabo de rede. Na ponta do cabo UTP que chega até a câmera, é utilizado um adaptador que separa a alimentação dos dados, chamado de *splitter*. Na outra ponta do cabo UTP que vai para o *switch* é conectado um adaptador para o recebimento da energia, chamado de *injector*, como pode ser visto na Figura 4.1.

O tipo de cabo UTP definido para a arquitetura é o cabo do tipo CAT 5e, adequado para as taxas de *bit* e compatibilidade com o padrão de rede *Ethernet* 802.3 100BASE-TX.

A fonte de alimentação utilizada deve ser compatível com a câmera, geralmente 12 V.

Os resultados resumidos de todas as características de *hardware* podem ser observados nas tabelas do apêndice A.

4.3 PARÂMETROS DE AMBIENTAÇÃO

A região de atuação de captura de imagens deve ser compatível com a resolução da câmera escolhida, 1280x720 *pixels*, uma vez que o tamanho da imagem da face capturada deve ser de tamanho suficiente para alcançar uma taxa de acerto satisfatória. Para determinar a região de captura, foram geradas equações como resultados que auxiliam no dimensionamento da área.

A partir da constante de largura média de 160 mm de uma face e 40 *pixels* por face, como sendo a quantidade mínima de *pixels* que uma face deve ser representada em uma imagem (BEZERRA, 2012), aqui se define a resolução facial na equação 4.1.

$$RES_{face} = \frac{Px_{face}}{L_{face}} \tag{4.1}$$

onde: RES_{face} – Resolução da face (*pixels*/mm); Px_{face} – Constante de *pixels* por face (*pixels*); L_{face} – Constante de largura da face (mm). Substituindo pelos valores do protótipo: RES_{face} = 40 / 160 = 0,25 *pixels* por milímetro.

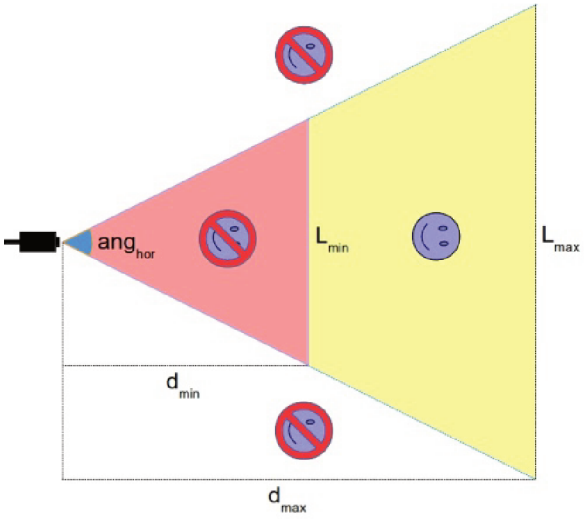
Utilizando a resolução da face e a resolução horizontal da câmera, é possível determinar a largura da imagem que contém a face, na distância máxima, utilizando a equação 4.2.

$$L_{max} = \frac{RES_{camera}}{RES_{face}} \tag{4.2}$$

onde: L_{max} – Largura máxima do ambiente na imagem que contém a face (mm); RES_{camera} – Resolução horizontal da câmera (*pixels*); RES_{face} – Resolução da face (*pixels*/mm). Aplicando o valor da resolução horizontal do protótipo: L_{max} = 1280 / 0,25 = 5120 mm.

Observa-se na Figura 4.2 que, dado o ângulo de abertura horizontal da câmera e a largura do ambiente, é possível descobrir a distância entre a lente e o objeto, através da equação 4.3.

Figura 4.2: Visão superior da cena. Distâncias (d) e Larguras (L) mínimas e máximas.



Fonte: Autor

$$d = \frac{L_{imagem}/2}{\tan\left(ang_{hor}/2\right)} \tag{4.3}$$

onde: d – Distância entre a face e a lente (mm); L_{imagem} – Largura do ambiente da imagem que contém a face (mm). ang_{hor} – Ângulo horizontal de abertura da lente (graus). Aplicando a equação 4.3, fazendo L_{imagem} = L_{max} e

utilizando o valor do ângulo horizontal do campo de visão da câmera do protótipo, obtém-se a distância máxima d_{max} entre a lente e a face: d_{max} = 5120/2 / tan(56,82/2) = 4732,64 mm.

Aplicando-se a equação 4.2 em 4.3, sendo $L_{imagem} = L_{max}$, pode se obter a equação resumida 4.4 para determinar a distância máxima entre a câmera e a face capturada:

$$d_{max} = \frac{RES_{camera} \times L_{face}}{Px_{face} \times \tan\left(ang_{hor}/2\right) \times 2} \quad (4.4)$$

onde: d_{max} – Distância máxima entre a face e a lente (mm). Aplicando as variáveis do protótipo na equação 4.4, obtém-se o mesmo valor encontrado anteriormente de d_{max}, 4732,64 mm, valor que determina a distância máxima que uma pessoa deve estar da lente para que se tenha uma face com pelo menos 40 *pixels*.

Dada a configuração do protótipo, a pessoa deve estar entre a distância mínima, que é a distância hiperfocal, calculada para a câmera do protótipo, d_{min} = 1148,03 mm e d_{max} = 4732,64 mm de distância da lente, para que se obtenha uma imagem de boa qualidade para o RF.

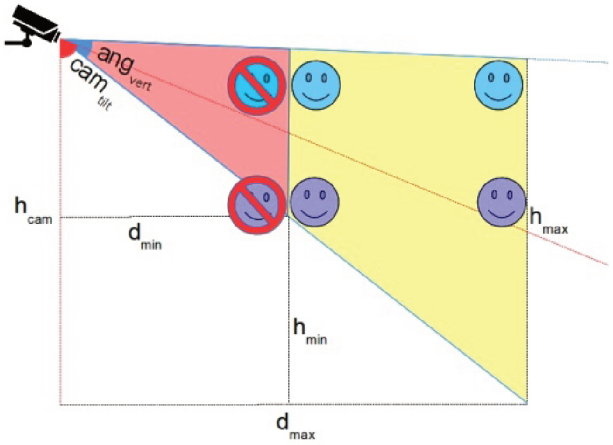
A largura mínima do ambiente é determinada pela distância mínima que um objeto pode estar posicionado, que é a distância hiperfocal. Logo, aplicando equação 4.3, a largura mínima do ambiente, utilizando $d_{min} = H$ = Distância Hiperfocal = 1148,03 mm: L_{min} = 2 x d_{min} x TAN(ang_{hor}/2) = 1241,99 m, e a largura máxima do ambiente, fazendo L_{max} = 2 x d_{max} x TAN(ang_{hor}/2) = 5,12 m, mesmo valor obtido através da equação 4.2.

A lente da câmera está posicionada a uma altura h_{cam} do chão, e o centro de foco de sua lente forma um ângulo cam_{tilt} com o chão. O ângulo de visão vertical do protótipo (ang_{vert}) é de 35,07°. A altura da câmera (h_{cam}) pode ser determinada como sendo a altura máxima das pessoas. No protótipo foi utilizado o valor 2,25 m.

O ângulo cam_{tilt} deve ser determinado de forma que a cena capture uma pessoa alta, de altura h_{max} à distância $d_{max.}$, sem prejudicar a captura de pessoas baixas a partir da distância mínima d_{min} considerando as limitações da técnica de Viola e Jones (2001). A Figura 4.3 exibe a visão lateral da cena, onde é possível

visualizar as distâncias e ângulos descritos neste parágrafo.

Figura 4.3: Visão lateral da cena. Distâncias (*d*) e alturas (*h*) mínimas e máximas.



Fonte: Autor

Dado que o centro de visão vertical da câmera está entre o ângulo que determina a altura máxima e o ângulo que determina a altura mínima, e ainda utilizando o triângulo formado entre a altura da câmera h_{cam}, a altura máxima de uma pessoa h_{max} e a distância máxima entre a lente e a face capturada d_{max}, é possível chegar à equação que determina o ângulo cam_{tilt}, em graus, entre o centro de visão vertical da câmera e o chão através da equação 4.5:

$$cam_{tilt} = 90 - \tan^{-1}\left(\frac{h_{cam} - h_{max}}{d_{max}}\right) - \frac{ang_{vert}}{2} \quad (4.5)$$

onde: cam_{til} – Ângulo entre a câmera e o chão; h_{max} – altura de captura; h_{cam} – altura da câmera; ang_{vert} – ângulo vertical da câmera. Aplicando os valores do protótipo na equação 4.5: h_{cam} = 2250; h_{max} = 2250; d_{max} = 4732,64; ang_{vert} = 35,07°, obtém-se cam_{tilt} = 72,46°, ângulo entre o centro de visão vertical da câmera e o chão.

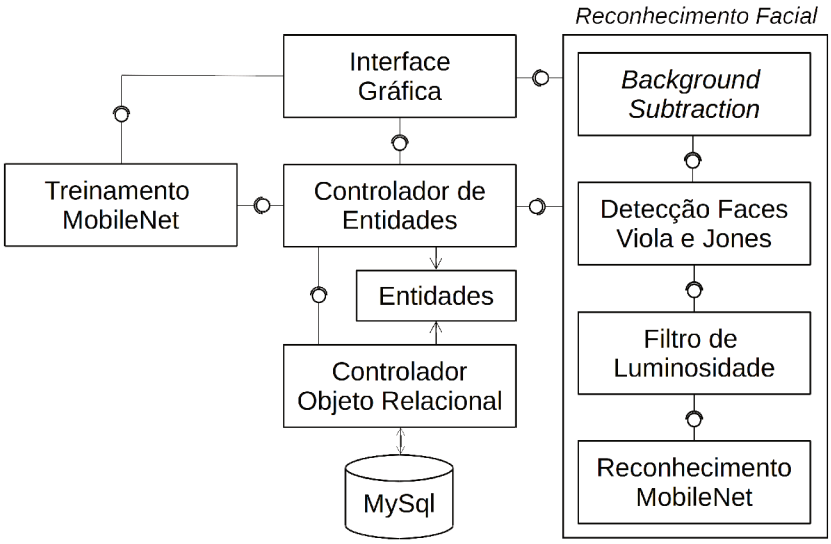
Os valores de configurações do ambiente, utilizando a câmera do protótipo, estão resumidas na tabela do apêndice B.

4.4 MODELO DE ARQUITETURA DE *SOFTWARE* PARA DETECÇÃO DE HUMANOS, DETECÇÃO DE FACES E DE RECONHECIMENTO DE FACES

A arquitetura de *software* utiliza uma abordagem Orientada a Objetos com padrões de projetos conhecidos e a linguagem de programação Python 3.5 para a implementação. O sistema operacional utilizado é o Ubuntu Linux 16.04.

A arquitetura proposta é composta por um componente de Interface Gráfica, que acessa os componentes de Treinamento MobileNet, o componente de Reconhecimento Facial e o componente Controlador de Entidades, como pode ser observado na Figura 4.4.

Figura 4.4 – Diagrama de componentes da arquitetura de *software*



Fonte: Autor

O componente que representa a interface gráfica foi desenvolvido utilizando as ferramentas e bibliotecas de interfaces gráficas Qt 5.10.1.

O componente Controlador de Entidades foi desenvolvido para manipular as entidades do sistema, definidas no componente de Entidades. São utilizados padrões de projeto, como *Controller*, *Adapter* e *Facade,* para reduzir o acoplamento entre os componentes Entidades e os componentes que o manipulam, Treinamento MobileNet, Interface Gráfica e Reconhecimento Facial, fazendo com que as funcionalidades do Controlador de Entidades possam ser reutilizadas em outros sistemas.

O Componente de Entidades foi criado para se armazenar os objetos das classes, tais como Câmeras e suas configurações de subtração de fundo, detecção e reconhecimento, Pessoas, Suspeitos, Registros de Faces, Registros de Passagens, Configuração de Treinamento e Resultados de Treinamento.

O componente Controlador Objeto Relacional utiliza a biblioteca SqlAlchemy versão 1.2.7 para realizar o mapeamento objeto-relacional das Entidades com o banco de dados MySql, versão 14.14. Dessa forma, todas as informações das Entidades são armazenadas no banco de dados, com exceção das imagens e arquivos contendo resultados de treinamentos, que são armazenadas em disco.

Para construir o componente de Treinamento MobileNet, que utiliza os modelos pré-treinados da RNCP MobileNet, foi desenvolvido um adaptador para acessar as interfaces da RNCP MobileNet. Esse componente tem como função classificar as faces e gerar as entidades de Resultados de Treinamento, que são utilizados para o RF. A RNCP MobileNet foi construída utilizando a biblioteca TensorFlow (ABADI et al., 2016), versão 1.6.

O componente de Reconhecimento Facial é composto de 4 camadas, sendo cada camada um componente que se comunica com a próxima camada, através de uma interface que a próxima camada disponibiliza.

Na primeira camada, um componente realiza a DH, definindo a RI, através da técnica *Background Subtraction* e outras técnicas disponibilizadas na biblioteca OpenCV 3.4, descritas na seção 4.4.1. As imagens que contém a RI são armazenadas em uma matriz na memória, que é utilizada como entrada da segunda camada.

A segunda camada utiliza o método de Viola e Jones (2001), disponibilizado na biblioteca OpenCV 3.4, para extrair as faces das imagens da RI, que são armazenadas em matrizes na memória, e utilizadas na terceira camada.

A terceira camada, Filtro de Luminosidade, filtra as matrizes que contém as faces, em memória, de forma que, somente as faces com boa luminosidade são consideradas para se realizar o reconhecimento. Essa funcionalidade está melhor detalhada na seção 4.6.1, e está prevista em futuras implementações do protótipo.

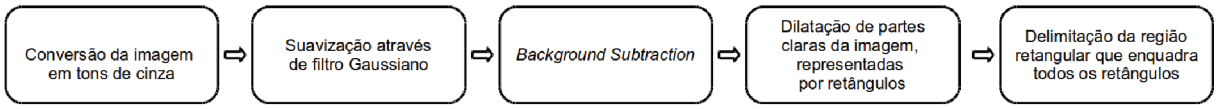
Em seguida, as imagens que contém as faces, armazenadas em matrizes na memória, são, então, armazenadas em disco. As informações das faces selecionadas, como o caminho em disco das imagens e a identificação da pessoa ou suspeito, são armazenas em entidades do tipo Registro de Faces.

Por fim, o componente de Reconhecimento MobileNet utiliza os arquivos que contém as imagens com faces para fazer o reconhecimento e gerar as informações de TRF, que são armazenadas nas entidades Registro de Passagem.

4.4.1 Detecção de humanos

A DH, etapa anterior à captura de faces para o reconhecimento, foi simplificada através da utilização da técnica *Background Subtraction*, por ser uma técnica extremamente rápida que detecta movimento, como pode ser visto na Figura 4.5.

Figura 4.5 – Etapas da detecção de humanos



Fonte: Autor

A técnica *Background Subtraction* utiliza imagens em tons de cinza para realizar a subtração, por isso, a primeira etapa da técnica requer a transformação da imagem em tons de cinza.

Adicionalmente, em seguida, uma filtragem gaussiana é feita para suavizar a imagem, através da convolução de cada ponto na matriz de entrada e, então, todos os pontos são somados para produzir a matriz de saída.

Após, uma subtração da imagem capturada pela imagem de fundo (*Background Subtraction)*, que resulta em uma imagem com os *pixels* mais claros representando as maiores diferenças. A imagem resultante com as diferenças é convertida em preto e branco (“binarizada”), através de um limiar configurável na interface gráfica do protótipo.

Em seguida, uma técnica foi implementada para determinar as maiores áreas subtraídas, através da execução de uma operação de dilatação na imagem “binarizada”. Esta operação consiste na convolução da imagem com alguma matriz (*kernel*). Quando o *kernel* é escaneado sobre a imagem, o valor máximo dos *pixels* sobrepostos pelo *kernel* é calculado e, então, o valor resultante substitui o *pixel* da imagem na posição do ponto de ancoragem com esse valor máximo. Essa operação de maximização faz com que regiões de cor branca “cresçam” dentro de uma

imagem, fazendo com que as diferenças sejam representadas em uma área retangular maior (do que um *pixel*, por exemplo). Dessa forma, diversas regiões retangulares representam as alterações.

Após determinar as regiões retangulares que enquadraram as partes em movimento, um novo enquadramento é feito, na última etapa, de forma que todas as regiões retangulares sejam enquadradas em um único retângulo, que finalmente determina a região de interesse (RI). A RI é utilizada na detecção das faces, desprezando as áreas da imagem que não contém humanos.

4.4.2 Detecção de faces

De acordo com a atual literatura, a técnica apresentada por Viola e Jones (2001), utilizando características Haar, é uma técnica muito utilizada em trabalhos que requerem a DF. Características de forma, especialmente o HGO, mostrou ser uma das principais preferências em DH, nas pesquisas realizadas a partir de 2005. No entanto, o HGO requer um poder de processamento relativamente alto. No que diz respeito às características Haar, são características de aparência robustas em cenas dinâmicas que consomem menos tempo de processamento, ao comparar com a técnica HGO. Por esse motivo, a técnica de Viola e Jones (2001) foi escolhida.

Os testes realizados, utilizando o protótipo, indicaram uma ótima velocidade, que pode se adequar aos custos do projeto através da configuração de parâmetros da técnica de Viola e Jones (2001), utilizando a biblioteca OpenCV. Os seguintes parâmetros foram ajustados para atingir uma menor taxa de FP e uma maior quantidade de faces detectadas corretamente:

a) *cascade* – arquivo "haarcascade_frontalface_default.xml", melhor configuração de arquivo de *cascade* encontrada nos experimentos, para obter a maior quantidade de faces detectadas;
b) *scaleFactor* – 1.1: definido para um maior ganho na quantidade de faces detectadas;
c) *minNeighbors* – 5: definido para diminuir a taxa de FP;
d) *minSize* – 40: valor adequado para uma boa qualidade de reconhecimento, como demonstrado nos experimentos do protótipo.

Nos testes realizados com a câmera do protótipo, de resolução 720p, foi verificada uma velocidade média de detecção de 14 FPS, utilizando a combinação

de técnicas de DH e DF desta seção. A combinação de técnicas propostas supera os resultados de outros métodos, que representam o estado da arte, como o trabalho de Li et al. (2015), que atingiu a mesma marca (14 FPS), porém com uma resolução inferior de 480p.

4.4.3 Reconhecimento de faces

A arquitetura MobileNet, baseada em RNCP, é rápida, se adapta às diferentes necessidades de processamento, elimina a necessidade do tratamento de imagens antes da etapa do reconhecimento, se adapta a variação de luminosidade, pose, rotação e expressão facial, o que a torna muito eficiente em sistemas de segurança, sendo então definida como a arquitetura utilizada no RF do protótipo.

Os 16 modelos pré-treinados da RNCP MobileNet, podem ser configurados no protótipo e selecionados para gerar diferentes resultados de treinamentos utilizando faces, que podem ser escolhidos para o RF.

4.4.4 Resultados em banco de dados de faces popular

Foi utilizado o banco de dados *Labeled Faces in the wild* para realizar testes com as diferentes configurações da RNCP MobileNet, a fim de se testar os métodos de treinamento e o RF do sistema de controle de acesso. De acordo com a arquitetura MobileNet, são necessárias ao menos 20 imagens de cada classe de objeto e no máximo 2^{27}-1 imagens para que a RNCP funcione adequadamente. Por esse motivo, e para restringir a quantidade máxima de pessoas (31), foram selecionadas apenas as pessoas que continham a quantidade mínima de 30 imagens (maior número após 20 imagens) e a quantidade máxima de 500 imagens por pessoa, durante o treinamento realizado.

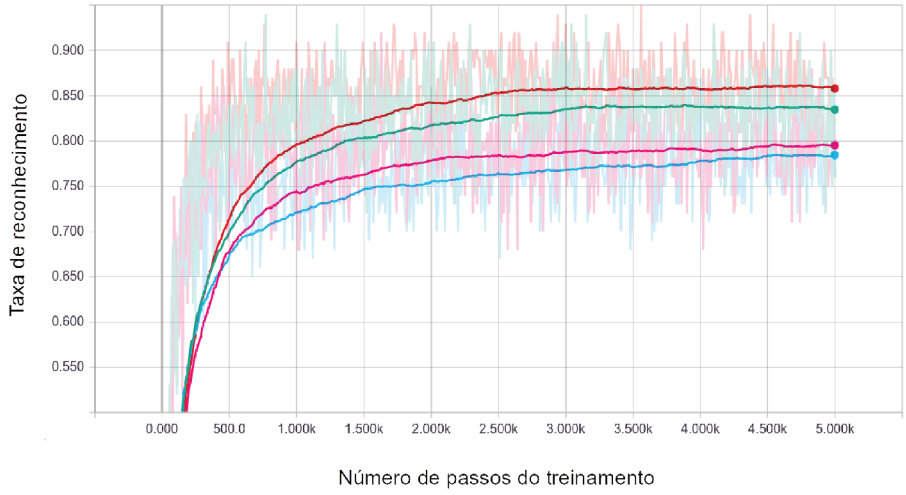
Com o auxílio de uma ferramenta de visualização de resultados de treinamento da biblioteca TensorFlow (ABADI et al., 2016), foi possível verificar quando o treinamento deveria ser parado ao atingir bons resultados (5000 passos).

Os treinamentos foram executados três vezes, sendo que cada treinamento utilizou todas as combinações de Multiplicador de Resolução da RNCP MobileNet, isto é, 128, 160, 192, 224, e todas as combinações de Multiplicador de Largura, 1.0,

0.75, 0.50 e 0.25. Foram utilizadas 10% das imagens do banco de dados para teste (imagens não utilizadas no treinamento).

O Gráfico 4.1 exibe quatro melhores resultados após 5000 passos de treinamento, em uma das 3 execuções realizadas. As 4 linhas do Gráfico 4.1 foram suavizadas para uma melhor visualização. O gráfico real, com a variação da precisão, é exibido com transparência.

Gráfico 4.1 – Acompanhamento da precisão do treinamento (Linhas de baixo para cima, com legendas no formato "Multiplicador de Resolução – Multiplicador de Largura": 224 – 1.0, 128 – 1.0, 160 – 0.75, 192 – 1.0).



Fonte: Autor.

Após três execuções do treinamento, os três primeiros resultados com a melhor taxa de precisão TRF foram:

a) Multiplicador de Resolução 192, tamanho relativo 1.0, TRF: 86.1272%;

b) Multiplicador de Resolução 192, tamanho relativo 1.0, TRF: 86.7052%;

c) Multiplicador de Resolução 192, tamanho relativo 1.0, TRF 86.1272%

Portanto, a média da TRF, dentre as três execuções, foi de aproximadamente 86,32%.

No trabalho de Shafey (2017), utilizando técnicas tradicionais como *PCA*, *LDA*, filtros Gabor, *LBP* e *ISV*, atingiu-se a taxa máxima de TRF de 76%, nesse mesmo banco de dados, porém utilizando todas as imagens. O trabalho de Schroff e Philbin (2015), considerado o estado da arte no reconhecimento de faces, utilizando RNP, atingiu a impressionante marca de 99,63%, porém pagando o preço de 7,5

milhões de parâmetros e 1,6 bilhão de *MACs*, enquanto que a melhor configuração MobileNet alcançou a taxa de 86,70%, utilizando 4,24 milhões de parâmetros e 418 milhões de *MACs*.

4.5 IMPLEMENTAÇÃO DO *SOFTWARE* DO SISTEMA DE CONTROLE DE ACESSO

O sistema de RF consiste no armazenamento das faces classificadas, por pessoa, para que seja possível realizar o reconhecimento das pessoas que estão transitando no ambiente delimitado.

Dado o conjunto de todas as faces do sistema, existe um subconjunto de faces que pertence a um indivíduo cadastrado no sistema. Após o cadastramento de algumas faces, de, pelo menos, dois indivíduos, executa-se a classificação das faces através da utilização do componente de Treinamento, armazenando os resultados da classificação, para que possam ser utilizados posteriormente no reconhecimento.

A captura de faces é realizada continuamente durante a passagem dos indivíduos, de forma que o sistema seja cada vez mais capaz de reconhecer as mesmas pessoas ou reconhecer novas pessoas.

Inicialmente o protótipo de sistema foram implementados na língua inglesa, mas a internacionalização pode ser implementada futuramente. O guia de utilização e as telas do protótipo podem ser vistos no Apêndice C.

4.5.1 Cadastros

Uma instância do sistema, configurada em um ambiente específico, como por exemplo a porta de entrada de uma empresa, é denominada Experimento do sistema. Diferentes Experimentos permitem utilizar o sistema em diferentes lugares, com diferentes configurações e pessoas. Dentro de cada instância de Experimento, é possível cadastrar Câmeras, Registros de Pessoas e Configurações de Treinamentos.

A tela “Experimentos”, permite configurar a largura de uma cabeça em milímetros, utilizada para se calcular a resolução da face (*pixels*/mm); os diretórios dos experimentos, que são definidos a partir do diretório raiz, configurado na tela

"Configuração Geral". Nesta tela também é possível cadastrar a TRF de referência, que indica o valor mínimo necessário para que, novas faces, capturadas durante o reconhecimento, sejam automaticamente incorporadas a um Registro de Pessoa conhecida, ou para que, em caso contrário, as faces sejam incorporadas em um Registro de Pessoa "anônima".

O cadastro de Câmeras permite configurar os parâmetros de conexão da câmera, uma descrição da câmera e o tipo da câmera quanto à localização: entrada ou saída do ambiente. Outras propriedades de câmeras, tais como sensor de imagem, distância focal, abertura, ângulos de visão horizontal e vertical e altura do chão podem ser configurados opcionalmente, para que o sistema auxilie o usuário na configuração do ambiente projetado, informando dados sobre distância e largura máxima do campo de visão e ângulo de posicionamento da câmera. Outras configurações possíveis, no cadastro de câmeras, são os parâmetros de Subtração de Fundo e os parâmetros de DF, como podem ser observados na tela "Câmeras".

Nas configurações da DH, é possível determinar o limiar de sensibilidade da Subtração de Fundo e o tempo de reinício de fundo, isto é, o tempo para que a imagem de fundo seja capturada novamente, sem humanos, para que pequenas diferenças sejam descartadas. Também foi desenvolvido um delimitador de largura do ambiente, para que não seja necessário procurar faces em regiões nas laterais da imagem, que podem conter paredes ou plantas que se movem com o vento.

A configuração de DF permite configurar o tamanho mínimo de captura de uma face; a quantidade mínima de faces vizinhas necessárias, que melhora a rejeição de FP, quando é aumentada; o fator de escala de tamanho da imagem, que torna a detecção mais rápida ao aumentar seu valor. Esses parâmetros são definidos e passados para a interface de DF de Viola e Jones (2001).

Também foi implementado um parâmetro que serve para determinar o tamanho do recorte em volta de uma face, para que seja possível realizar experimentos utilizando informações de uma área maior, que compreende a face e cabelo, por exemplo, ou uma região menor, que compreende somente a região dos olhos, nariz e boca.

Foram desenvolvidos dois parâmetros para o controle de tempo de DF e RF: um determina o intervalo de tempo de reconhecimento entre duas faces capturadas, para que o sistema consiga processar o reconhecimento, sem sobrecarregar o processamento.

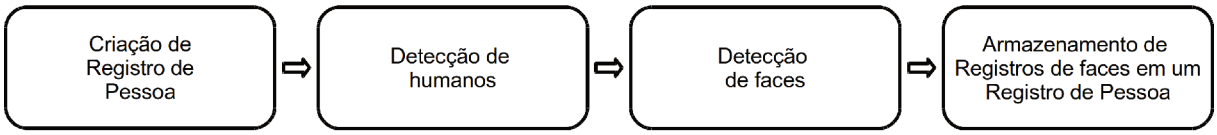
Outro parâmetro determina a quantidade de *threads* (execuções simultâneas) de reconhecimento. O processo de reconhecimento foi paralelizado, de forma que, mais *frames* sejam processados durante a passagem do indivíduo, possibilitando uma maior quantidade de imagens reconhecidas.

Adicionalmente, foi implementado um parâmetro que possibilita realizar o reconhecimento somente no final da passagem do indivíduo, após coletar todas as faces durante a passagem. Desta forma, é possível coletar um número maior de faces, pois não está sendo feito o reconhecimento durante a passagem, mas somente o processo de DF, que é bastante rápido.

O cadastro de um Registro de Pessoa é iniciado pelo nome, conforme figura da tela “Pessoas Reconhecidas”. Em seguida, suas faces são capturadas, de forma controlada, durante a passagem do indivíduo. Quanto mais faces e mais diferentes poses, melhor será a generalidade do reconhecimento.

As faces detectadas durante a passagem do indivíduo são armazenadas no banco de dados em Registros de Faces, contidos em um Registro de Pessoa, como pode ser visto no processo da Figura 4.6.

Figura 4.6 – Etapas para armazenar as faces em um Registro de Pessoa.



Fonte: Autor

Após o cadastro das pessoas conhecidas em Registros de Pessoas, contendo Registros de Faces, cadastra-se uma Configuração de Treinamento, que permite a selecionar os parâmetros de treinamento da RNCP MobileNet, como pode ser visto na tela “Treinamentos”. Os principais parâmetros, que podem ser configurados, são o Multiplicador de Resolução (128, 160, 192 ou 224), o Multiplicador de Largura (0.25, 0.50, 0.75 e 1.0), a quantidade de passos de treinamento, a porcentagem da base utilizada para validação, e a porcentagem da base para testes (faces não utilizadas no treinamento).

Outros parâmetros opcionais servem para realizar alterações aleatórias nas imagens, de forma que se obtenham mais variações de luminosidade, tamanho e escala das imagens durante o treinamento. A utilização desses parâmetros pode propiciar uma maior variação, que pode melhorar a qualidade do reconhecimento,

antecipando imagens em diferentes contextos de luminosidade ou diferentes imagens, que são geradas utilizando um recorte aleatório ou uma escala aleatória em cima da imagem original.

4.5.2 Classificação das faces

Após o cadastro de, ao menos, duas pessoas, é possível acionar o processo de classificação das faces de um experimento, através da execução do treinamento de uma Configuração de Treinamento, que gera um Resultado de Treinamento. A execução do treinamento utiliza somente os Registros de Faces de Pessoas conhecidas, gerando um Resultado de Treinamento, que armazena as características das faces, que serão utilizadas no processo de reconhecimento. A Configuração de Treinamento pode ser feita na tela "Treinamento".

A constante captura de faces, requer que novos treinamentos sejam feitos, de forma que o RF reconheça novas faces, em diferentes situações.

4.5.3 Reconhecimento das faces

O processo de DH e DF atua continuamente sobre um ambiente, controlando quem está entrando, através da câmera de entrada, e quem está saindo, através da câmera de saída. Tal processo pode ser acompanhado na tela "Controle de Suspeitos".

Quando uma pessoa entra no ambiente, um Registro de Passagem de Entrada é gerado e, quando essa pessoa sai do ambiente, é gerado um Registro de Passagem de Saída, podendo ser um registro de pessoa reconhecida ou anônima.

Um Registro de Passagem contém os Registros de Faces, a data e hora da passagem, a identificação da pessoa, nos casos em que foi reconhecida, ou a identificação "anônima", quando não foi reconhecida, como se pode observar na tela "Registros de Suspeitos". Essa tela permite visualizar os Registros de Passagens, visualizar e excluir as faces relacionadas. Também permite realizar o reconhecimento de todos os registros novamente, utilizando um Resultado de Treinamento diferente.

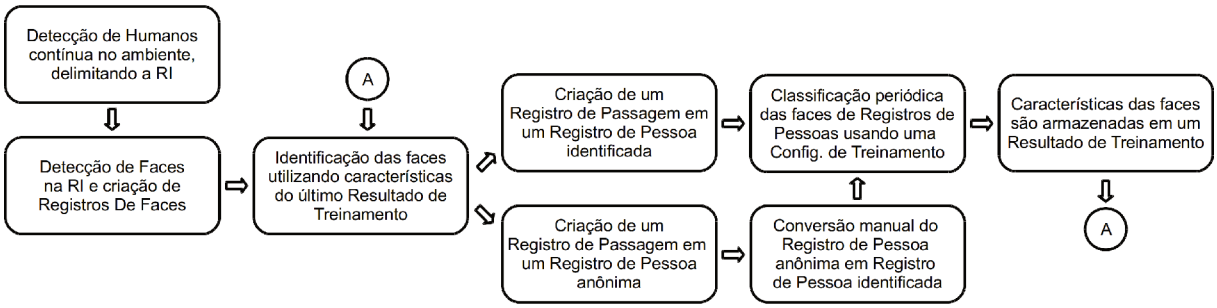
A TRF é calculada como sendo a média das taxas de reconhecimento de cada imagem, registradas durante uma passagem. Quando uma pessoa que está

registrada no sistema é identificada, durante a passagem, com uma taxa de confiabilidade superior a taxa de referência, que pode ser configurada na tela “Experimento”, as novas faces registradas são automaticamente adicionadas aos registros de faces da pessoa conhecida.

Quando o indivíduo ainda não foi registrado no sistema ou a TRF foi inferior à taxa de referência, a palavra “anônima” é atribuída ao Registro de Pessoa. Para os casos de anônimos, o sistema permite converter, manualmente, o Registro de Pessoa “anônima” e seus Registros de Faces em um Registro de Pessoa conhecida, ao atribuir um nome e clicar no botão para tal fim, na tela de “Registro de Pessoa”.

Após a conversão do Registro de Pessoa “anônima” em um Registro de Pessoa conhecida, deve ser executado novamente um treinamento, para que o sistema passe a reconhecer essa pessoa. Uma nova execução do treinamento também introduz as novas faces registradas de indivíduos conhecidos, fazendo com que o reconhecimento seja cada vez mais preciso, já que uma maior quantidade de faces, incluindo faces mais recentes e uma maior variação de poses e oclusões está sendo utilizada no treinamento. A Figura 4.7 ilustra os processos de classificação e reconhecimento das faces.

Figura 4.7. Processos de classificação e reconhecimento de faces



Fonte: autor

4.6 TESTES E RESULTADOS EM CENÁRIO REAL

O experimento foi realizado em uma sala de aula, durante a noite, em um espaço com 8 metros de comprimento por 2 de largura, fundo branco com alguns objetos e iluminação variando de 250 a 300 lux, de acordo com o posicionamento das lâmpadas fluorescentes tubulares e dos refletores, ou com valor máximo de 400 lux, quando próximo de refletores de LED. A medição de luminosidade foi realizada

com o auxílio de um aplicativo de celular, que utiliza a câmera para calcular o número em lux, podendo não representar uma medida precisa. Foram utilizadas duas câmeras com resolução 720p, com características definidas na arquitetura de *hardware*.

Para otimizar a quantidade de faces capturadas, foram realizadas quatro sessões de gravações de faces por pessoa, totalizando de 50 pessoas e 9698 imagens contendo faces.

Na primeira sessão de gravações, na entrada no ambiente, foi utilizada uma câmera com zoom fixo, com características da câmera definida na arquitetura de *hardware*

Para a segunda sessão, na saída do ambiente, foi utilizada uma câmera com com mesmas características, mas com ajuste de zoom e foco, onde optou-se por aumentar um pouco o zoom e acertar o foco para a região de captura, a mesma definida para a primeira câmera, calculada utilizando 40 *pixels* de largura de face. Foi verificada uma pequena variação de luminosidade e de cor com relação à primeira sessão, devido ao zoom aplicado e as diferenças de materiais das câmeras.

Na terceira sessão, utilizando a câmera da entrada no ambiente, foi utilizado um refletor de LED de 4500 lumens, com ângulo de abertura de 120 graus, posicionado logo abaixo da câmera, projetando a luz na mesma direção da câmera, onde observou-se um aumento de luminosidade ao aproximar da câmera, principalmente nos dois metros mais próximos da câmera.

Na quarta sessão, na saída do ambiente, também foi utilizado o refletor abaixo da câmera de saída. A variação de luminosidade também foi percebida principalmente nos dois metros próximos da câmera.

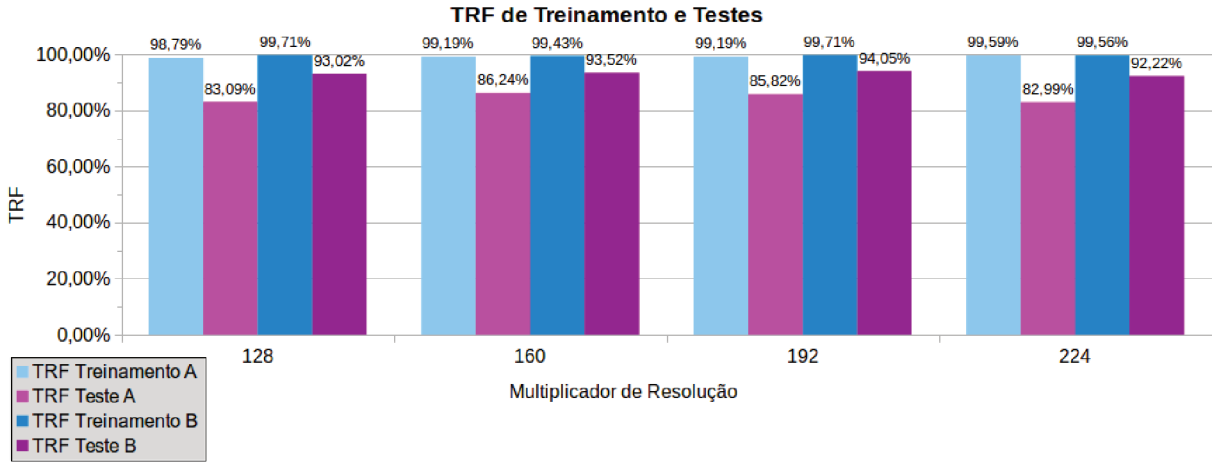
Na primeira bateria de testes de reconhecimento (Testes A), a primeira sessão de captura de faces foi utilizada para realizar os treinamentos (Treinamento A), e as demais três sessões foram utilizadas para testar o reconhecimento, onde foram testadas 7206 imagens (74,31% do total).

A segunda bateria de testes de reconhecimento (Testes B) utilizou as três primeiras sessões para treinamento (Treinamento B) e a quarta seção para testes, a fim de se avaliar a capacidade de generalização do sistema, onde foram testadas 2699 imagens (27,83% do total).

Para cada um dos Treinamentos, A e B, foram realizados quatro treinamentos, com Multiplicadores de Resolução 128, 160, 192 e 224, Multiplicador de Largura 1.0,

10% da base para testes e 5000 passos de treinamento. Os resultados podem ser observados no Gráfico 4.2

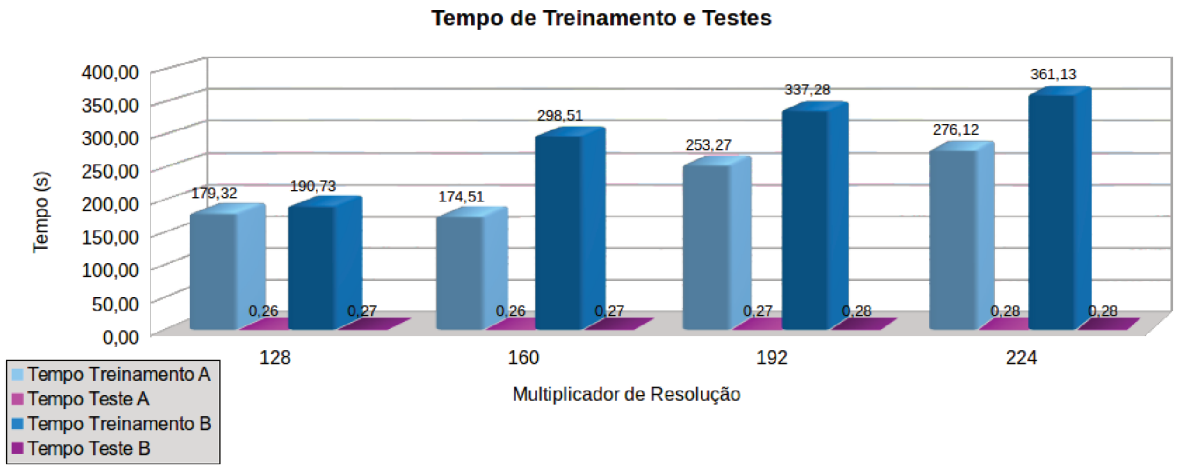
Gráfico 4.2 – TRF de treinamento e testes



Fonte: Autor

A TRF considerada foi calculada como sendo a média das TRFs de cada passagem de pedestre. Medições nos tempos de treinamento e testes foram feitas a fim de se avaliar a progressão, como se observa no Gráfico 4.3.

Gráfico 4.3 – Tempo de treinamento e testes



Fonte: Autor

Nos treinamentos e testes foi observado que, o tempo de treinamento e o tempo de reconhecimento aumentaram com o aumento do Multiplicador de Resolução e também com o aumento da base treinada.

4.6.1 Resultados utilizando filtros de iluminação em imagens com faces

A iluminação de uma face, em uma imagem digital, depende de diversos fatores, principalmente da iluminação do ambiente, que depende da quantidade de luz do Sol e de lâmpadas. A quantidade de luz exposta na face, emitida por uma lâmpada, depende do seu posicionamento, ângulo de incidência da luz na face, potência, tipo e cor.

Também há de se considerar os parâmetros da câmera, como sensibilidade do sensor no momento da captura, abertura da lente, utilização de compensação de luz de fundo, controle de ganho etc. Ainda, a tonalidade da cor da pele produz distintas luminosidades, o que torna o cálculo da iluminação da face ainda mais complexo.

Por outro lado, a iluminação nas imagens com faces pode ser avaliada utilizando algum fator estatístico, calculado sob a Luminância Relativa da imagem, como a média, mediana, variação ou desvio padrão. Esse fator pode ser calculado, para cada pessoa, como sendo a média dos fatores estatísticos de cada face de uma pessoa.

Foram realizados testes para identificar alterações na quantidade de faces detectadas incorretamente (FP) e testes para identificar alterações na TRF, quando são selecionadas parte das imagens que serão treinadas, aquelas que estão dentro de um intervalo de valores de um fator estatístico da Luminância Relativa.

O valor mínimo do intervalo de valores, para cada pessoa, é definido como sendo o fator médio de cada pessoa menos uma porcentagem sobre o fator médio. O valor máximo do intervalo de valores, para cada pessoa, é definido como sendo o fator médio de cada pessoa mais o valor da mesma porcentagem sobre o fator médio.

Para cada teste é definida uma porcentagem, um valor entre 0% e 100%, sendo um teste a cada 10%. Logo, foram realizados 11 testes por fator estatístico e por tipo de teste, TRF ou FP.

Dois tipos de fatores estatísticos foram utilizados nos testes. Os fatores do primeiro tipo, atuam sobre a Luminância Relativa da imagem inteira, que é representada por uma matriz bidimensional da imagem. Os fatores são: média, mediana, variância e desvio padrão. Os fatores do segundo tipo, atuam sobre o histograma de Luminância Relativa da imagem.

Os fatores do segundo tipo são: fator Tons Escuros e fator Tons Claros, que foram desenvolvidos neste trabalho para identificar a proporção entre tons médios, representados por faces, em sua maioria, e tons escuros ou claros.

O histograma de Luminância Relativa é normalizado com 256 posições, que podem variar em 256 níveis cada (tons de Luminância). Para o cálculo dos fatores do segundo tipo, o histograma é dividido em três partes. As áreas mais escuras, com baixa Luminância Relativa, são representadas pelos níveis do primeiro terço do histograma. As áreas de meia tonalidade, de Luminância Relativa média, são representadas pelos níveis do segundo terço do histograma. As áreas mais claras, com alta Luminância Relativa, são representadas pelos níveis do terceiro terço do histograma.

O fator Tons Escuros é calculado sob o histograma, como sendo a divisão da soma das áreas de meia tonalidade pela soma das áreas escuras, e procura encontrar imagens com áreas de tonalidades médias mais escuras, em relação às áreas de tonalidade escuras, como fundos, cabelo ou acessórios escuros. Por esse motivo, esse fator atrai imagens com faces mais escuras.

O fator Tons Claros é calculado sob o histograma, como sendo a divisão da soma das áreas de meia tonalidade pela soma das áreas claras, e procura selecionar imagens com áreas de tonalidades médias mais claras, em relação às áreas de tonalidades claras, como fundos claros ou partes mais claras nas faces. Por esse motivo, esse fator atrai imagens com faces mais claras.

Foram realizados testes para a avaliação de remoção de FP, utilizando a segunda e terceira seções do experimento (4507 imagens), de onde foram selecionados somente indivíduos que participaram de todas as seções, 43 no total. Adicionalmente, foram inseridas, para cada pessoa, 20 imagens, de tamanho 80x80, contendo somente objetos ou parte de objetos, com diferentes luminosidades.

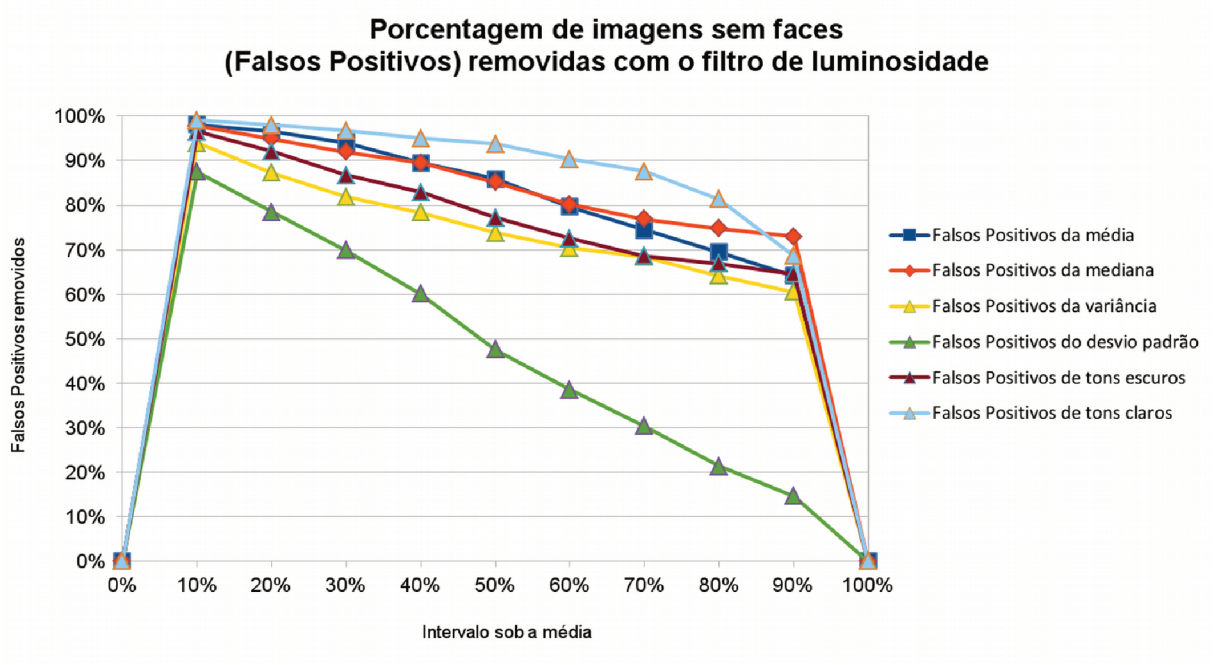
Para a avaliação dos testes de FP removidos, foi gerado o Gráfico 4.4, que exibe a porcentagem de FP, removidos do grupo de imagens cujo fator estatístico está dentro do intervalo do fator.

Observa-se que, a utilização da maioria dos filtros atingiu mais de 95% de remoção de FP, ao utilizar as imagens dentro do intervalo de 10% sobre a média, sendo a maior marca de 99%, ao utilizar o filtro fator Tons Claros.

Para os testes de avaliação da qualidade da TRF, com filtros de luminosidade, foi utilizada a primeira seção de gravações (2239 imagens) para treinamento da

base, e a segunda e terceira seções para validação (4507 imagens). Não foram adicionadas imagens com FP. Em todos os treinamentos, foi utilizado o Multiplicador de Resolução 224, Multiplicador de Largura 1.0, 10% da base para testes do treinamento e 5000 passos de treinamento.

Gráfico 4.4 – Falsos Positivos removidos utilizando o filtro de luminosidade.



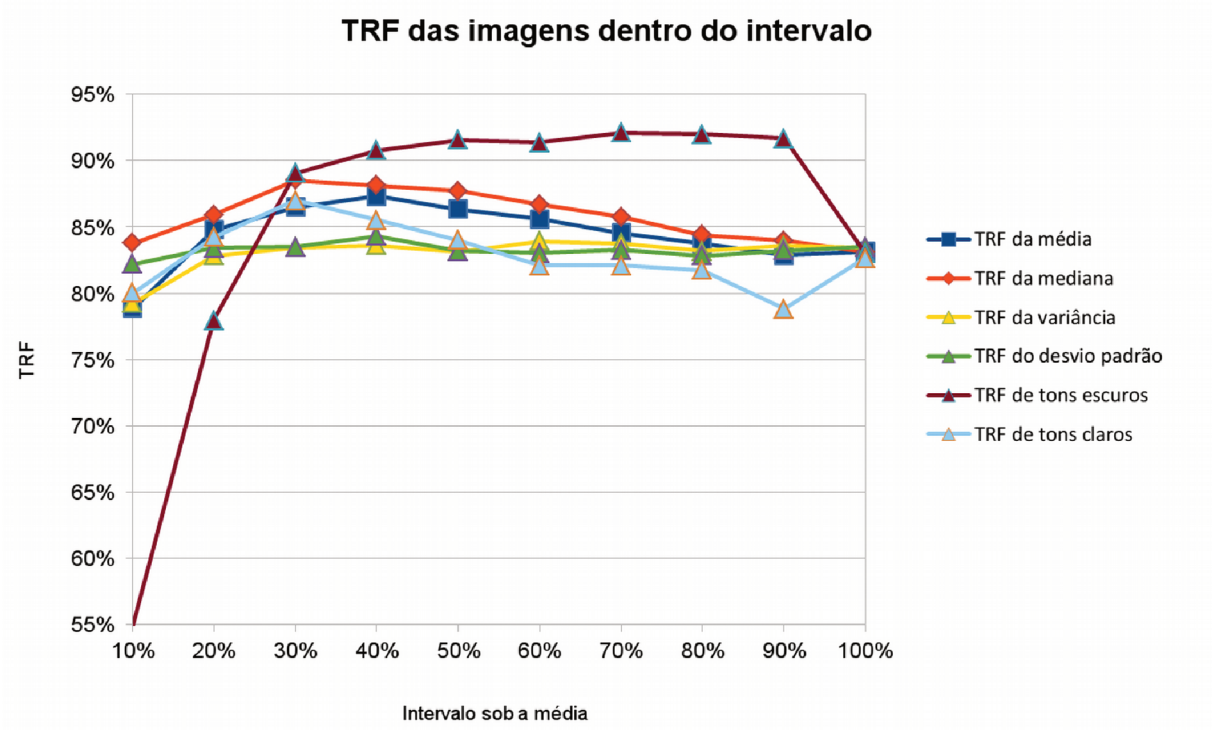
Fonte: autor.

Cada teste de TRF realiza um treinamento, utilizando as imagens que foram filtradas dentro de um determinado intervalo. A mesma porcentagem do intervalo de treinamento é utilizada para os testes de validação, gerando dois grupos de validação. O primeiro grupo de validação é formado por imagens cujo fator estatístico está dentro do intervalo do fator, e o segundo grupo de validação seleciona imagens cujo fator estatístico está fora do intervalo do fator.

Para a avaliação da TRF, foram gerados três gráficos, que sintetizam os resultados dos testes com seleção de imagens. O Gráfico 4.5 apresenta a TRF de validação do grupo de imagens cujo fator estatístico está dentro do intervalo do fator, isto é, as imagens que devem ser armazenadas no sistema.

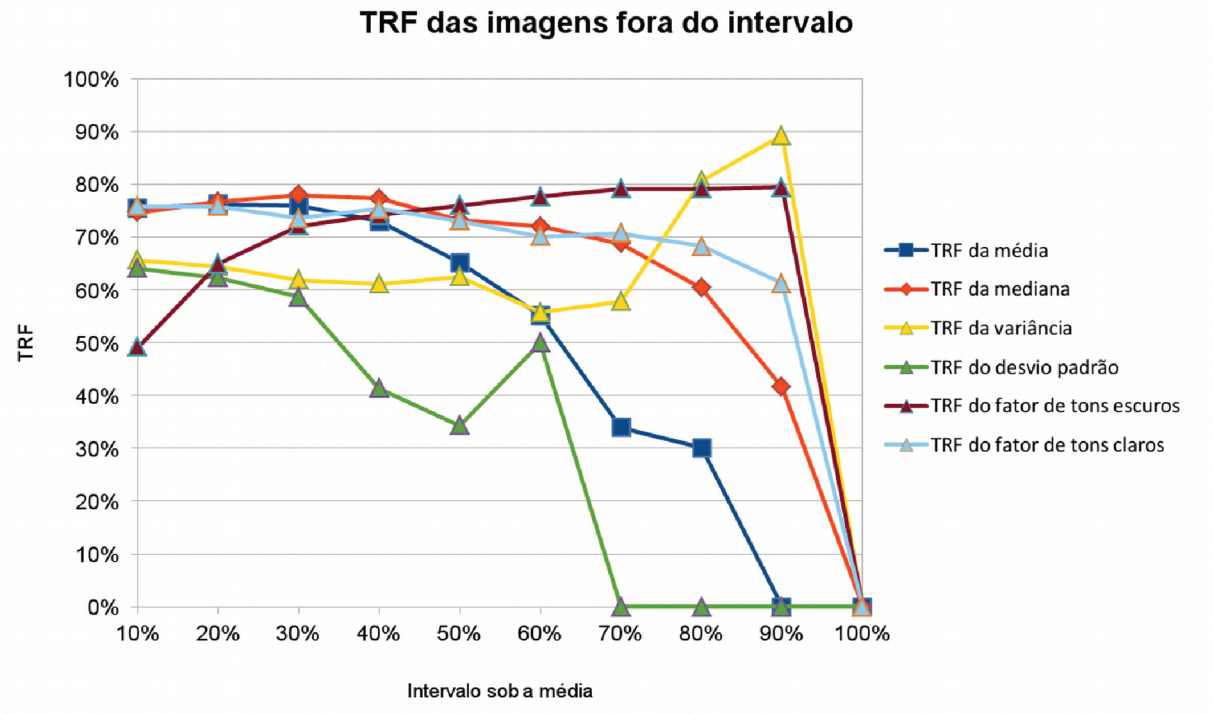
O Gráfico 4.6 apresenta a TRF de validação do grupo de imagens cujo fator estatístico está fora do intervalo do fator, ou seja, as imagens que serão descartadas pelo sistema após a detecção.

Gráfico 4.5 – TRF de validação das imagens dentro do intervalo.



Fonte: autor

Gráfico 4.6 – TRF de validação das imagens fora do intervalo.



Fonte: autor

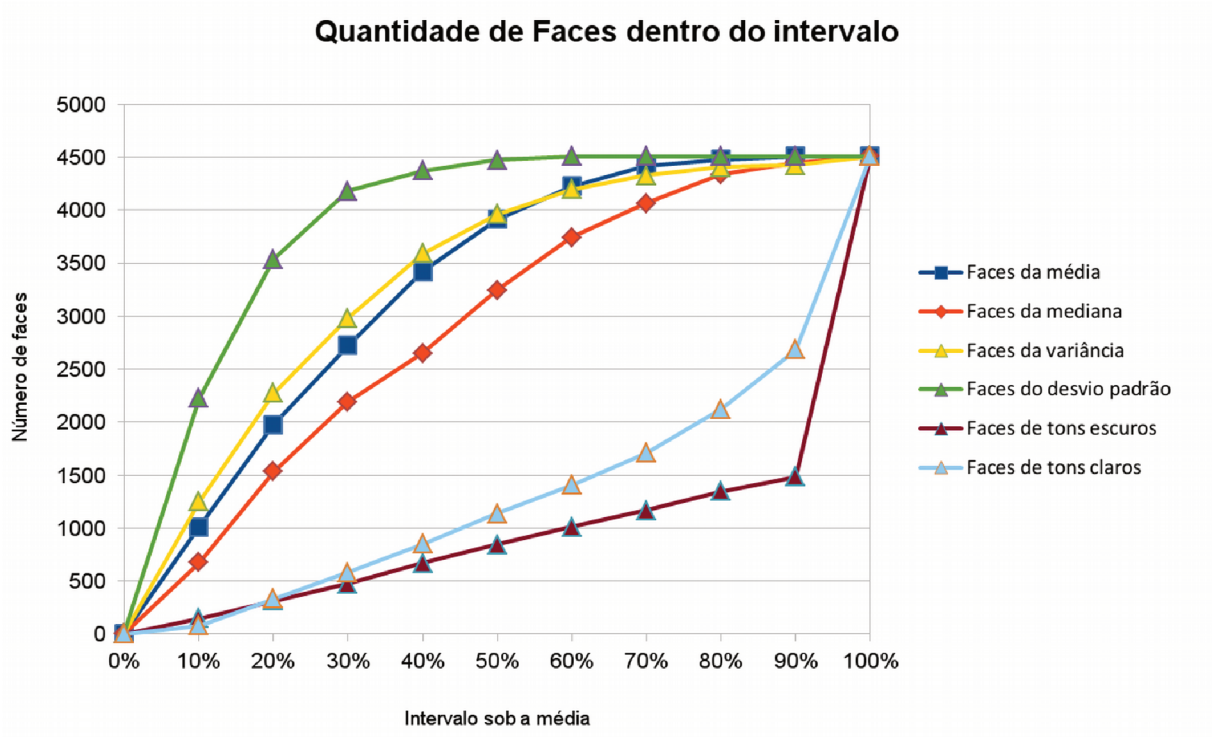
O Gráfico 4.7 apresenta a quantidade de imagens que foi utilizada pelo grupo de imagens cujo fator estatístico está dentro do intervalo do fator.

No gráfico 4.5 verifica-se que, ao utilizar 100% das imagens para treinamento e testes, cenário no qual não está sendo aplicado nenhuma seleção de imagens, a TRF é de, em média, 83,4%.

No Gráfico 4.6, observa-se que, ao selecionar imagens através dos fatores estatísticos, a melhor TRF foi de 92,11%, atingida ao se utilizar o fator Tons Escuros, onde, para cada indivíduo, é utilizado um intervalo de 70% das imagens em torno da média do fator de cada indivíduo, onde se verifica uma melhora de 8,71%.

O Gráfico 4.7 mostra que, a utilização do fator Tons Escuros, com intervalo de 70%, seleciona aproximadamente um terço (1170) das imagens para testes. Ao passo que o intervalo aumenta, é possível selecionar até 1500 imagens, quando o intervalo se aproxima dos 90%, com uma pequena diminuição da TRF (91,70%).

Gráfico 4.7 – Quantidade de imagens dentro do intervalo.



Fonte: autor

O fator mediana também consegue atingir boas TRF, acima de 85%, utilizando um intervalo entre 20% e 70%. A melhor TRF foi de 88,52%, no intervalo de 30%, selecionando 2192 imagens para testes. A utilização desse fator, seleciona uma maior quantidade de imagens, com relação ao fator Tons Escuros, o que pode ser útil quando se deseja ter uma base de dados maior, com uma maior generalização.

O fator média, atinge uma TRF não tão expressiva com relação às suas superiores, no entanto, a quantidade de faces utilizadas é bastante expressiva, e as faces removidas têm uma TRF mais baixa com relação aos fatores Tons Escuros e mediana.

Os outros fatores estatísticos sobre a imagem, variância e desvio padrão, não demonstraram ganhos significativos para a TRF, porém conseguiram uma TRF baixa das imagens fora do intervalo, mas tal fato pode ter ocorrido devido ao maior número de imagens dentro do intervalo.

O fator Tons Claros, não demonstra resultados de melhoria muito expressivos, 87% ao utilizar o intervalo de 30%. Tal fato acontece por haver luminosidade excedente em poucas das fotos de teste. No entanto, esse fator pode ser útil quando se têm muita luminosidade em faces ou uma quantidade expressiva de FP, como demonstra o Gráfico 4.4.

A utilização dos fatores de Tons Claros e Tons Escuros revela que, as faces selecionadas, possuem tamanho semelhante com relação à imagem inteira, pois as faces são representadas, na maior parte, pelas meias tonalidades, o que faz com que seja criada uma relação de tamanho uniforme, entre a área da face e outras áreas da imagem, mais escuras, como cabelo ou acessórios de tons escuros, ou áreas mais claras, como paredes, reflexos ou acessórios de tons claros.

Dependendo do fator e porcentagem escolhidos, a utilização do filtro pode reduzir significativamente a quantidade de imagens para treinamento, o que deve ser levado em consideração, pois a capacidade de generalização de reconhecimento pode evoluir apenas com as imagens com menos variação de luminosidade, ao passo que, conjuntos de faces, capturadas com distintas exposições de luz, em diferentes dias, podem não entrar no rol de imagens de treinamento.

Por fim, o tempo de cálculo da utilização dos fatores estatísticos é muito pequeno, apenas 2 milésimos de segundo em uma imagem de 80x80 *pixels*, utilizando os fatores levam mais tempo para serem calculados.

4.7 CUSTO DO PROTÓTIPO E COMPARATIVO COM OUTROS SISTEMAS DO MERCADO

As Tabelas 4.1 e 4.2 apresentam, respectivamente, orçamentos de duas empresas: FaceMatch e TecLink.

Tabela 4.1 – Proposta de materiais e licenças da empresa Facematch.

Material	Valor	Quantidade	Total (R$)
Licença Base + Licença para Banco de Faces (até 500 faces)	US$640	1	2.560,00
Pacote adicional para cada 500 faces (opcional)	US$200	0	0
Licença para duas (02) câmeras de reconhecimento facial	US$1920	1	7.680,00
Câmera IP vari focal 3M Pixel POE AXIS M1125	R$1183	2	2.366,00
Fonte 12 V para câmera	R$15	2	30,00
Switch 8 portas POE Ativo	R$727	1	727,00
Cabo de rede (25 m) e RJ45	R$30	1	30,00
Computador CPU I7 2.7GHZ 8 GB	R$3254,07	1	3254,00
TOTAL			**16.647,00**

Fonte: autor

A empresa FaceMatch forneceu os valores da licença de *software*, licença das câmeras e especificações das câmeras, que utilizam *POE*. Por esse motivo, foi incluído o valor do *switch POE*.

Tabela 4.2 – Proposta de materiais e licenças da empresa TeckLink.

Material	Valor	Quantidade	Total (R$)
Licença ITechFace	R$4000	1	4000,00
Licença servidor / Suporte mensal	R$300	12	3600,00
Câmera IP AXIS POE (modelo não especificado)	R$1500	2	3000,00
Fonte 12 V para câmera	R$15	2	30,00
Switch 8 portas POE Ativo	R$727	1	727,00
Cabo de rede (25 m) e RJ45	R$30	1	30,00
PLCS e *Hardwares* periféricos	?	-	-
TOTAL			**11.387,00**

Fonte: autor

A empresa TecLink enviou orçamento estimado da licença de *software* e de suporte, mas não especificou modelos de câmeras, apenas informou o fabricante das câmeras, o tipo POE e os valores estimados. Também foi informado que mais componentes seriam necessários, tais como placas controladoras, roteadores Wi-Fi

e *switches*, pois o processamento é realizado no servidor, encarecendo bastante a solução para um único ambiente. Devido à arquitetura centralizada, não é necessário um computador, no entanto é necessário o pagamento do suporte mensal.

Em todas as cotações foram adicionados os mesmos valores de fontes de energia, cabos de rede e plugues que foram utilizados no protótipo, pois não foram informados pelas empresas.

Alguns valores foram informados em dólar e convertidos para o valor aproximado na data de cotação (R$4,00/US$), realizada no mesmo dia para todos os equipamentos.

Utilizando materiais de menor custo, com menor poder de processamento e resolução de imagem, projetados nas arquiteturas e ambiente aqui definidos para esse fim, sistemas operacionais e bibliotecas de *software* de código aberto, é possível determinar uma arquitetura com valor bastante reduzido, como se observa na Tabela 4.3.

Tabela 4.3 – Proposta de materiais da arquitetura de baixo custo.

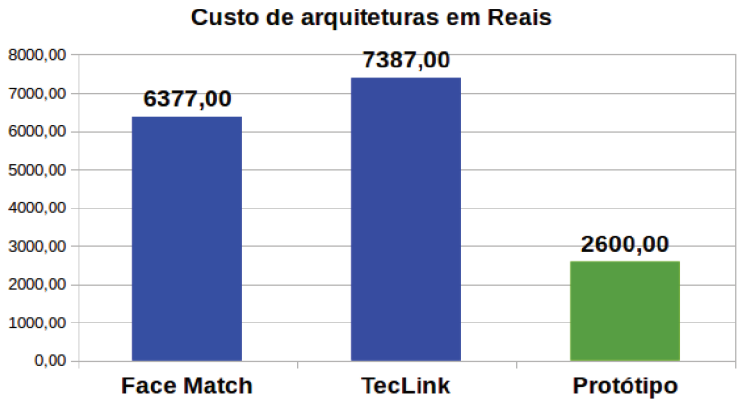
Material	Valor	Qtdd	Total(R$)
Câmera Ip Intelbras Bullet Vip 1120 B 1mp 720p	R$116	2	232,00
Par Splitter/Injector	R$6	2	12,00
Fonte 12 V para câmera	R$15	2	30,00
Switch 8 portas S/ POE Ativo	R$58	1	58,00
Cabo de rede (25 m) e RJ45	R$30	1	30,00
Computador CPU I5 2.7GHZ 4GB	R$2238	1	2.238,00
TOTAL			**2.600,00**

Fonte: autor

O custo das arquiteturas das empresas FaceMatch e TecLink, sem considerar as licenças é de R$6377,0 e R$7387, respectivamente, incluindo o valor da licença do servidor para a empresa TecLink, por utilizar um sistema de processamento central, enquanto que o custo da arquitetura de hardware proposta é de R$2.600,00 como pode ser visto no Gráfico 4.8.

Observa-se que, o custo da arquitetura proposta é 59,23% menor do que o custo da arquitetura da empresa Face Match e 64,80% menor do que o custo da arquitetura da empresa TecLink.

Gráfico 4.8 – Comparativo entre custos de arquiteturas



Fonte: Autor

Sumariamente, a arquitetura proposta representa 37,78% do valor médio das duas arquiteturas encontradas no mercado, aqui pesquisadas.

5 CONCLUSÃO

A maioria dos métodos disponíveis para detectar humanos e faces, e métodos para reconhecer faces, procuram alcançar as mais altas taxas de detecção e de reconhecimento. No entanto, não foram encontrados trabalhos na literatura que abordem a definição de um modelo de sistema de controle de acesso, composto de uma arquitetura de *hardware* e *software*, que forneçam as bases para a implementação de um sistema completo de controle de acesso, para ambientes internos com pouco controle de iluminação e sem controle de pose, capaz de realizar o RF com eficiência e rapidez a um baixo custo.

O projeto de uma plataforma de RF robusta, rápida e adaptável à diferentes condições de iluminação e pose é fator determinante para atingir ótimos resultados de qualidade e velocidade, em ambientes não controlados, como os obtidos com os testes do protótipo. Além disto, um sistema de controle de acesso deve possibilitar a configuração de uma grande quantidade de parâmetros, disponibilizados nos métodos das técnicas utilizadas, para seja possível se adaptar às diferentes necessidades de qualidade e rapidez de reconhecimento.

Com o objetivo de reduzir os custos do *hardware* do projeto é preciso diminuir o volume dos dados utilizados no RF, através do descarte das áreas da imagem que não são processadas, isto é, a delimitação da RI. Na primeira etapa, a RI é rapidamente delimitada pela região que pode conter humanos, através da técnica de Subtração de Fundo, reduzindo a RI para no mínimo metade da imagem (quando o indivíduo está muito próximo da câmera).

Em seguida, a busca por faces é feita somente na área que contem humanos, reduzindo o processamento proporcionalmente ao tamanho da área delimitada na etapa anterior. Portanto, a utilização destes métodos combinados reduz o tempo de detecção em 50%, no pior caso.

À medida que as faces são coletadas, o RF foi implementado para ser executado paralelamente em algumas t*hreads*, o que resultou em uma maior quantidade de faces reconhecidas, otimizando a utilização das CPU's disponíveis e, portanto, reduzindo o custo do projeto.

O protótipo se mostrou muito eficiente no RF, principalmente com o aumento da base treinada. No primeiro treinamento, com poucas imagens por pessoa, atingiu uma TRF média de 86,24%, com apenas 1 pessoa reconhecida incorretamente, ao

utilizar o cálculo que considera a TRF média por passagem. Após a gravação de mais duas passagens por pessoa, a TRF aumentou significativamente para 94,05%, levando apenas 0,2765 segundos para se reconhecer uma face, sem apresentar reconhecimentos errôneos.

Com o aumento da base treinada em 180,86%, houve um aumento de 6,36% no tempo de treinamento, para a configuração com Multiplicador de Resolução 128, e de 30,86%, para a configuração com Multiplicador de Resolução 224. Portanto, a escolha de uma melhor configuração deve ser ponderada ao implementar as janelas de tempo de treinamento dos processos automatizados de treinamento da base. Também há de se considerar o tempo de reconhecimento médio, que normalmente aumenta com o aumento dos multiplicadores de resolução e de largura.

A variação de luminosidade em um ambiente pode ser verificada através da representação da Luminância Relativa em imagens digitais. Experimentos mostraram que, a escolha de um filtro, que utiliza um fator estatístico sobre a Luminância Relativa, pode servir para remover quase 100% de FP, detectados indevidamente ao utilizar a técnica de Viola e Jones (2001).

Ainda, utilizando filtros de Luminância Relativa, verifica-se uma melhora significativa na TRF, de 83,4% para 92,11%, em testes realizados utilizando somente imagens com faces. Portanto, pode-se dizer que a variação de luminosidade no ambiente prejudica consideravelmente o RF, mesmo ao utilizar uma RNCP que se adapta às variações de luminosidade.

Utilizando os métodos deste trabalho, constatou-se que, imagens maiores requerem um maior poder de processamento, desde a captura de imagens, DF e RF. Quanto maior a imagem, melhor deve ser o *hardware* da câmera para capturar e processar as imagens internamente, o que implica um maior custo da câmera e dos equipamentos de rede e do *hardware.*

As imagens das faces capturadas devem possuir um tamanho mínimo para que se obtenha uma boa qualidade de reconhecimento. Por isso, a área de captura deve ser projetada de acordo com o tamanho da imagem capturada, determinadas pelos parâmetros da câmera, tais como resolução, distância focal, abertura e tamanho do sensor. Por esses motivos, deduz-se que o custo do projeto está intimamente relacionado com a área de cobertura das câmeras, que podem ser

calculadas previamente, através das equações definidas nos resultados deste trabalho.

Ao lado do custo dos equipamentos, não foram necessários investimentos em compras de *softwares*, devido à grande disponibilidade de bibliotecas que implementam os mais recentes avanços na detecção e RF, reduzindo o tempo de implementação do sistema de controle de acesso e, consequentemente, o custo de mão de obra para a implementação do sistema.

Finalmente, o aumento da disponibilidade de equipamentos, com bom poder de processamento a um custo acessível e a grande quantidade de bibliotecas de *software,* que implementam métodos rápidos de detecção e reconhecimento, ativamente desenvolvidas pela comunidade de desenvolvedores de *software* livre, torna possível a criação de sistemas de controle de acesso de baixo custo, qualificados para realizar o RF em ambientes que, de certa forma, são delimitados à capacidade dos equipamentos, mas que, em contrapartida, podem ser muito eficientes com a modelagem adequada da arquitetura de *hardware* e *software*.

Em trabalhos futuros, deseja-se criar um processo dentro do sistema que controle o volume de dados da base. Com o aumento da base de dados, o tempo de treinamento aumenta consideravelmente, o que pode ser controlado com uma limitação da quantidade de dados treinados, através de um processo para manter poucas imagens de cada passagem, mantendo a base de treinamento cada vez mais heterogênea e generalista.

Deseja-se implementar no protótipo uma funcionalidade que possibilite a escolha de um filtro de fator estatístico de Luminância Relativa, ou vários filtros combinados e suas respectivas porcentagens de intervalo. A análise da combinação dos filtros, através de experimentos, com uma maior variação de luminosidade, também é desejada para o melhor entendimento dos resultados das combinações. Dessa forma, espera-se reduzir a maioria dos FP, atingir melhores TRF e, até mesmo, utilizar o filtro para controlar o volume da base de dados.

Outra importante implementação desejada é uma funcionalidade que torna o sistema capaz de rastrear as pessoas que estão passando pelo ambiente. Assim, é possível identificar mais de uma pessoa ao mesmo tempo, pois as faces que estão sendo capturadas, podem ser agrupadas por pessoa que está sendo rastreada.

REFERÊNCIAS

ABADI, M. et al. TensorFlow : A System for Large-Scale Machine Learning. In: PROC of 12TH USENIX CONFERENCE ON OPERATING SYSTEMS DESIGN AND IMPLEMENTATION 2016, Savannah. **Anais**... Savannah, 2016. p. 265–283.

ABATE, A. F. et al. 2D and 3D face recognition: A survey. **Pattern Recognition Letters**, v. 28, n. 14, p. 1885–1906, 2007.

AHONEN, T.; HADID, A.; PIETIKÄINEN, M. Face description with local binary patterns: Application to face recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 28, n. 12, p. 2037–2041, 2006.

ANGADI, S. A.; KAGAWADE, V. C. A robust face recognition approach through symbolic modeling of Polar FFT features. **Pattern Recognition**, v. 71, p. 235–248, 2017.

BALLARD, D. H. Generalizing the Hough transform to detect arbitrary shapes. **Pattern Recognition**, v. 13, n. 2, p. 111–122, 1981.

BARBOSA, A; BONADIO, I. **Princípios de Processamento de Imagens: Uma introdução à Convolução.** 2018. Disponível em: <https://engenharia.elo7.com.br/convolucao/>. Acesso em: 13 dez. 2018.

BELEZNAI, C.; BISCHOF, H. Fast human detection in crowded scenes by contour integration and local shape estimation. In: 2009 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2009, Miami, FL. **Anais**... Miami, FL, 2009, p. 2246–2253.

BELHUMEUR, P. N.; HESPANHA, J. P.; KRIEGMAN, D. J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 19, n. 7, p. 711–720, 1997.

BEZERRA, R. A. **Proposta de critérios para câmeras de vigilância em aplicações de cftv**. Universidade de Brasília, Brasília, 2012.

BRAGA, L. F. Z. **Sistemas de reconhecimento facial**. 2013. Escola de Engenharia de São Carlos, São Carlos, 2013.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CHEN, L.H.; WANG, L.Y.; SU, C.W. Human Detection in Surveillance Video. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 28, n. 02, p. 1455003, 2014.

CIE. **International Commission on Illumination.** 2019. Disponível em: <http://www.cie.co.at/>. Acesso em: 17 jan. 2019.

CROW, F. C. Summed-area tables for texture mapping. **ACM SIGGRAPH Computer Graphics**, New York, NY, v. 18, n. 3, p. 207–212, 1984.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: PROCEEDINGS 2005 IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, CVPR 2005, San Diego, CA. **Anais**... San Diego, CA, 2005, p. 886–893.

DALAL, N.; TRIGGS, B.; SCHMID, C. Human detection using oriented histograms of flow and appearance. In: LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS) 2006, Berlin, Heidelberg. **Anais**... Berlin, Heidelberg, 2006, p. 428–441.

DATA SCIENCE ACADEMY. **Deep Learning (Book)**. 2018. Disponível em: <http://deeplearningbook.com.br/funcao-de-ativacao/>. Acesso em: 2 ago. 2018.

FELZENSZWALB, P. F. et al. Object Detection with Discriminatively Trained Part Based Models. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 9, p. 1–20, 2009.

FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Distance transforms of sampled functions. **Cornell Computing and Information Science Technical Report TR20041963**, v. 4, p. 1–15, 2004.

FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Pictorial structures for object recognition. **International Journal of Computer Vision**, v. 61, n. 1, p. 55–79, 2005.

FELZENSZWALB, P.; MCALLESTER, D.; RAMANAN, D. A discriminatively trained, multiscale, deformable part model. In: 26TH IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, CVPR 2008, Anchorage. **Anais**... Anchorage, 2008, p. 1–8.

FISCHLER, M. A.; ELSCHLAGER, R. A. The Representation and Matching of Pictorial Structures Representation. **IEEE Transactions on Computers**, v. C–22, n. 1, p. 67–92, 1973.

GAVRILA, D. M. The Visual Analysis of Human Movement: A Survey. **Computer Vision and Image Understanding**, v. 73, n. 1, p. 82–98, 1999.

GAVRILA, D. M. Pedestrian Detection from a Moving Vehicle. In: ECCV EUROPEAN CONFERENCE ON COMPUTER VISION 2000, Antibes. **Anais**... Antibes, 2000, p. 37–49.

GAVRILA, D. M. A Bayesian, exemplar-based approach to hierarchical shape matching. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 29, n. 8, p. 1408–1421, 2007.

GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2014, Columbus. **Anais**... Columbus, 2014, p. 863–879.

GRGIC, M.; DELAC, K.; GRGIC, S. SCface – Surveillance cameras face database. **Multimedia Tools and Applications**, v. 51, n. 3, p. 863–879, 2011.

GROSS, R. et al. Multi-PIE. **Image and Vision Computing**, v. 28, n. 5, p. 807–813, 2010.

HE, K. et al. Deep Residual Learning for Image Recognition. In: 2016 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR) 2016, Las Vegas. **Anais**... Las Vegas, 2016, p. 770–778.

HE, Z. et al. Robust FEC-CNN: A High Accuracy Facial Landmark Detection System. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS 2017, Honolulu. **Anais**... Honolulu, 2017, p. 2044–2050.

HOWARD, A. G. et al. **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications**. 2017a. Disponível em: <https://arxiv.org/abs/1704.04861>. Acesso em: 7 jun. 2017.

HOWARD, A. G. et al. **Google AI Blog: MobileNets: Open-Source Models for Efficient On-Device Vision**. 2017b. Disponível em: <https://ai.googleblog.com/2017/06/mobilenets-open-source-models-for.html>. Acesso em: 15 maio. 2018.

HU, C. et al. A new face recognition method based on image decomposition for single sample per person problem. **Neurocomputing**, v. 160, p. 287–299, 2015.

HUANG, G. B. et al. **Labeled Faces in the Wild: A Database for studying Face Recognition in Unconstrained Environments**. Massachussetts. Disponível em: <https://hal.inria.fr/inria-00321923>.

HUANG, G. B.; LEARNED-MILLER, E. **Labeled faces in the wild : Updates and new reporting procedures**. Massachussetts. Disponível em: <https://pdfs.semanticscholar.org/2d34/82dcff69c7417c7b933f22de606a0e8e42d4.pdf>

HUSSAIN, S. U.; TRIGGS, B. Feature Sets and Dimensionality Reduction for Visual Object Detection. In: BRITISH MACHINE VISION CONFERENCE 2010, Aberystwyth. **Anais**... Aberystwyth, 2010, p. 112.1–112.10.

IANDOLA, F. N. et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. **arXiv**, [s. l.], p. 13, 2016. Disponível em: <http://arxiv.org/abs/1602.07360>

IEC. **International Electrotechnical Commission**. 2018. Disponível em: <http://www.iec.ch/>. Acesso em: 1 mar. 2018.

IEEE, C. S. IEEE Standard for Ethernet - Section Two. **IEEE Standard for Ethernet**, [s. l.], v. 2012, n. December, p. 1–400, 2012.

IEEE, E. W. G. **IEEE 802.3 ETHERNET WORKING GROUP**. 2018. Disponível em: <http://www.ieee802.org/3>. Acesso em: 1 mar. 2018.

IOFFE, S.; FORSYTH, D. A. Probabilistic methods for finding people. **International Journal of Computer Vision**, v. 43, n. 1, p. 45–68, 2001.

IOFFE, S.; SZEGEDY, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. **arXiv:1502.03167 [cs]**, 2015.

ISO. **ISO/IEC 14496-10:2014 - Information technology - Coding of audio-visual objects - Part 10: Advanced Video Coding**. 2014. Disponível em: <https://www.iso.org/standard/66069.html.>. Acesso em: 1 mar. 2018.

JALALI, A.; MALLIPEDDI, R.; LEE, M. Sensitive deep convolutional neural network for face recognition at large standoffs with small dataset. **Expert Systems with Applications**, v. 87, p. 304–315, 2017.

JONATHON PHILLIPS, P. et al. The FERET evaluation methodology for face-recognition algorithms. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 10, p. 1090–1104, 2000.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. 1 ImageNet Classification with Deep Convolutional Neural Networks. **Advances In Neural Information Processing Systems**, p. 1097–1105, 2012.

LADES, M. et al. Distortion invariant object recognition in the dynamic link architecture. **IEEE Transactions on Computers**, v. 42, n. 3, p. 300–311, 1993.

LEE, Roger. **Software engineering research, management and applications**. 1ª edição, Berlin, Heidelberg: Springer, 2018.

LEIBE, B.; LEONARDIS, A.; SCHIELE, B. Robust object detection with interleaved categorization and segmentation. **International Journal of Computer Vision**, v. 77, n. 1–3, p. 259–289, 2008.

LEIBE, B.; SEEMANN, E.; SCHIELE, B. Pedestrian detection in crowded scenes. In: COMPUTER VISION AND PATTERN RECOGNITION, 2005. CVPR 2005. IEEE COMPUTER SOCIETY CONFERENCE ON 2005, San Diego, CA. **Anais**... San Diego, CA, 2005, p. 878–885.

LI, H. et al. Efficient boosted exemplar-based face detection. In: PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2014, Columbus. **Anais**... Columbus, 2014, p. 1843–1850.

LI, H. et al. A convolutional neural network cascade for face detection. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2015, Massachussetts. **Anais**... Massachussetts, 2015, p. 5325–5334.

LIN, Z. et al. Hierarchical part-template matching for human detection and segmentation. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION 2007, **Anais...** Rio de Janeiro, 2007, p. 1–8.

LUO, P. et al. Switchable deep network for pedestrian detection. In: PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2014, Columbus. **Anais**... Columbus

MALLICK, S. **Histogram of Oriented Gradients | Learn OpenCV**. 2016. Disponível em: <https://www.learnopencv.com/histogram-of-oriented-gradients/>. Acesso em: 8 ago. 2018.

MAZZA, L. O. **Aplicacão De Redes Neurais Convolucionais Densamente Conectadas No Processamento Digital De Imagens Para Remoção De Ruído Gaussiano**. 2017. Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2017.

MIKOLAJCZYK, K.; SCHMID, C.; ZISSERMAN, A. Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. In: PROC. EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV) 2004, Prague. **Anais**... Prague Disponível em: <http://link.springer.com/10.1007/978-3-540-24670-1_6>

NGUYEN, D. T.; LI, W.; OGUNBONA, P. O. Human detection from images and videos: A survey. **Pattern Recognition**, [s. l.], v. 51, p. 148–175, 2016. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2015.08.027>

OLIVIERO, A. W.; BILL. **Cabling: the complete guide to copper and fiber-optic networking**. 4. ed Indianapolis, Wiley, 2014.

OPENCV. **Open Source Computer Vision Library**. 2018. Disponível em: <http://opencv.org/>. Acesso em: 1 mar. 2018.

OTT, P.; EVERINGHAM, M. Implicit color segmentation features for pedestrian and object detection. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION 2009, Kyoto. **Anais**... Kyoto, 2009, p. 723–730.

OUYANG, W.; WANG, X. A discriminative deep model for pedestrian detection with occlusion handling. In: PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2012, Providence. **Anais**... Providence, 2012, p. 3258–3265.

OUYANG, W.; WANG, X. Joint deep learning for pedestrian detection. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION 2013, Portland. **Anais**... Portland, 2013, p. 2056–2063.

PAPAGEORGIOU, C. P.; OREN, M. A general framework for object detection. In: COMPUTER VISION, IEEE INTERNATIONAL CONFERENCE ON 1998, Bombay. **Anais**... Bombay, 1998, p. 555–562.

PONTI, M. A.; DA COSTA, G. B. P. Como funciona o Deep Learning. In: SBC, S. B. de C. (Ed.). **Tópicos em Gerenciamento de Dados e Informações 2017**. Uberlândia. p. 63–93.

REDMON, J. et al. You Only Look Once: Unified, Real-Time Object Detection. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2016, **Anais**... Las Vegas, 2016, p. 779–788.

ROBOTICS, A. **Nao Robot**. 2014. Disponível em: <https://www.softbankrobotics.com/emea/en/robots/nao>. Acesso em: 1 mar. 2018.

RUSSAKOVSKY, O. et al. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision**, v. 115, n. 3, p. 211–252, 2015.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. 3. ed., Rio de Janeiro, Campus, 2013.

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. FaceNet: A unified embedding for face recognition and clustering. In: PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2015, Boston. **Anais**... Boston, 2015, p. 815–823.

SHAFEY, L. El. **2D Face Recognition: an Experimental and Reproducible Research Survey**, IDIAP, Martigny, 2017.

SHEN, X. et al. Detecting and aligning faces by image retrieval. In: PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2013, Portland. **Anais**... Portland, 2013, p. 3460–3467.

SIMONYAN, K.; ANDREW ZISSERMAN; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: INTERNATIONAL CONFERENCE

ON LEARNING REPRESENTATIONS 2015, San Diego, CA. **Anais**... San Diego, CA, 2015, p. 1–14.

SWGIT. **Recommendations and Guidelines for Using Closed-Circuit Television Security Systems in Commercial Institutions**. 2014. Disponível em: <https://www.swgit.org/>. Acesso em: 21 ago. 2018.

SZEGEDY, C. et al. Going Deeper with Convolutions. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2015, Boston. **Anais**... Boston, 2015, p. 1–9.

TADEU, E.; SEARA, R. **Codificação de vídeo h.264 – estudo de codificação mista de macroblocos.** Universidade Federal de Santa Catarina, Florianópolis, 2007.

TAN, X. et al. Face recognition from a single image per person: A survey. **Pattern Recognition**, v. 39, n. 9, p. 1725–1745, 2006.

TIA, T. I. A. **TIA Standards.** 2018. Disponível em: <https://www.tiaonline.org/>. Acesso em: 1 mar. 2018.

TRIANTAFYLLIDOU, D.; NOUSI, P.; TEFAS, A. Fast Deep Convolutional Face Detection in the Wild Exploiting Hard Sample Mining. **Big Data Research**, v. 11, p. 65–76, 2018.

TRIGO, T. **Equipamento Fotográfico, teoria e prática**. 5ª edição, São Paulo, Senac, 2012.

TURK. Face recognition using eigenfaces. In: PROCEEDINGS OF IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 1991, Maui. **Anais**... Maui, 1991, p.586–591.

VIOLA, P.; JONES, M. Rapid Object Detection Using A Boosted Cascade of Simple Features. In: PROCEEDINGS OF THE 2001 IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2001, Kauai. **Anais**... Kauai, 2001, p. 511–518.

VIOLA, P.; JONES, M. Robust real-time face detection. **International journal of computer vision**, v. 57, n. 2, p. 137–154, 2004.

VIOLA, P.; JONES, M. J.; SNOW, D. Detecting pedestrians using patterns of motion and appearance. **International Journal of Computer Vision**, v. 63, n. 2, p. 153–161, 2005.

WAGNER, A. et al. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 34, n. 2, p. 372–386, 2012.

WITTEN, I. H.; FRANK, E.; HALL, M. a. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. Burlington, Morgan Kaufman, 2011.

WOLF, L.; HASSNER, T.; MAOZ, I. Face recognition in unconstrained videos with matched background similarity. In: PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2011, **Anais...** Colorado, 2011, p. 529–534.

WU, B.; NEVATIA, R. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. **International Journal of Computer Vision**, v. 75, n. 2, p. 247–266, 2007.

YANG, H. et al. Recent advances and trends in visual tracking: A review. **Neurocomputing**, v. 74, n. 18, p. 3823–3831, 2011.

YANG, M. et al. Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person. **Pattern Recognition**, v. 66, p. 117–128, 2017.

ZAFEIRIOU, S.; ZHANG, C.; ZHANG, Z. A survey on face detection in the wild: Past, present and future. **Computer Vision and Image Understanding**, v. 138, p. 1–24, 2015.

ZENG, X. et al. Deep Learning of Scene-Specific Classifier for Pedestrian Detection. In: PROCEEDINGS OF THE EUROPEAN CONFERENCE ON COMPUTER VISION 2014, Zurich. **Anais**... Zurich, 2014, p. 472–487.

ZHANG, T. **Exploring The Frontier of Smart Video Surveillance : Novel Domains and Fine-Grain Event Understanding**. 2017. University of Queensland, Queenslad, 2017.

ZHAO, L.; THORPE, C. E. Stereo- and Neural Network-Based Pedestrian Detection. **IEEE Transactions on Intelligent Transportation Systems**, v. 1, n. 3, p. 148–154, 2000.

ZHU, X.; RAMANAN, D. Face detection, pose estimation, and landmark localization in the wild. In: PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2012, Providence. **Anais**... Providence, 2012, 2879-2886.

APÊNDICE A – RESUMO DA ARQUITETURA DE *HARDWARE*

Tabela A.1 – Características das lentes das câmeras.

Descrição	Valor
Distância focal	3,6 mm
Abertura máxima	F2.0
Ângulo de visão	H: 56,82° / V: 35,07°
Tipo de lente	Fixa

Fonte: autor

Tabela A.2 – Características de vídeo das câmeras.

Descrição	Valor
Resolução de imagem	720p (1.280 × 720) / 16:9
Compressão de vídeo	H.264H
Throughput	RTP: 24 Mbps
Taxa de *bit*	7 kbps a 2626 kbps

Fonte: autor

Tabela A.3 – Características gerais das câmeras.

Descrição	Valor
Sensor de imagem	1/4″ 1 megapixel
Iluminação mínima	0,1 lux: colorido
Controle de ganho	Desejável
Compensação de luz de fundo	Desejável
Modos de vídeo	Colorido/PB/IR
Temperatura de operação	-10 °C ~ +50 °C
Umidade relativa	<90%
Nível de proteção mínimo	IP66
Alimentação	100 ~ 240 V

Fonte: autor

Tabela A.4 – Características de rede das câmeras.

Descrição	Valor
Interface	RJ45 (10/100BASE-T)
Protocolos e serviços mínimos	TCP/IP, IPv4, IPv6, DHCP, DNS, RTSP

Fonte: autor

Tabela A.5 – Características do *switch.*

Descrição	Valor
Padrão *Ethernet*	802.3 100BASE-TX
Interface	RJ45 (10/100BASE-T)
Protocolos e serviços mínimos	TCP/IP, IPv4, IPv6, DHCP, DNS, RTSP
Alimentação	100 ~ 240 V

Fonte: autor

Tabela A.6 – Características do cabo de rede.

Descrição	Valor
Categoria	CAT5 E
Interface	RJ45 (10/100BASE-T)

Fonte: autor

Tabela A.7 – Características dos adaptadores de rede.

Descrição	Valor
Injector	Entradas com Plugue RJ45 macho e energia P4 fêmea com saída RJ45 fêmea
Splitter	Entrada RJ45 fêmea com saídas RJ45 macho + energia P4 macho
Nível de proteção mínimo	IP66
Suporte para alimentação	5 V, 12 V, 24 V, 48 V

Fonte: autor

Tabela A.8 – Características do computador.

Descrição	Valor
Porta *Ethernet*	802.3 100BASE-TX
CPU	2.50GHz
Memória	4GB

Fonte: autor

APÊNDICE B – RESUMO DAS CARACTERÍSTICAS DO AMBIENTE

Tabela B.1 – Características do Ambiente.

Descrição	Valor
Altura das câmeras	2250 mm
Ângulo Vertical das câmeras	72,46°
Distância mínima de captura	1148,03 mm
Distância máxima de captura	4732,64 mm
Largura Mínima de visão	1241,99 mm
Largura Máxima de visão	5120 mm
Altura Máxima de visão	2250 mm
Iluminação mínima	110 lux
Iluminação máxima	1080 lux
Iluminação Ideal	500 Lux

Fonte: autor

APÊNDICE C – GUIA DE UTILIZAÇÃO DO PROTÓTIPO DO SISTEMA

O protótipo do sistema é composto por uma tela principal, dividida em abas, onde cada aba contém uma subtela do sistema.

A tela principal também contém uma barra de status, na parte de baixo da tela, que exibe informações relevantes, como mensagens de sucesso, de erro, ou de início e fim de processos.

A ordem dos cadastros requeridos pelo sistema seguem a ordem em que estão posicionadas as subtelas, da esquerda para a direita. Desta forma, inicialmente deve ser configurados os parâmetros gerais na tela "*General Configuration*", em seguida, um experimento na tela "*Experiment*" e assim por diante.

Com exceção das telas "*General Configuration*" e "*Experiment*", todas as outras telas manipulam Entidades que são ligadas a um experimento. Um experimento deve ser selecionado na tela "*Experiments*", para que seja possível utilizar essas telas. Nesses casos, o experimento selecionado é exibido no canto superior esquerdo de cada subtela.

Todas as telas que manipulam entidades, como Câmeras, Pessoas, Treinamentos e Registros de Passagens, possuem um padrão de criação e exclusão dessas entidades.

Para criar uma Entidade, deve-se clicar no botão "*New*", entrar com os parâmetros obrigatórios e, opcionalmente, os não obrigatórios. Em seguida, clica-se no botão "*Save*". A Entidade criada é exibida na lista do lado esquerdo da tela, e pode ser selecionada, para que os campos referentes a ela sejam carregados.

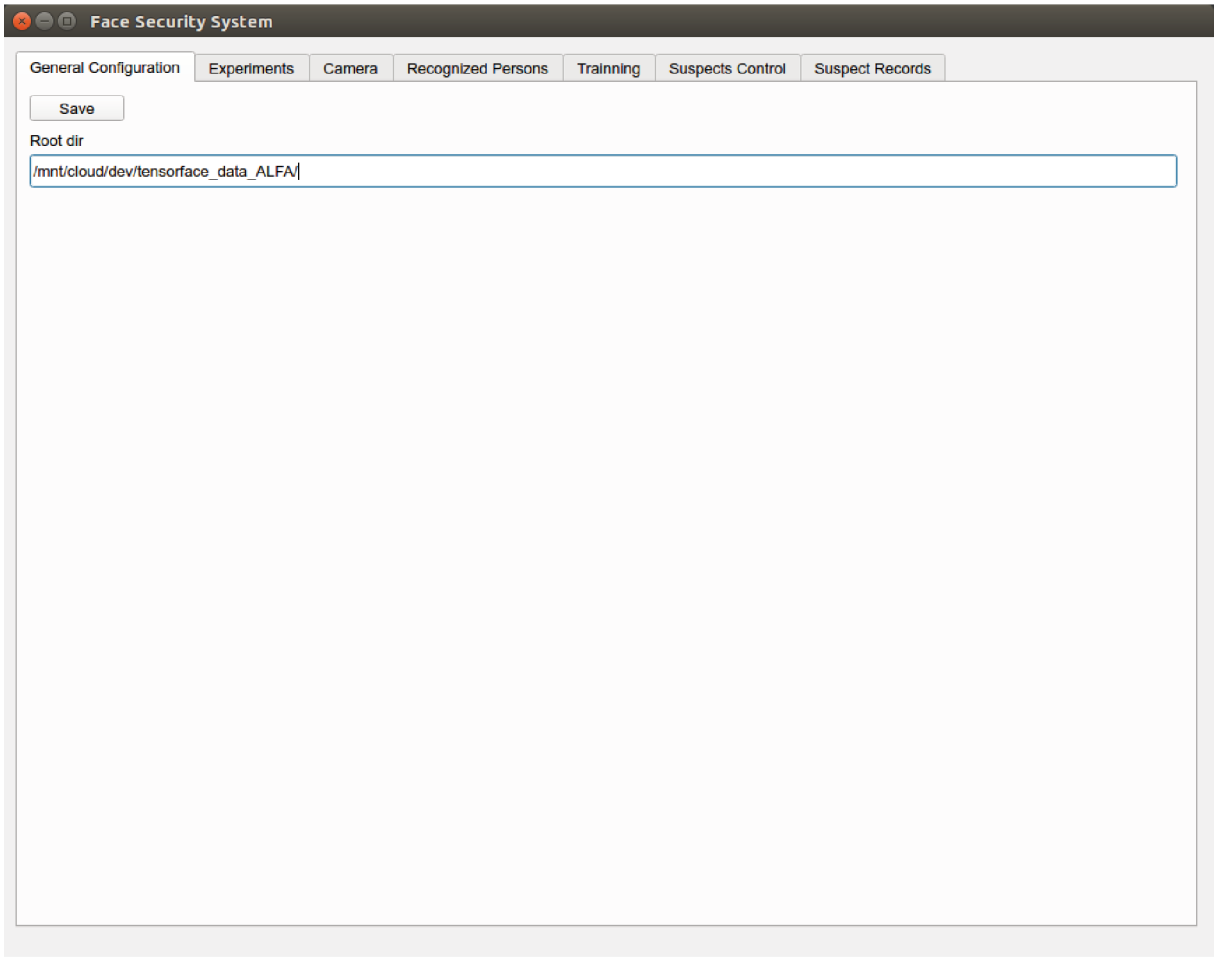
Para alterar uma entidade, altera-se os campos desejados e clica-se no botão "*Save*". Para excluir uma Entidade, seleciona-se a linha correspondente na lista e clica-se no botão "*Delete*".

CONFIGURAÇÕES GERAIS

A tela "Configurações Gerais" ("*General Configuration*" – Figura C.1) possibilita configurar, até o presente trabalho, apenas um parâmetro geral do sistema:

a) “*Root dir*” (obrigatório): diretório raiz do sistema, onde são gravadas todas as informações dos experimentos. Cada experimento é configurado em um subdiretório dentro deste.

Figura C.1 – Tela “Configuração Geral”. Permite configurar o caminho inicial da aplicação



Fonte: Autor

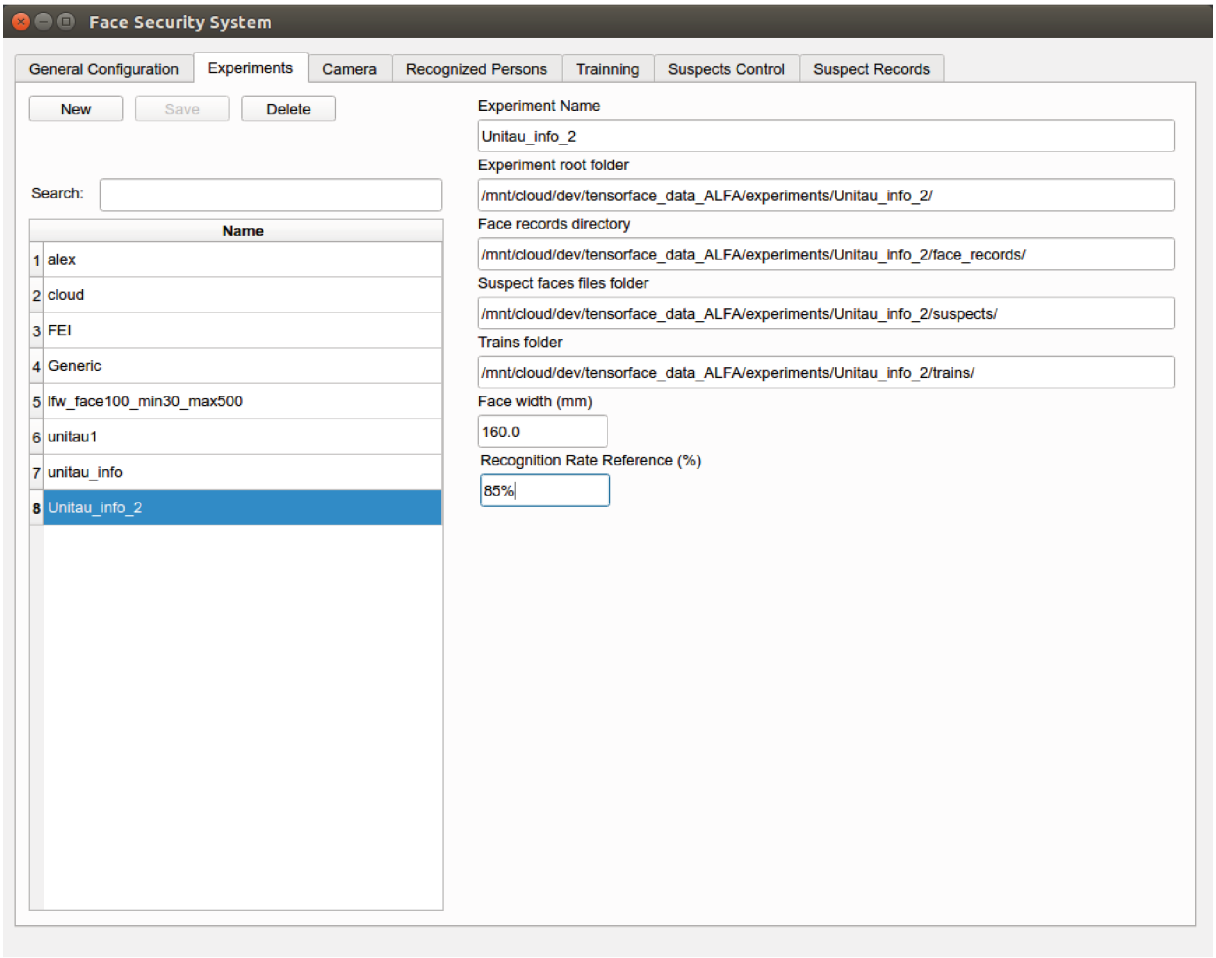
EXPERIMENTOS

A tela “Experimentos” (“*Experiments*” – Figura C.2) permite criar, modificar e excluir os Experimentos, que devem ser configurados para configurar um ambiente, no qual são configuradas as Câmeras, Pessoas e Registros de Suspeitos, que estão relacionados com o Experimento selecionado nesta tela.

Os parâmetros da tela estão listados abaixo:

a) “*Experiment Name*” (obrigatório): nome do experimento, também utilizado como diretório raiz do experimento, criado a partir do diretório raiz da aplicação, configurado na tela “*General Configuration*”;

Figura C.2 – Tela “Experimentos”. Permite configurar os diretórios do experimento, a largura de cabeça, utilizada no cálculo da resolução da face (*pixels*/mm), e a TRF de referência.



Fonte: Autor

b) “*Experiment root folder*” (automático): diretório raiz do experimento criado;

c) “*Face Records folder*” (automático): diretório onde são gravadas as imagens de faces de Pessoas conhecidas;

d) “*Suspect faces files folder*” (automático): diretório onde são gravadas as imagens de faces de Suspeitos;

e) “*Trains folders*” (automático): diretório onde são armazenadas as informações dos treinamentos;

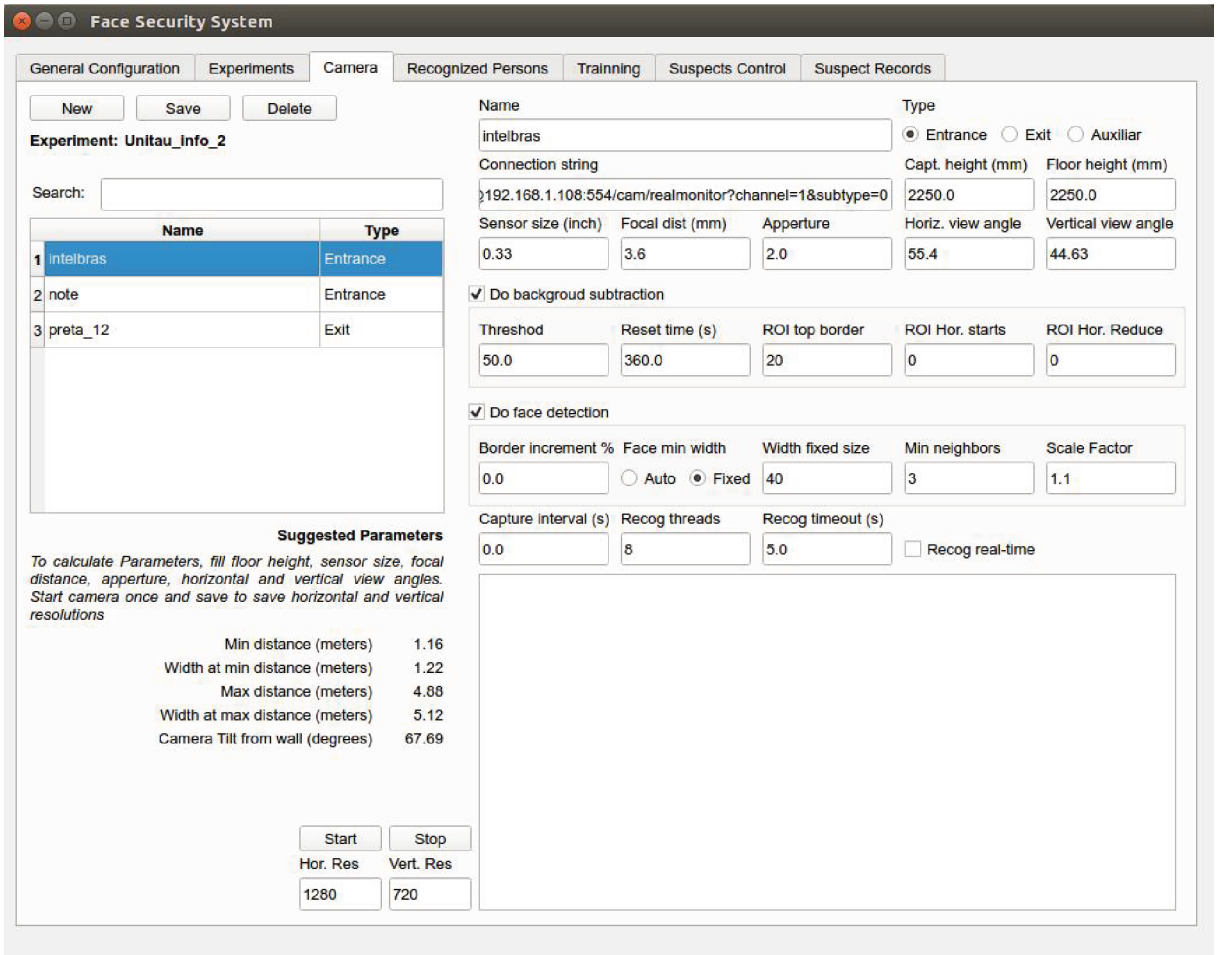
f) “*Face Width*” (obrigatório; padrão: 160 mm): a largura da face, em milímetros, utilizada para o cálculo do dimensionamento do ambiente;

g) “*Recognition Rate Reference”* (obrigatório; unidade: %; padrão: 85): TRF mínima que deve ser atingida, calculada em uma passagem, para considerar um suspeito como uma Pessoa conhecida.

CAMERA

A tela “Câmera” (*Camera* – Figura C.3) permite configurar a maioria dos parâmetros do sistema, que podem ser configurados individualmente, para cada câmera.

Figura C.3 – Tela “Câmera”: principais configurações do sistema, que podem ser feitas por câmera



Fonte: autor

Os seguintes parâmetros podem ser configurados:

a) “*Name*” (obrigatório): descrição da câmera;

b) “*Type*” (obrigatório): determina se uma câmera é de entrada, de saída, ou uma câmera auxiliar do ambiente, que não registra se ocorreu uma entrada ou saída do ambiente. Os registros de passagem armazenam as informações de entrada e saída do ambiente, utilizando essa configuração para indicar se um suspeito está entrando ou saindo do ambiente;

c) "*Connection String*" (obrigatório): *string* de conexão da câmera, informada no manual da câmera;
d) "*Capt. Height*" (unidade: mm): altura da câmera, que foi instalada no ambiente;
e) "*Floor Height*" (unidade: mm): altura máxima de captura;
f) "*Sensor Size*" (unidade: polegadas): tamanho da diagonal do sensor digital, em polegadas, informado no manual da câmera;
g) "*Focal dist*" (unidade: mm): distância focal da lente da câmera, informada no manual da câmera;
h) "*Apperture*": abertura da câmera, informada no manual da câmera;
i) "*Horiz. view angle*" (unidade: graus): ângulo horizontal de visão da câmera, informado no manual da câmera;
j) "*Vert. view angle*" (unidade: graus): ângulo vertical de visão da câmera, informado no manual da câmera;
k) "*Do background subtraction*": ativa ou desativa a Subtração de Fundo antes de detectar faces. Fortemente recomendado para ganho de performance;
l) "*Threshold"* (obrigatório): valor entre 0 e 255, que indica o limiar de sensibilidade da subtração de fundo. Valores maiores indicam uma menor sensibilidade;
m) "*Reset time"* (obrigatório; unidade: segundos): intervalo de tempo em que a imagem utilizada como fundo será recapturada, para ser utilizada como referência para a subtração de fundo. A recaptura de tempos em tempos é aconselhada devido à variações na luminosidade que podem ocorrer com o tempo;
n) "*ROI top border"* (unidade: *pixels*; padrão: 20): quantidade de *pixels* acima da cabeça que deve ser considerada após a delimitação da *RI*. Aconselhável a utilização de um valor maior ou igual ao padrão, para se ter a garantia de que a cabeça será enquadrada por inteiro;
o) "*ROI Horiz. starts*" (unidade: *pixels*; padrão: 0): quantidade de *pixels* do lado esquerdo da imagem, a partir da qual a detecção de movimento atuará. Útil para se remover uma faixa do lado esquerdo da imagem, que não será considerada, ou que possa conter objetos que se movem com o vento;
p) "*ROI Horiz. Reduce"* (unidade: *pixels*; padrão: 0): quantidade de *pixels* do lado direito da imagem, que será descartada pela detecção de movimento. Útil

para se remover uma faixa à direita da imagem, que não será considerada, ou que possa conter objetos que se movem com o vento;

q) "*Do face detection*" (padrão: ativado): ativa ou desativa a DF. Se o parâmetro for desabilitado, o sistema não realizará a DF e, consequentemente, não executará o RF;

r) "*Border increment*" (unidade: pixels; padrão: 0.0): seleciona uma área maior ou menor do que a área capturada pelo detector padrão, onde um incremento (valor positivo) ou decremento (valor negativo) sobre a área, em *pixels*, é definido pelo valor deste parâmetro. O parâmetro pode ser utilizado para se fazer estudos de diferentes tamanhos de recortes sobre a face, de forma que a identificação da face possa ser feitas em áreas centrais da face ou em áreas que podem envolver a cabeça inteira;

s) "*Face min width*" (unidade: *pixels*: padrão: fixo, 40): largura mínima de uma face para que seja detectada. Pode-se utilizar um valor fixo, como padrão, ou o valor que é utilizado automaticamente pelo detector da bibloteca OpenCV (não informado na documentação);

t) "*Min. Neighbors*" (padrão: 5): quantidade mínima de faces, encontradas nas vizinhanças de uma face detectada durante a varredura da imagem, para que a face detectada seja considerada uma detecção correta;

u) "*Scale Factor*" (padrão: 1.1): fator que determina quanto a imagem é reduzida em cada escala de imagem; Um valor menor pode identificar uma maior quantidade de faces vizinhas;

v) "*Capture interval*" (unidade: segundos; padrão: 0): intervalo de tempo entre cada captura de imagem;

w) "*Recog. Threads*" (padrão: 4): quantidade de processos paralelos que irão ser executados durante o reconhecimento. O paralelismo de reconhecimento é especialmente útil quando se ativa o parâmetro "*Recog real-time*";

x) "*Recog. timeout*" (unidade: segundos; padrão: 3): o início de uma passagem inicia uma contabilização do tempo de passagem do indivíduo. Após o tempo em segundos, determinado por esse parâmetro, as faces param de ser detectadas durante a passagem e o registro de passagem é criado;

y) "*Recog real-time*" (padrão: desativado): quando ativado, reconhece as faces em tempo de passagem, exibindo o nome do suspeito, de cada face reconhecida, na tela de Controle de Suspeitos, em "tempo real". Quando o

parâmetro está desativado, as faces são reconhecidas somente ao final da passagem, possibilitando salvar uma grande quantidade de faces antes do reconhecimento, que foram detectadas durante a passagem, pois a utilização do processo de detecção sem o reconhecimento leva apenas alguns milésimos de segundo. Nesse caso, o processo de reconhecimento no final da passagem pode levar um tempo maior, devido a uma grande quantidade de faces detectadas. Logo, deve-se ajustar o parâmetro "*Capture Interval*" para controlar a quantidade de faces detectadas.

PESSOAS RECONHECIDAS

A tela "Pessoas Reconhecidas" ("*Recognized Persons*" – Figura C.4) possibilita cadastrar pessoas e registrar suas faces manualmente, através do registro controlado por câmera ou através da importação de imagens de faces. Os suspeitos não identificados também são exibidos nesta tela, onde também é possível convertê-los para pessoas reconhecidas. Nesta tela, também é possível realizar a exclusão manual de faces.

Após a criação de uma Pessoa conhecida, um *nickname* é automaticamente gerado para a pessoa e é exibido no campo "*Nickname".*

Para registrar as faces de uma pessoa, através da câmera, em um modo controlado, quando a pessoa que está sendo registrada é uma pessoa conhecida, seleciona-se a linha da pessoa correspondente, na lista ao lado esquerdo, e clica-se no botão "*Start Rec Faces*".

Ao pressionar o botão, faces serão capturadas através da câmera que foi selecionada ao lado esquerdo do botão, a qual terá suas imagens exibidas na tela logo abaixo do botão "*Start Rec Faces*". Para parar de registrar as faces, clica-se em "*Stop Rec. Faces*".

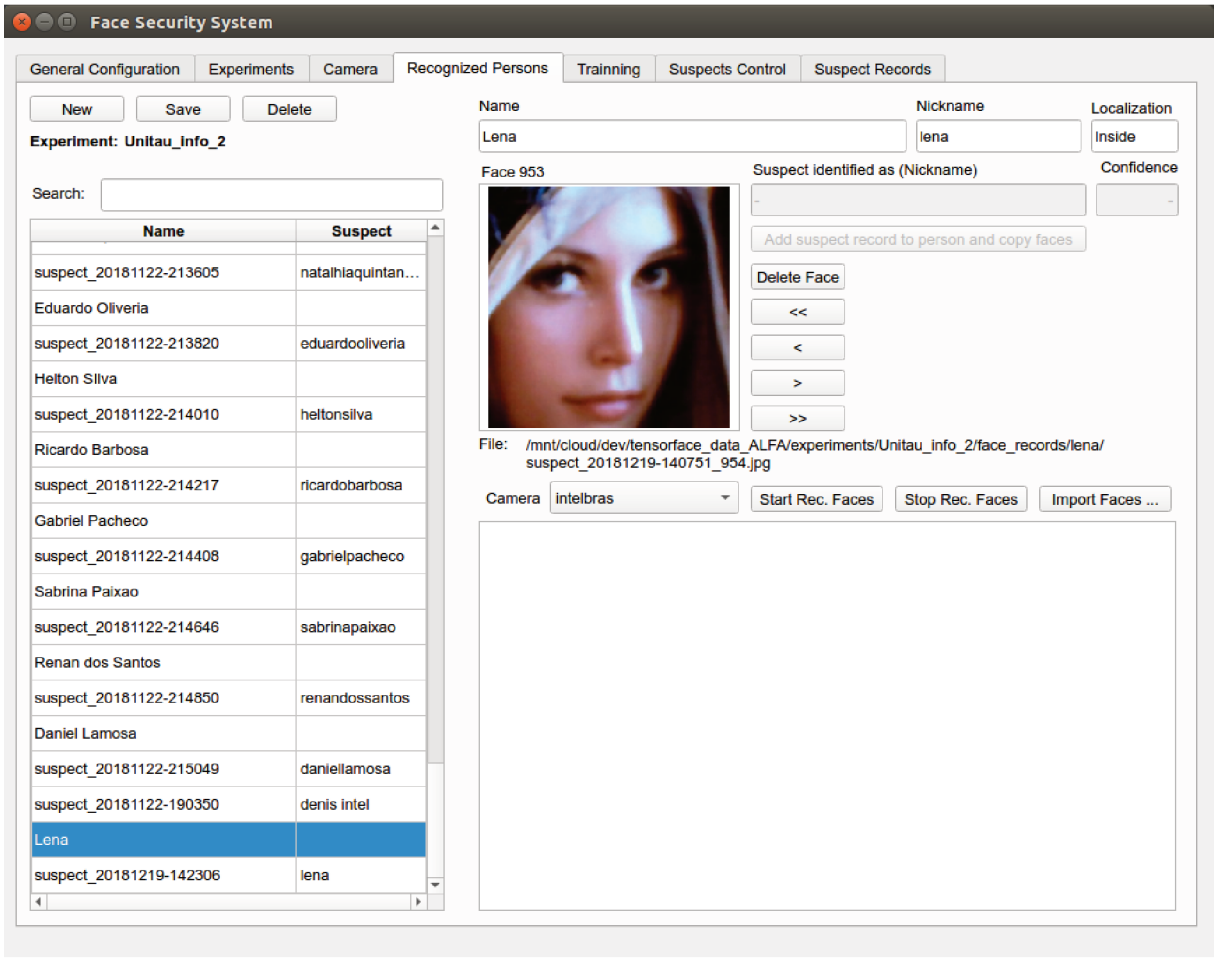
Após finalizada a captura de faces, é possível visualizá-las e excluí-las manualmente. Para excluir manualmente as faces de uma pessoa ou suspeito, deve-se encontrar a face desejada, através dos botões com setas "<<", "<", ">", ">>", que servem para encontrar, respectivamente, a primeira face, a face anterior, a face seguinte e a última face e, em seguida, clicar no botão "*Delete Face".*

Para importar faces, que estão previamente armazenadas em um diretório, seleciona-se a linha da pessoa correspondente, na lista ao lado esquerdo, clica-se

no botão "*Import faces*" e seleciona-se o diretório que contém as faces a serem importadas para a pessoa selecionada.

Para converter um suspeito, que foi identificado com uma baixa TRF, em uma pessoa conhecida, este deverá ser selecionado na lista ao lado esquerdo, o campo "*Suspect identified as (Nickname)*" deverá ser preenchido com o *nickname* da pessoa reconhecida e, em seguida, o botão "*Add suspect records to person and copy faces*" deve ser pressionado. Após o pressionamento do botão, todas as faces e registros do suspeito serão movimentados para a pessoa reconhecida, e o registro do suspeito será excluído.

Figura C.4 – Tela "Pessoas Reconhecidas". Permite registrar pessoas, manipular as imagens e migrar as imagens de registros de suspeitos para Registros de Pessoas conhecidas.



Fonte: autor

Dois campos informativos são exibidos na tela:

a) "*Localization*": indica a localização em que o suspeito se encontra. Se o último Registro de Passagem foi feito através de uma câmera de entrada, sua

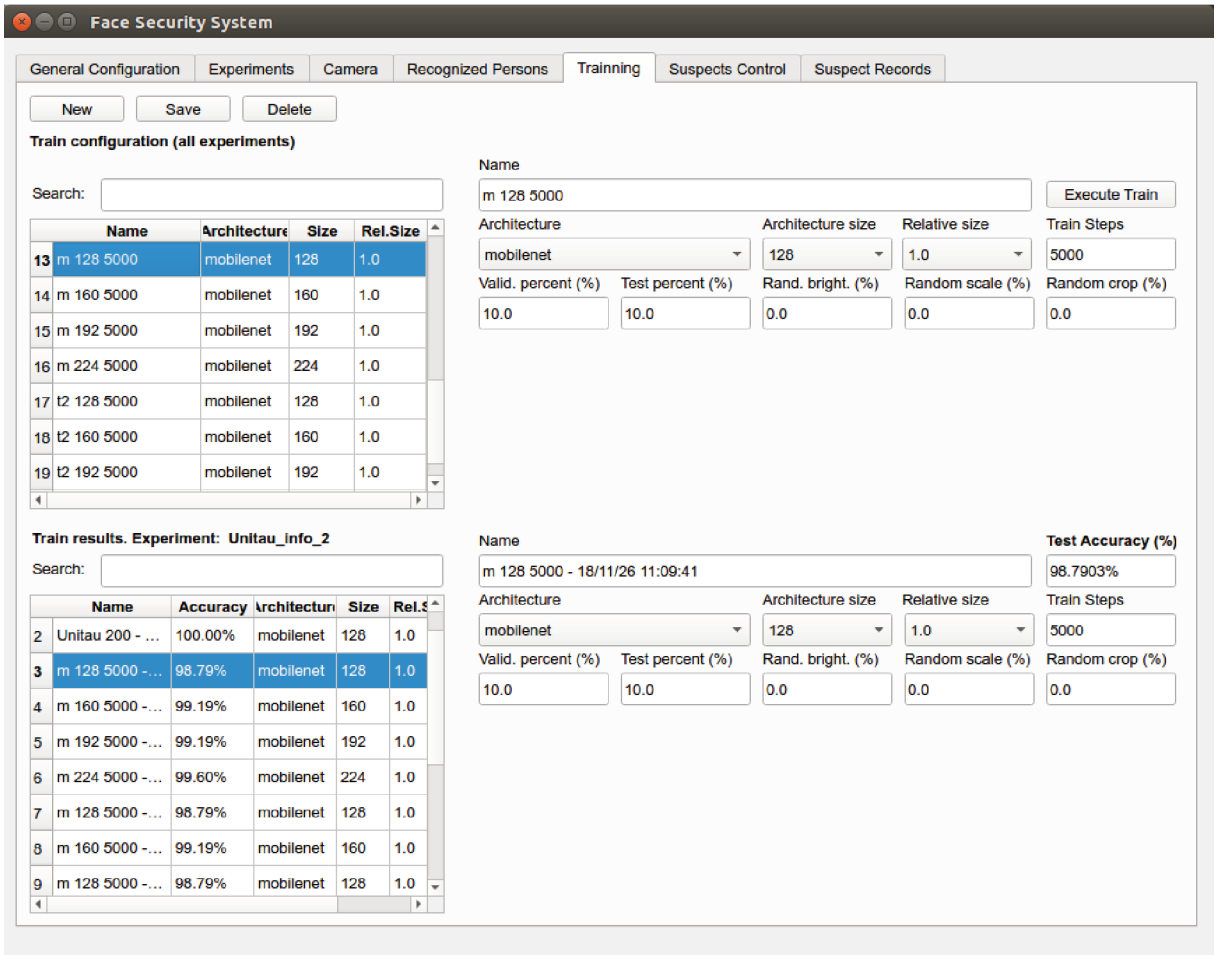
localização será exibida como "*Inside*". Analogamente, se o último Registro de Passagem foi realizado através de uma câmera de saída, sua localização será exibida como "*Outside*". Se o último Registro de Passagem foi realizado através de uma câmera auxiliar, o valor exibido será referente à última passagem que foi registrada em uma câmera de entrada ou de saída;

b) "*Confidence*": em casos de suspeitos, exibe a TRF de seu último Registro de Passagem;

TREINAMENTO: CLASSIFICAÇÃO DAS FACES

A tela "Treinamento" ("*Training*" – Figura C.5) permite realizar a classificação das faces que estão registradas em pessoa conhecidas. Ao contrário de outras telas, os registros de Configuração de Treinamento não são específicos de experimentos.

Figura C.5 – Tela "Treinamento": permite gerenciar Configurações de Treinamento e executar a classificação de faces, que gera um Resultado de Treinamento.



Fonte: autor

A partir de uma Configuração de Treinamento, um Resultado de Treinamento é gerado, quando uma Configuração de Treinamento é selecionada e o botão "*Execute Train*" é pressionado. Os Resultados de Treinamento são gerados para o experimento sendo utilizado, de forma que possa ser utilizado para o reconhecimento desse experimento.

As configurações de Treinamento são listadas na parte superior e os Resultados de Treinamento gerados são armazenados na parte inferior da tela.

Para criar uma Configuração de Treinamento, os campos abaixo devem ser preenchidos:

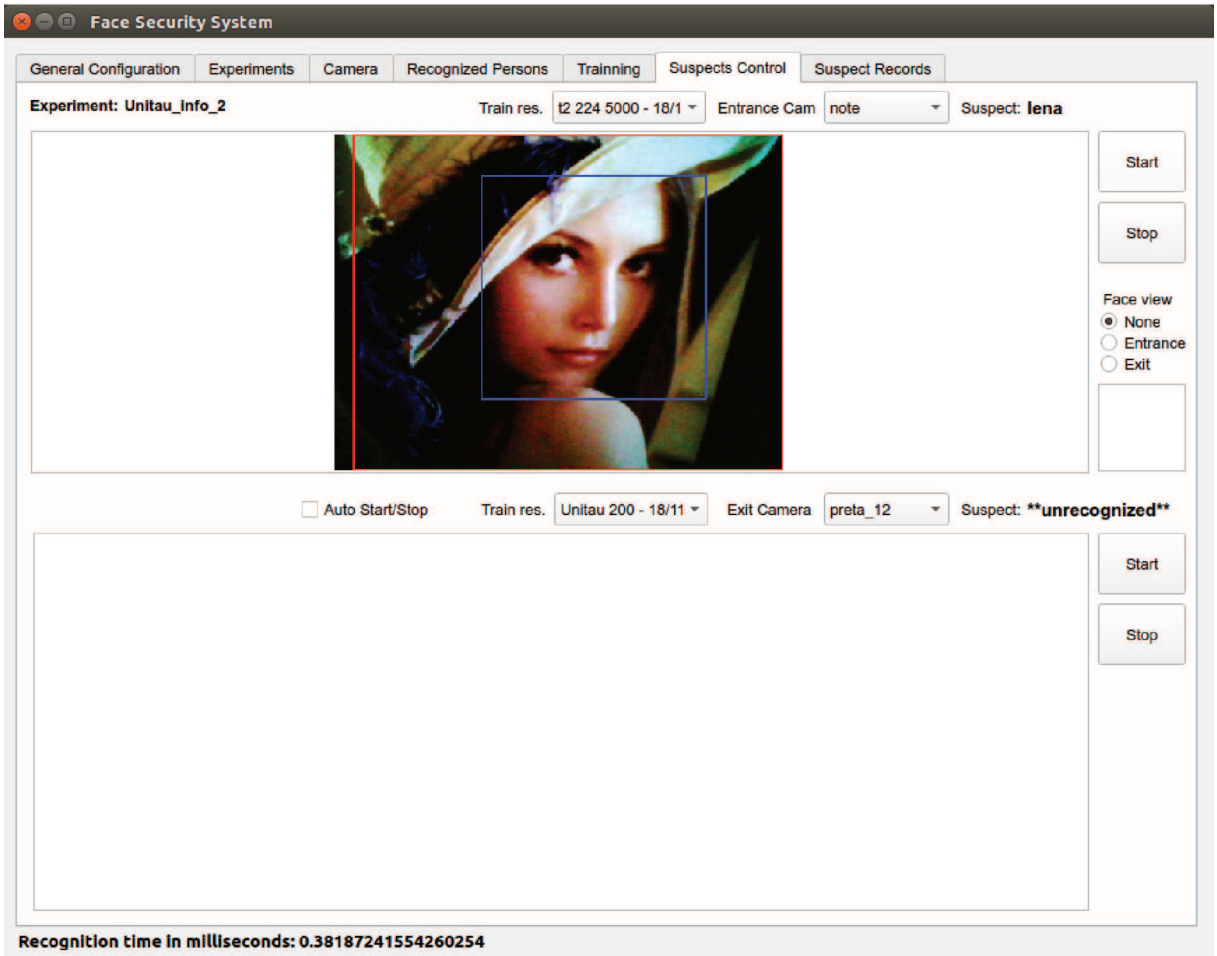
a) "*Architecture*" (obrigatório): arquitetura de rede. Até o momento, duas arquiteturas podem ser utilizadas: MobileNet e Inception. A arquitetura Inception não foi utilizada nos experimentos;
b) "*Architecture Size*" (obrigatório): tamanho da arquitetura, representado pelo parâmetro Multiplicador de Resolução MobileNet, quando selecionada a arquitetura Mobilenet;
c) "*Relative Size*" (obrigatório): tamanho relativo da arquitetura, representado pelo parâmetro Multiplicador de Largura MobileNet, quando selecionada a arquitetura Mobilenet;
d) "*Train Steps*" (unidade: passos de treinamento; obrigatório): quantidade de passos utilizados na execução do treinamento;
e) "*Valid. Percent*" (unidade: percentual; obrigatório): percentual da base que será utilizada para validação do treinamento (percentual da base testado, incluído no treinamento);
f) "*Test. Percent*" (unidade: percentual; obrigatório): percentual da base que será utilizada para teste do treinamento, isto é, o percentual da base que não estará incluído durante o treinamento;
g) "*Rand. Bright*" (unidade: percentual; opcional): porcentagem que determina quanto multiplicar aleatoriamente os *pixels* da imagem de treinamento para cima ou para baixo. A utilização deste parâmetro serve para realizar testes com simulações de variações na intensidade dos *pixels* das faces;
h) "*Random Scale*" (unidade: percentual; opcional): porcentagem que determina quanto aumentar ou diminuir aleatoriamente o tamanho das imagens de treinamento. A utilização deste parâmetro serve para realizar testes com simulações de variações dos tamanhos das imagens que contém faces;

i) "*Random Crop*" (unidade: percentual; opcional): porcentagem que determina o tamanho de uma margem, sobre as imagens de treinamento, que será removida da imagem, utilizando um percentual de aleatoriedade sobre o tamanho da margem. A utilização deste parâmetro serve para realizar testes com simulações de variações dos tamanhos de recorte das imagens;

CONTROLE DE SUSPEITOS

A tela "Controle de Suspeitos" (*Suspets Control* – Figura C.6) permite monitorar a captura das câmeras. Durante a captura, é exibido um retângulo vermelho que delimita a RI nas imagens capturadas, e um retângulo azul, que delimita a face capturada, como pode ser observado na Figura desta tela.

Figura C.6 – Tela: "Controle de Suspeitos". Permite selecionar o Resultado de Treinamento, utilizado no reconhecimento, e as câmeras de entrada e saída do ambiente.



Fonte: autor

Para iniciar a captura e o reconhecimento facial através de uma câmera de entrada, deve-se selecionar o Resultado de Treinamento e a câmera de entrada, na parte de cima da tela e, em seguida, clica-se no botão "*Start*". O mesmo procedimento deve ser realizado para o registro através da câmera de saída, mas utilizando os controles da parte de baixo da tela. Para parar o reconhecimento, o botão "*Stop*" deve ser pressionado.

A partir do início da captura, as faces detectadas começam a ser armazenadas em um Registro de Passagem. Ao final da passagem de uma pessoa pelo ambiente, o nome da pessoa reconhecida é inserida no Registro de Passagem e uma TRF é atribuída ao registro.

Quando o suspeito é reconhecido com uma TRF menor do que a taxa de referência cadastrada no Experimento em questão, um Registro de Passagem de pessoa anônima é criado. Em caso contrário, um Registro de Passagem de pessoa conhecida é criado.

Pode-se visualizar a face sendo detectada de uma das câmeras, em uma pequena tela ao lado direito da tela de Controle de Suspeitos. Para visualizar as faces detectadas na câmera de entrada, seleciona-se a opção "*Entrance*". Para visualizar as faces detectadas na câmera de saída, seleciona-se a opção "*Exit*". Ou se pode selecionar a opção "*None*", para não visualizar as faces.

O suspeito que está sendo reconhecido terá seu nome exibido nos campos "*Suspect*" de cada câmera, quando a opção "*Recog. Real Time*" é selecionada para a câmera, indicando o reconhecimento de cada face. O valor "***unrecognized***" será exibido quando a face não é reconhecida.

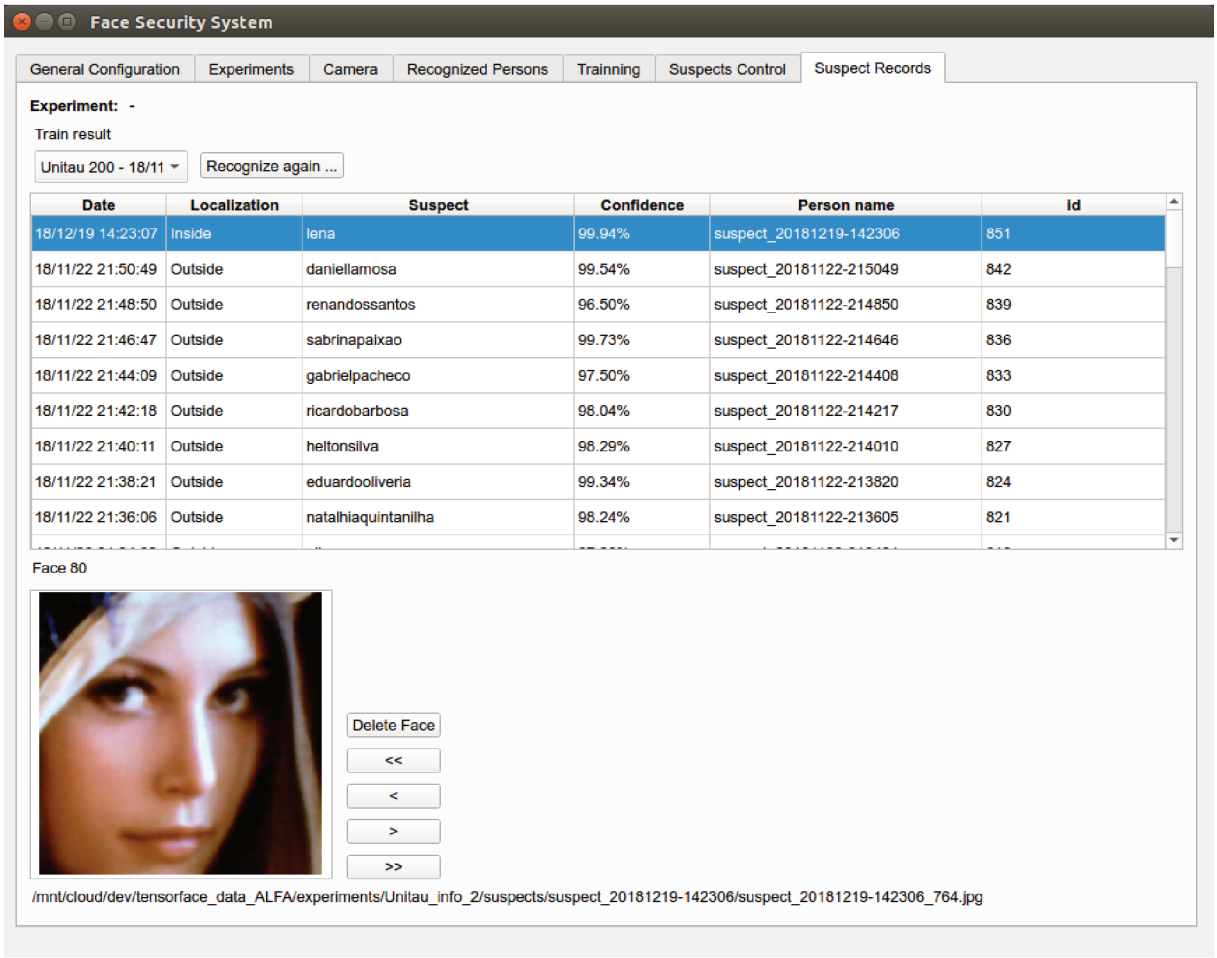
Durante o reconhecimento, é exibido na barra de status, o tempo de reconhecimento que foi necessário para reconhecer cada face.

REGISTROS DE SUSPEITOS

Os Registros de Passagens criados podem ser visualizados na tela "Registros de Suspeitos" ("*Suspects Records*" – Figura C.7). Esta tela permite remover faces de suspeitos, através dos botões com setas "<<", "<", ">", ">>", que servem para encontrar, respectivamente, a primeira face, a face anterior, a face seguinte e a última face e, em seguida, clicar no botão "*Delete Face*".

A tela também permite selecionar um ou mais Registros de Passagens, através do clique do mouse selecionando a tecla “Ctrl”, para serem reconhecidos novamente, utilizando um Resultado de Treinamento diferente. Para isso, deve-se selecionar o Resultado de Treinamento desejado no campo “*Train Result*” e o botão “*Recognize again*” deve ser pressionado.

Figura C.7 – Tela “Registros de Suspeitos”. Permite visualizar os Registros de Passagens e gerenciar suas faces. Também permite realizar o reconhecimento de todos os registros novamente, utilizando um Resultado de Treinamento diferente.



Fonte: autor

A lista com os Registros de Passagens exibe as seguintes colunas:

a) “*Date*”: exibe a data e hora de criação do Registro de Passagem;
b) “*Localization*”: exibe a localização da pessoa que passou pelo ambiente.
c) “*Suspect*”: exibe o *nickname* da Pessoa conhecida, com a qual o suspeito que foi identificado mais se assemelha;
d) “*Confidence*”: exibe a TRF com a qual o suspeito foi identificado;

e) "*Person Name*": nome da Pessoa conhecida que foi identificada, a qual serão incorporadas as faces registradas durante a passagem, caso a TRF da passagem atinja a taxa mínima de reconhecimento cadastrada no experimento. Caso a taxa mínima não seja atingida, exibe um nome de suspeito, gerado automaticamente. Este nome deve ser utilizado para a identificação do suspeito na tela de Pessoas conhecidas e para que suas faces possam sejam transferidas manualmente para uma pessoa conhecida;

ANEXO A – ARQUITETURA MOBILENET

Tabela – Arquitetura MobileNet.

Tipo / *Stride*	Formato do Filtro	Tamanho da entrada
Conv / s2	3×3×3×32	224×224×3
Conv dw / s1	3×3×32 dw	112×112×32
Conv / s1	1×1×32×64	112×112×32
Conv dw / s2	3×3×64 dw	112×112×64
Conv / s1	1×1×64×128	56×56×64
Conv dw / s1	3×3×128 dw	56×56×128
Conv / s1	1×1×128×128	56×56×128
Conv dw / s2	3×3×128 dw	56×56×128
Conv / s1	1×1×128×256	28×28×128
Conv dw / s1	3×3×256 dw	28×28×256
Conv / s1	1×1×256×256	28×28×256
Conv dw / s2	3×3×256 dw	28×28×256
Conv / s1	1×1×256×256	14×14×256
5x Conv dw / s1 Conv / s1	3×3×512 dw 1×1×512×512	14×14×512 14×14×512
Conv dw / s2	3×3×512 dw	14×14×512
Conv / s1	1×1×512×1024	7×7×512
Conv dw / s2	3×3×1024 dw	7×7×1024
Conv / s1	1×1×1024×1024	7×7×1024
Avg Pool / s1	Pool 7×7	7×7×1024
TC / s1	1024×1000	1×1×1024
Softmax / s1	Classifier	1×1×1000

Fonte: (HOWARD et al., 2017a)