

INSTITUTO FEDERAL DO ESPÍRITO SANTO

ENGENHARIA DE CONTROLE E AUTOMAÇÃO

CONRADO COSTA

**ENSAIOS DE DETECÇÃO FACIAL EM DISPOSITIVOS DE BORDA,
PARA MONITORAMENTO E CONTROLE DE ACESSO EM ESPAÇOS
ABERTOS E INTELIGENTES**

SERRA – ES

2021

CONRADO COSTA

**ENSAIOS DE DETECÇÃO FACIAL EM DISPOSITIVOS DE BORDA,
PARA MONITORAMENTO E CONTROLE DE ACESSO EM ESPAÇOS
ABERTOS E INTELIGENTES**

Trabalho de conclusão de curso, apresentada como parte das atividades para obtenção do título de bacharel em Engenharia de Controle e Automação, do curso de Engenharia de Controle e Automação do Instituto Federal do Espírito Santo.

Orientador: Prof. Rafael Emerick Z de Oliveira

SERRA – ES

2021

Aqui entra a Ficha catalográfica.

Será gerada pela biblioteca!

CONRADO COSTA

**ENSAIOS DE DETECÇÃO FACIAL EM DISPOSITIVOS DE BORDA,
PARA MONITORAMENTO E CONTROLE DE ACESSO EM ESPAÇOS
ABERTOS E INTELIGENTES**

Texto submetido ao Curso de Graduação em Engenharia de Controle e Automação do Instituto Federal do Espírito Santo como requisito parcial para obtenção do título de Bacharel em Engenharia de Controle e Automação.

Aprovada em XX de XXXXX de XXXX.

COMISSÃO EXAMINADORA

Prof. Rafael Emerick Z de Oliveira
Instituto Federal do Espírito Santo - *campus Serra*

Prof. XXXXX
Instituto Federal do Espírito Santo - *campus Serra*

Prof. XXXXX
Instituto Federal do Espírito Santo - *campus Serra*

SERRA – ES

2021

DECLARAÇÃO DO AUTOR

Declaro, para fins de pesquisa acadêmica, didática e técnico-científica, que a presente Dissertação de Mestrado pode ser parcial ou totalmente utilizada desde que se faça referência à fonte e aos autores.

Conrado Costa

Serra, XX de XXXXX de XXXX.

Resumo

<a preencher>

Palavras-chave: Processamento Digital de Imagens; Inteligência Artificial; Rede Neural; Máquina de Vetor de Suporte;

Abstract

<a preencher>

Keywords: Digital Image Processing; Artificial Intelligence; Neural Network; Support Vector Machine.

Listas de ilustrações

Figura 1 – Fluxo de processo de reconhecimento de face.	19
Figura 2 – Características geométricas (destaque em branco), usada em experimentos de reconhecimento de faces.	20
Figura 3 – Arquitetura de computação distribuída	24
Figura 4 – Interface e seus parâmetros.	27
Figura 5 – Exemplo de resultado retornado.	28
Figura 6 – Exemplo de resultado com poucas faces detectadas.	31
Figura 7 – Exemplo de resultado com várias faces detectadas e alguns falsos positivos.	32
Figura 8 – Exemplo de resultado possivelmente satisfatório.	33
Figura 9 – Exemplo de matriz de resultado com limites ajustados.	34
Figura 10 – Exemplo de resultado com as métricas.	35
Figura 11 – Imagem selecionada para testes da cena 1.	37
Figura 12 – Exemplo de variação de cena com redução de 40 faces.	39
Figura 13 – Otimização Cena 1 - resolução 1440p.	45
Figura 14 – Otimização Cena 1 - resolução 1080p.	46
Figura 15 – Otimização Cena 1 - resolução 720p.	47

Sumário

Lista de ilustrações	6
Sumário	7
 1 INTRODUÇÃO	9
 1.1 Objetivo Geral	11
1.1.1 Objetivos Específico	11
 1.2 Estrutura do Texto	12
 2 REVISÃO TEÓRICA	13
 2.1 O problema da segurança em espaços públicos e privados	13
 2.2 Smart cities e videomonitoramento inteligente	14
 2.3 Redes Neurais e Deep Learning	16
 2.4 Processamento de imagens, detecção e reconhecimento de faces	17
 2.5 A Internet das Coisas, Edge e Fog Computing	22
 3 DESENVOLVIMENTO	25
 3.1 O algoritmo de detecção de objetos	25
 3.2 Ferramenta de parametrização, otimização e obtenção de resultados	26
 3.3 Cenários de testes	36
3.3.1 Cena 1	37
3.3.1.1 Variação de quantidade de faces	38
3.3.1.2 Variação de resolução da imagem	39
3.3.1.3 Tempo médio de captura	40
3.3.2 Cena 2	41
3.3.2.1 Captura das imagens para teste	41
3.3.2.2 Tempo médio de captura	42
3.3.2.3 Variações de resolução	42
3.3.3 Dispositivos testados	42

4	EXPERIMENTOS REALIZADOS	43
4.1	Cena 1	43
4.1.1	Otimização de parâmetros	43
4.1.1.1	Resolução 1440p	45
4.1.1.2	Resolução 1080p	46
4.1.1.3	Resolução 720p	47
5	CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS	48
	REFERÊNCIAS	49

1 INTRODUÇÃO

O problema de segurança pública sempre foi um problema sério no Brasil. Constantes crimes de furtos e roubos geram grande danos, principalmente financeiros, tanto para o indivíduo, no caso das vítimas, quanto para a sociedade (CERQUEIRA et al., 2007; G1, 2013). Um dos reflexos gerados por essa insegurança está no fato de que vários locais públicos devem ter seu acesso controlado e vigiado para garantir a segurança patrimonial.

As *smart cities* (cidades inteligentes) estão emergindo como uma prioridade para pesquisa e desenvolvimento em todo o mundo. Elas abrem oportunidades significativas em várias áreas, como crescimento econômico, saúde, bem-estar, eficiência energética e transporte, para promover o desenvolvimento sustentável das cidades (SONG et al., 2017). O conceito e oportunidades das *smart cities* são escaláveis para outros conceitos 'smart' como o *smart room*, *smart home*, *smart building*, etc (PACHECO et al., 2018). A proliferação de tecnologias de informação e comunicação possibilita o desenvolvimento de diversos serviços inteligentes. E um dos serviços comunitários mais essenciais é justamente a vigilância inteligente (CHEN et al., 2016; NIKOUEI et al., 2018).

Nos últimos anos, aplicações de reconhecimento facial a partir de imagens geradas por câmeras de videomonitoramento têm ganhado relevância, sendo largamente utilizadas para a verificação ou identificação de indivíduos em locais públicos. No âmbito da segurança, o reconhecimento facial em vídeo permite agilidade nas situações em que muitos indivíduos devem ser identificados rapidamente (QUIRITA, 2014).

Apesar de estarmos longe de conseguir com que a IA (inteligência artificial) se aproxime da performance humana, em algumas áreas, como reconhecimento de imagem, carros autônomos e jogos eletrônicos, ela se mostra equivalente, ou até mesmo superior (AGGARWAL, 2018). Em tarefas de visão computacional é possível rastrear o movimento de uma pessoa em um plano de fundo complexo. E com moderado sucesso, é possível tentar localizar e nomear todas as pessoas em uma fotografia,

através da detecção e reconhecimento de faces, roupas e cabelos (Szeliski, 2011).

O *deep learning* (aprendizagem profunda), ramo do *machine learning* (aprendizagem de máquina), tornou-se imensamente popular no reconhecimento de imagens, bem como em outras tarefas de reconhecimento e correspondência de padrões (Verhelst; Moons, 2017).

As redes neurais artificiais simulam o sistema nervoso humano com base em aprendizado de máquina, tratando as unidades computacionais em um modelo de aprendizado de maneira semelhante aos neurônios humanos. Não é uma tarefa fácil pois o poder computacional do computador mais rápido atualmente equivale a uma pequena fração do poder computacional de um cérebro humano (Aggarwal, 2018).

As *deep neural networks* (redes neurais profundas) envolvem uma complexidade computacional significativa, fazendo com que, até recentemente, seu processamento fosse viável apenas em plataformas de potentes servidores disponíveis na ‘nuvem’ (Verhelst; Moons, 2017). Quando é necessário armazenamento e computação de dados em larga escala, a computação em nuvem tem sido a solução. Porém, com o grande crescimento de dispositivos móveis e inteligentes, juntamente com as tecnologia de IoT (Internet das Coisas), o foco mudou para se obter respostas em tempo real. (Dolui; Datta, 2017).

Nos últimos anos, vê-se uma tendência de se incorporar o processamento de aprendizado profundo em dispositivos de borda, como celulares, dispositivos móveis e nos nós da IoT. Isso torna possível a análise de dados localmente, em tempo real, além de mitigar problemas de privacidade dos dados (Verhelst; Moons, 2017). Outro benefício da computação na borda (*Edge Computing*) é o descongestionamento da rede de dados, pois permite que o processamento seja feito próximo das fontes dos dados (Merenda; Porcaro; Iero, 2020). Assim, evita-se a comunicação desnecessária, que sobrecarrega não só a rede principal como também o datacenter na nuvem (Azam; Hu, 2014).

1.1 Objetivo Geral

Com este trabalho objetiva-se testar o desempenho de dispositivos de borda no processo de detecção de faces, avaliando sua capacidade de detecção e tempo de resposta para diferentes cenários, com possíveis aplicações de monitoramento inteligente e controle de acesso. Com os resultados obtidos, espera-se determinar, para cada cenário definido, se o dispositivo é capaz de processar de forma satisfatória a etapa de detecção de faces, e as vantagens de se realizar esse processo na borda, em uma arquitetura de processamento distribuído.

1.1.1 Objetivos Específico

Em um sentido mais estrito, pretende-se

- Definir diferentes cenários (com aplicabilidade para monitoramento inteligente e controle de acesso) e os requisitos a serem cumpridos, como tempos de respostas e capacidade de reconhecimento. Serão utilizadas imagens estáticas que representem cada cenário para os testes.
- Desenvolver uma ferramenta cliente-servidor para auxiliar na parametrização do algoritmo de detecção, buscando melhor otimização para cada cena e dispositivo, e na obtenção das métricas de desempenho.
- Analisar os resultados obtidos e determinar para quais cenários os dispositivos podem executar a detecção de face de forma satisfatória e quais os ganhos em se executar tal processamento na borda, principalmente no que se refere à utilização de banda na rede de um sistema com arquitetura de processamento distribuído, sua escalabilidade e tempos de resposta.

1.2 Estrutura do Texto

<a preencher>

2 REVISÃO TEÓRICA

2.1 O problema da segurança em espaços públicos e privados

O problema de segurança pública no Brasil é algo que está sempre em evidência. O Programa das Nações Unidas para o Desenvolvimento (Pnud), em seu relatório divulgado em 12/11/2013, constata que o Brasil apresentou a maior taxa de roubo da América Latina, segundo dados de 2011 repassados pelos países. Os dados apontam que para cada 100 mil habitantes no Brasil, há 572,7 ocorrências de roubo. E sabe-se que, na realidade, esse número tende a ser maior, tendo em vista que nem todos os roubos são reportados às autoridades (G1, 2013).

Estima-se um total de 15 milhões de ocorrências de roubos e furtos no Brasil no ano de 2003, incluído os casos que não foram notificados. E como parte da consequência, estima-se uma perda material de R\$ 8,4 bilhões (CERQUEIRA et al., 2007). Se corrigido para o ano de 2020 com base no IPCA (Índice de Preços ao Consumidor Amplo), esse valor seria de aproximadamente R\$ 20,2 bilhões. Esse cálculo foi feito com base em uma calculadora online disponível no site do Banco Central do Brasil (BCB). Esse é um problema que gera não só perdas para as vítimas dos roubos e furtos, mas também indiretamente para outros indivíduos da sociedade, uma vez que, essa transferência de valor pode ser considerada como recurso de oportunidade a serem aplicados no setor de crimes. Este, por sua vez, demanda recurso público para o seu combate, tornando o problema um causador não só de dano ao indivíduo, como também de custo social (ANDERSON, 1999; CERQUEIRA et al., 2007).

Os fatores que motivam ou favorecem esse tipo de crime são vários. A teoria das janelas quebradas, testada em um experimento por Philip Zimbardo, psicólogo de Stanford, propõe-se a explicar um deles. O experimento consistiu em deixar dois carros similares abandonados nas ruas de dois bairros diferentes de Nova Iorque, um nobre e outro na periferia. O que se observou foi quem o carro deixado na periferia foi atacado por vândalos nos primeiros 10 minutos, enquanto o segundo carro, deixado

no bairro nobre, ficou intocado por mais de uma semana. Então Zimbardo com uma marreta danificou parte do carro e o que se observou em seguida foi que várias pessoas que estavam transitando se juntaram ao carro, que, em poucas horas, estava completamente destruído. O que o experimento transmite é que a desordem e o crime estão de certa forma ligados. A desordem passa uma impressão de descuido, de forma que um indivíduo mal intencionado se sentirá muito mais à vontade em cometer algum delito, por ter a sensação de que ninguém irá notar, ou se importar. E por isso, há também uma sensação de impunidade. Uma propriedade mal cuidada se torna ideal para os que saem na intenção de vandalizar ou saquear, e até mesmo para aqueles quem nem pensariam em tais atitudes, mas as cometem ao enxergar o delito como uma oportunidade (WILSON; KELLING, 1982).

Sistemas de câmeras de vigilância têm sido cada vez mais implantados em muitos lugares, como prédios, ruas, instalações industriais e comerciais, escolas, shoppings, aeroportos e residências, provendo, segurança pública, monitoramento de ambientes internos, monitoramento de tráfego e proteção de infraestrutura (PUVVADI et al., 2015).

Em uma das formas de monitoramento, as imagens da câmera são monitoradas em tempo real por seguranças. Como outra forma, é possível registrar a saída de cada câmera no um gravador (VCR), para futura análise. Porém, na primeira forma de monitoramento, uma ocorrência ou incidente de segurança pode acabar não sendo verificado, devido, por exemplo, a uma falha humana. E no segundo caso, o momento da verificação do ocorrido pode não acontecer em um tempo satisfatório (OLSON, 2006; Ramos Lima; Marques Ciarelli, 2019). Mas hoje, com o avanço da tecnologia, temos opções mais inteligentes de monitoramento, como é tratado no item 1.2.

2.2 *Smart cities* e videomonitoramento inteligente

“Smart City” é um poderoso paradigma que aplica as mais avançadas tecnologias de comunicação aos ambientes urbanos, com o objetivo de melhorar a qualidade de vida nas cidades e fornecer um amplo conjunto de serviços de valor tanto para os cidadãos quanto à administração (CENEDESE et al., 2014). O recente conceito de Smart Cities, impulsionado pelo rápido crescimento da IoT (Internet das Coisas), atraiu

a atenção de planejadores urbanos e pesquisadores para aumentar a segurança e o bem-estar dos residentes. A proliferação de tecnologias de informação e comunicação conecta sistemas ciber-físicos e entidades sociais, bem como possibilita muitos sistemas inteligentes. Um dos serviços comunitários inteligentes mais essenciais é a vigilância inteligente (CHEN et al., 2016; NIKOUEI et al., 2018).

Nos últimos tempos, os sistemas de câmeras de vigilância evoluíram de simples aquisição de vídeo e sistemas de exibição para sistemas semiautônomos inteligentes, capazes de realizar procedimentos complexos. Hoje em dia, um sistema de vigilância por vídeo pode integrar alguns dos algoritmos de análise de imagem e vídeo mais sofisticados como de classificação (por exemplo, redes neurais), reconhecimento de padrões, tomada de decisão, aprimoramento de imagem e vários outros (TSAKANIKAS; DABIUKLAS, 2018). Isso permite uma grande possibilidade de aplicações como controle de acesso em áreas de interesse, reconhecimento de faces humanas, detecção de padrões e objetos, reconhecimento de comportamento, estatísticas de fluxo de multidões, análise de congestionamento, etc (HU et al., 2004).

Um sistema de vigilância moderno compreende não só dispositivos de aquisição de imagem e vídeo para exibição, mas também dispositivos para processamento de dados e unidades de armazenamento, componentes cruciais para a execução da tarefa (TSAKANIKAS; DABIUKLAS, 2018). Além disso, têm estado cada vez mais disponíveis dispositivos de vigilância com conectividade de rede que suportam o protocolo IP. Isso abre uma ainda maior gama de possibilidades já que os dados podem ser enviados praticamente para qualquer de equipamento onde quer que esteja localizado. Porém, isso traz junto a preocupação com a privacidade, pois os dados de imagem trafegados em rede estão sujeitos a interceptações. Uma abordagem utilizada é a criptografia dos dados em tráfego, porém isso traz uma carga maior de processamento que pode acabar prejudicando a performance do monitoramento em tempo real (PUVVADI et al., 2015).

Muitas das aplicações de videomonitoramento inteligente requerem recursos computacionais e de armazenamento significativos, para ser capaz de lidar com a grande quantidade de dados gerada pelos sensores de vídeo. De acordo com o estudo recente, os dados de vídeo dominam o tráfego em tempo real e criam uma carga de trabalho

pesada nas redes de comunicação. Por exemplo, vídeo online responde por 74% de todo o tráfego online em 2017 e 78% do tráfego móvel será de dados de vídeo em 2021. O volume de dados são cada vez maiores, à medida que se têm maiores taxas de quadro e maiores resoluções (PORTER; FRASER; HUSH, 2010).

O paradigma da computação em nuvem, ou Cloud Computing, oferece excelente flexibilidade para lidar com essa grande quantidade de transferência de dados, além de também de ser escalável, correspondendo ao número crescente de câmeras de vigilância (NIKOUEI et al., 2018). Para tarefas de vigilância urbana que requerem a combinação de dados complexos, a computação em nuvem têm sido amplamente aceita como a solução (CHEN et al., 2016). No entanto, existem obstáculos significativos para a arquitetura de vigilância inteligente baseada em nuvem remota (NIKOUEI et al., 2018). Os delay adicional devido à comunicação em rede pode não ser tolerável em aplicações que sejam sensíveis a latências mais altas, como as aplicações de tempo real (CHEN et al., 2016). Uma grande distância entre o sensor de vídeo e os servidores da nuvem, além dos possíveis congestionamentos na rede, torna ainda mais inviável essa abordagem.

2.3 Redes Neurais e Deep Learning

Deep Learning, ou aprendizado profundo, é um campo do aprendizado de máquina baseado nas em redes neurais artificiais (BROWNLEE, 2019).

“Redes neurais artificiais são técnicas populares de aprendizado de máquina que simulam o mecanismo de aprendizado em organismos biológicos. O sistema nervoso humano contém células, conhecidas como neurônios. Os neurônios são conectados uns aos outros com o uso de axônios e dendritos, e as regiões de conexão entre os axônios e dendritos são chamadas de sinapses” (AGGARWAL, 2018).

Na programação convencional, o programador, através de várias linhas precisas de código, determina quais tarefas que o computador deve executar, à risca. Grandes problemas são quebrados em problemas menores nos quais os computadores conseguem performar. Já no paradigma de redes neurais, o programador não precisa

dizer exatamente o que a máquina deve fazer. Ao invés disso, a partir de algoritmos de aprendizado e dados observacionais, a máquina consegue aprender por si só como resolver determinado problema (NEAPOLITAN, 2018).

Existem várias diferentes arquiteturas de redes neurais que são comumente usadas em diferentes aplicações, como as *Restricted Boltzmann Machines* (RBM), as *Recurrent Neural Networks* (RNN) e as *Convolutional Neural Networks* (CNN). As mais utilizadas atualmente são as CNN e as RNN. As RNN, ou redes neurais recorrentes, são projetadas para trabalhar com dados sequenciais, como frases de texto, séries temporais e outras sequências discretas, como sequências biológicas. As CNN, ou redes neurais convolucionais, são redes inspiradas biologicamente e são usadas em visão computacional para classificação de imagens e detecção de objetos. (AGGARWAL, 2018).

As redes neurais convolucionais têm se mostrado a mais bem-sucedidas de todos os tipos de redes neurais. São amplamente usadas para o reconhecimento de imagem, detecção de objetos, rastreamento e até mesmo processamento de texto. O desempenho dessas redes chegou a superar o desempenho dos humanos no problema de classificação de imagens (HE et al., 2016).

Existem hoje disponíveis gratuitamente vários frameworks e estruturas de aprendizado de máquina incluindo redes neurais, tornando o trabalho de treinar uma rede neural mais simples. Entre as mais conhecidas estão Keras, PyTorch, TensorFlow, Scikit-learn. O treinamento de redes neurais exige muito poder de processamento. Os cálculos de deep learning tendem a ser mais rápidos quando feitos por GPUs (*Graphical Processing Units*) (HELLER, 2019).

2.4 Processamento de imagens, detecção e reconhecimento de faces

Pesquisadores têm desenvolvido na área de visão computacional técnicas matemáticas para recuperar a forma tridimensional e a aparência de objetos em imagens. Hoje, há técnicas confiáveis para calcular com precisão um modelo 3D parcial de um ambiente a partir de milhares de fotografias sobrepostas. A visão computacional

é usada hoje em uma ampla variedade de aplicativos do mundo real, que incluem reconhecimento óptico de caractere (OCR), segurança automotiva, videomonitoramento, reconhecimento de digitais, detecção e reconhecimento de faces, entre outros (Szeliski, 2011).

A detecção de faces é um problema já bem resolvido na área de visão computacional. Isso deve-se ao fato de que a detecção de faces é um dos processos mais utilizados em sistemas de videomonitoramentos e é exigido em vários tipos de aplicações como reconhecimento de faces, rastreamento, análise comportamental, etc (Zafeiriou; Zhang; Zhang, 2015).

“O objetivo da detecção de face é, em primeiro lugar, determinar se algum rosto está representado em uma cena e, em segundo lugar, calcular e retornar as coordenadas dos rostos detectados. Esta tarefa envolve muitas condições não triviais, como variações de escala, localização, orientação e pose, bem como condições de iluminação, expressões faciais e oclusões” (TSAKANIKAS; DAGIUKLAS, 2018).

Entre várias técnicas detecção de face, o trabalho inovador de Viola e Jones (Viola and Jones, 2001) baseado em melhorar o processo de detecção de faces, foi o primeiro algoritmo que tornou a detecção de rosto praticamente viável em aplicações do mundo real. Até hoje é amplamente aplicada em câmeras digitais e softwares de organização de fotos (Zafeiriou; Zhang; Zhang, 2015). A abordagem proposta por Viola e Jones para detecção de objetos minimiza o tempo de processamento ao mesmo tempo em que consegue grande acurácia na detecção. E quando aplicado na detecção de faces, se mostrou 15 vezes mais rápido que qualquer abordagem precedente (VIOLA; JONES, 2001). A biblioteca OpenCV, multiplataforma e de uso totalmente livre (OpenCV), disponibiliza uma função de detecção de objeto baseada no método proposto por Viola e Jones.

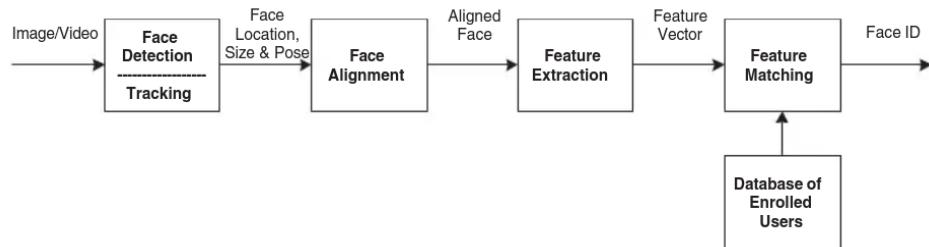
O reconhecimento de faces constitui o problema de identificar a face de uma pessoa mediante a comparação com uma base de dados de inúmeras outras faces previamente identificadas, a partir da qual obtém-se o grau de similaridade entre cada comparação (Quirita, 2014). O problema de reconhecimento de face requer que um rosto já tenha sido detectado em uma imagem, portanto, a detecção de face se torna um

pré-requisito para o processo de reconhecimento. Nos últimos tempos, os algoritmos de reconhecimento de faces evoluíram de tal forma que hoje podem ser usados em aplicativos do mundo real e em ambientes não controlados. (ZAFEIRIOU; ZHANG; ZHANG, 2015).

Uma das primeiras abordagens para reconhecimento de face baseia-se em localizar as características distintas da imagem, como olhos, nariz, boca, e medir a distância entre as posições de cada uma (FISCHLER; ELSCHLAGER, 1973; KANADE, 1977). Abordagens mais recentes baseiam-se na comparação de imagens em escala de cinza projetadas em subespaços dimensionais inferiores chamados de eigenfaces (Szeliski, 2011).

Um sistema de reconhecimento de faces, baseado em características, geralmente consiste de quatro partes: detecção, alinhamento, extração de características e combinação/verificação, conforme é esquematizado na figura 2.1 (Stan Z. Li, 2011).

Figura 1 – Fluxo de processo de reconhecimento de face



Fonte: (Stan Z. Li, 2011)

Na etapa de detecção “Face Detection”, temos um vídeo ou uma imagem como entrada. Essa etapa tem como função localizar e extrair uma ou mais faces que estejam presentes na imagem. No caso em que a entrada é um vídeo, se necessário, essa etapa também é responsável pelo rastreamento, “Tracking”, das faces, quadro a quadro. As faces extraídas passam então pela etapa de alinhamento, “Face Alignment”, que é responsável por normalizar a imagem em rotação e em escala, de forma que na imagem resultante a face esteja alinhada ao eixo horizontal do plano. (Stan Z. Li, 2011; QUIRITA, 2014).

Depois de já alinhada, a face passa pela etapa de extração de características, “Feature Extraction”. Essa etapa tem a função de identificar e extrair da face pontos

faciais distintos, como olhos, nariz, boca e outras marcas de referência que sejam suficientes para caracterizar a face, de forma que a mesma possa ser distinguida dentre as faces obtidas de outras pessoas. (Stan Z. Li, 2011). Em seguida são computadas as relações geométricas entre esses pontos faciais, reduzindo-se assim a imagem facial de entrada em um vetor de características geométricas (JAFRI; ARABNIA, 2009). A figura 2.2 destaca alguns pontos faciais característicos e as relações geométricas entre eles.

Os primeiros trabalhos realizados em reconhecimento de faces baseavam-se principalmente nessa técnica. Em uma das primeiras tentativas empregou-se um método simples de processamento de imagem para extrair um vetor de 16 parâmetros faciais, que eram relações de distâncias, áreas e ângulos, de forma a compensar o tamanho variável das imagens. Então usou-se uma medida de distância euclidiana simples para correspondência, chegando a um desempenho máximo de 75% em um banco de dados de 20 pessoas diferentes, usando 2 imagens por pessoa. (KANADE, 1977)

Figura 2 – Características geométricas (destaque em branco), usada em experimentos de reconhecimento de faces



Fonte: (Stan Z. Li, 2011)

Tendo as características geométricas da face calculadas e codificadas em um vetor, a próxima etapa para o reconhecimento da face é fazer a correspondência, ou “Feature Matching” (quarta e última etapa do processo de reconhecimento proposto na figura 1.2). Essa etapa consiste em comparar o vetor obtido da face que se deseja reconhecer com um banco de dados de vetores de outras faces que já são conhecidas e tiveram suas características extraídas pelo mesmo processo descrito até aqui. O reconhecimento é feito com base no grau de similaridades entre a face sendo verificada e as faces conhecidas (QUIRITA, 2014).

A principal vantagem oferecida pelas técnicas de reconhecimento baseadas em características é que, uma vez que a extração dos pontos característicos é anterior à análise feita para comparar a face com a de um indivíduo conhecido, esses métodos são relativamente robustos para variações de posição na imagem de entrada (JEBARA, 1996). A princípio, esquemas baseados em características podem ser invariáveis ao tamanho, orientação ou iluminação (COX; GHOSN; YANILOS, 1996). Outras vantagens de se utilizar técnicas baseadas em características faciais incluem a compactação de representação das imagens de face e a alta velocidade na tarefa de verificação de correspondência (identificação) (BRUNELLI; POGGIO, 1992).

A principal desvantagem dessa técnica é a dificuldade de detecção automática de características e o fato de que o desenvolvedor da aplicação deve tomar decisões arbitrárias sobre quais características são importantes para o processo de reconhecimento (CENDRILLON; LOVELL, 2000).

A biblioteca Dlib C++ disponibiliza algumas ferramentas para reconhecimento facial. Em seu site, está disponível gratuitamente um modelo pré-treinado para extração de características faciais, que alcançou uma marca de precisão de 99,38% no benchmark de reconhecimento facial da *Labeled Faces in the Wild* (LFW), que é comparável a outros métodos de última geração para reconhecimento facial em fevereiro de 2017. (DLIB...,). A LFW é uma referência pública para verificação de desempenho de reconhecimento facial (LABELLED...,).

Esse modelo mapeia a imagem de uma face humana em um vetor de 128 dimensões espaciais. Quando vetores de duas imagens diferentes são muito próximos, significa que a face tende a pertencer à mesma pessoa. Assim, o reconhecimento facial pode ser obtido mapeando-se várias faces em vetores de 128 dimensões e compará-las entre si calculando a distância Euclidiana (DLIB...,).

Esse modelo foi treinado a partir de um banco de cerca de 3 milhões de imagens. A precisão alcançada de 99,38% significa que, ao ser apresentada um par de imagens faciais, a ferramenta identificará corretamente se o par pertence à mesma pessoa ou é de pessoas diferentes em 99,38% das vezes (DLIB...,), o que o torna uma boa ferramenta para reconhecimento de faces de uma forma geral.

O grande avanço nas técnicas e ferramentas de aprendizado de máquina abriu um grande leque de possibilidades de aplicações de inteligência artificial nos últimos anos. Devido a isso, os recentes modelos de detecção e reconhecimento, não só de faces, mas de objetos em geral, estão cada vez mais precisos e rápidos. Isso, aliado ao expressivo avanço da Internet das Coisas (IoT), trazendo à tona o conceito de Edge Computing, torna cada vez mais viável as implementações de videomonitoramento automático.

2.5 A Internet das Coisas, Edge e Fog Computing

A Internet das Coisas, ou Internet of Things (IoT), é um paradigma de comunicação recente que prevê em um futuro próximo que os objetos da vida cotidiana serão equipados com microcontroladores, transceptores para comunicação digital e pilhas de protocolo adequadas que os tornarão capazes de se comunicarem entre si e com os usuários finais, tornando-se parte integrante da Internet (ATZORI; IERA; MORABITO, 2010).

O conceito de IoT, portanto, visa tornar a Internet ainda mais imersiva e abrangente. Além disso, ao permitir fácil acesso e interação com uma ampla variedade de dispositivos, como, por exemplo, eletrodomésticos, câmeras de vigilância, sensores de monitoramento, atuadores, monitores, veículos e assim por diante, a IoT promoverá o desenvolvimento de uma série de aplicações que fazem uso da quantidade e enorme variedade de dados gerados por tais objetos para fornecer novos serviços aos cidadãos, indústrias e administrações públicas. Esse paradigma encontra aplicação em muitos domínios diferentes, como automação residencial, automação industrial, assistência médica, saúde móvel, assistência a idosos, gerenciamento de energia inteligente e redes inteligentes, automotivo, gerenciamento de tráfego e muitos outros (Bellavista et al., 2013).

Com o advento da Internet das coisas nós estamos em uma era onde haverá uma grande quantidade de dados gerados por coisas que estão imersas em nosso dia a dia. Conforme estimado pelo Cisco Global Cloud Index, em 2019, os dados produzidos por pessoas, máquinas e “coisas” chegaria a 500 zetabytes (??). Para processar esses

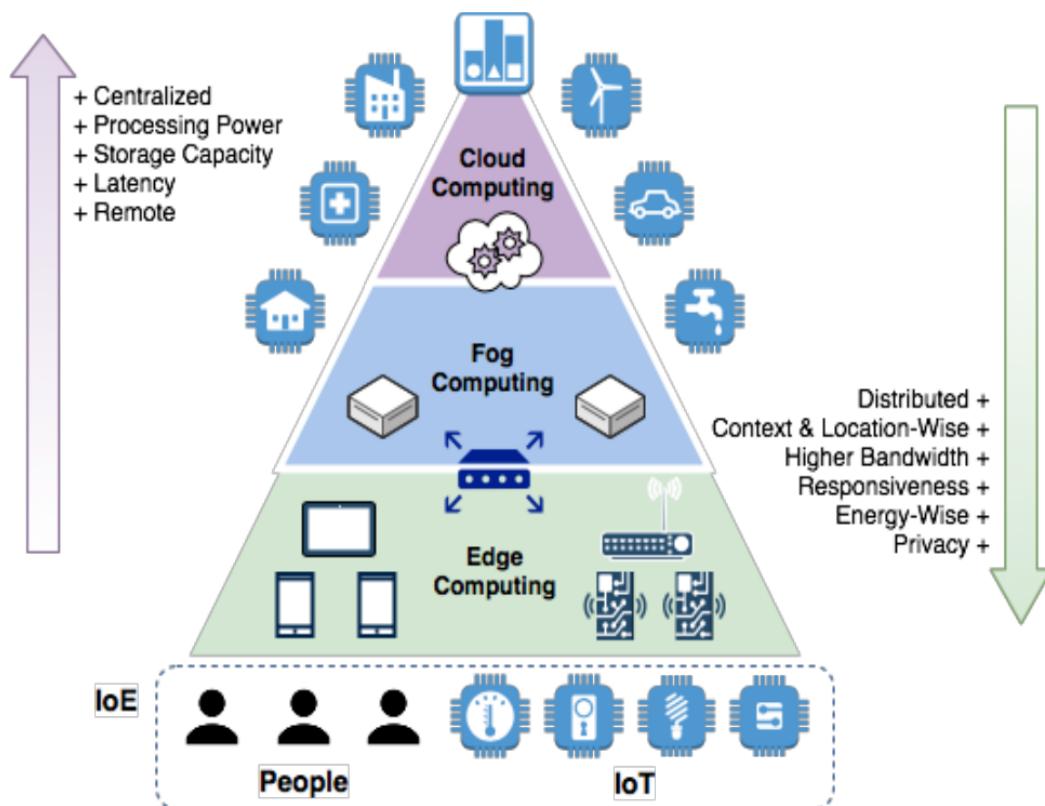
todos esses dados, é necessário muito poder computacional.

Hoje em dia, a computação em nuvem é uma plataforma econômica e prevalecente, oferecendo enorme poder de processamento e capacidade de armazenamento para treinamento de modelos de machine learning, reconhecimento facial, reconhecimento de fala, visão computacional, processamento automatizado de linguagem, classificação de texto e diversas aplicações de IoT e smart cities. No entanto, também tem algumas desvantagens importantes, como latência de resposta da rede e a segurança do sistema no que diz respeito a questões de privacidade, já que para chegar na nuvem os dados, possivelmente sensíveis, precisam trafegar por um longo caminho na rede mundial (PACHECO et al., 2018).

A maioria das ações de controle de IoT deve ser realizada em tempo real, portanto, o tempo de espera de processamento em nuvem, principalmente devido à latência da rede, não funciona bem para problemas de IoT (SINGH, 2017). Na tentativa de contornar algumas dessas limitações, aparecem alguns paradigmas recentes e complementares. São o Fog e o Edge Computing, que promete a capacidade de realizar tarefas de uma forma mais distribuída e responsiva, uma vez que os nós de IoT estão mais próximos das fontes de dados dos sensores. Além disso, também reduz o tráfego de rede e evita a exposição de dados privados do usuário (PACHECO et al., 2018).

A figura 2.3 demonstra muito bem a ideia de uma arquitetura de processamento distribuído, indicando as características que cada parte da rede dá à aplicação ao ser responsável por parte do processamento.

Figura 3 – Arquitetura de computação distribuída



Fonte: (PACHECO et al., 2018)

3 DESENVOLVIMENTO

3.1 O algoritmo de detecção de objetos

O algoritmo de detecção de objetos utilizado nesse estudo foi o *Haar cascade object detection*, proposto por Viola e Jones em sua pesquisa *Rapid Object Detection using a Boosted Cascade of Simple Features* (VIOLA; JONES, 2001).

<pending>explicar um pouco sobre o algoritmo (baseado na detecção com base em características, é treinado com imagens positivas e negativas, É rápido pois a classificação em camadas descarta rapidamente uma área em que não há face, etc).</pending>

Não é o algoritmo com melhor acurácia atualmente, se comparado com técnicas mais modernas que aplicam *deep learning*, porém é um algoritmo extremamente rápido e preciso, portanto ainda é muito útil para detecção de objetos em dispositivos com recursos limitados, como é o caso de dispositivos de borda, em geral.

Uma desvantagem nesse algoritmo é a tendência à detecção de falsos positivos e à necessidade de definição de alguns parâmetros, descritos resumidamente a seguir.

<pending>- parâmetros importantes (explicar cada) –scaleFactor Parameter specifying how much the image size is reduced at each image scale. –minNeighbors Parameter specifying how many neighbors each candidate rectangle should have to retain it. –minSize Minimum possible object size. Objects smaller than that are ignored. –maxSize Maximum possible object size. Objects larger than that are ignored. If maxSize == minSize model is evaluated on single scale.</pending>

O ajuste de parâmetros não é muito simples e depende do cenário da imagem, dos possíveis tamanhos de faces que se deseja detectar, etc. A escolha dos parâmetros influencia diretamente no resultado da detecção, na capacidade de detecção de determinados tamanhos de faces, na propensão em detectar falsos-positivos e no tempo de execução.

Devido a isso, viu-se necessário o desenvolvimento de uma ferramenta para se testar de uma só vez um range variável de valores de determinados parâmetros e verificar facilmente a qualidade e tempo de detecção para cada conjunto de valores testados.

3.2 Ferramenta de parametrização, otimização e obtenção de resultados

O *design* da ferramenta foi pensado de forma a facilitar a análise do resultado de diversas combinações de parâmetros ao mesmo tempo, de diferentes imagens e em dispositivos diferentes, através de uma interface web.

Contitui de duas partes: a parte cliente, uma interface web que pode ser executada em qualquer dispositivo através de um navegador, e a parte servidor, que deve rodar nos dispositivos que serão testados.

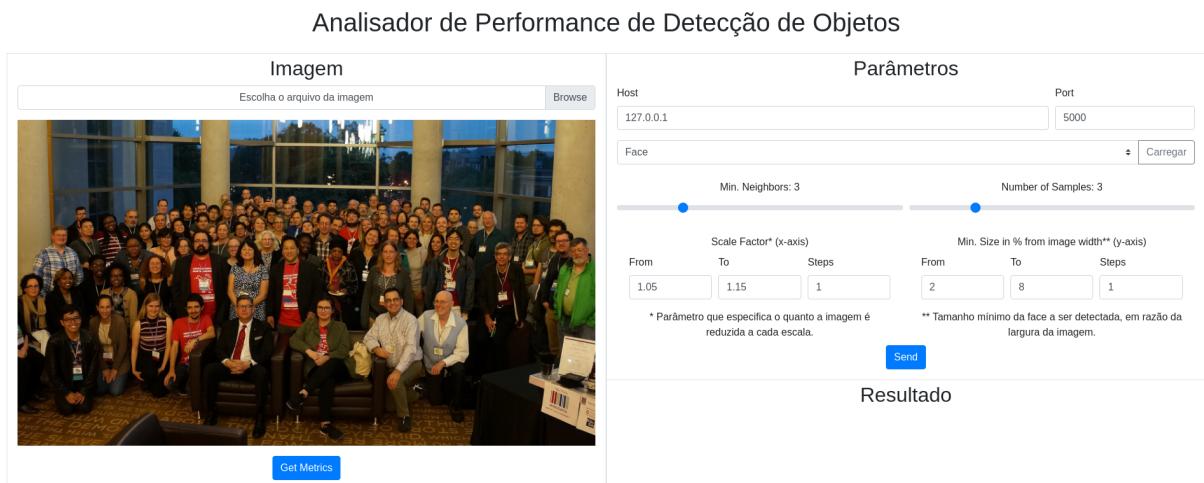
No cliente é feita a seleção da imagem a ser utilizada no teste, é definido o endereço do dispositivo a ser testado, rodando o servidor, e são definidos os parâmetros de testes. O cliente envia os dados para o servidor que executa o algoritmo de detecção conforme os parâmetros passados e retorna para o cliente o resultado de cada combinação de valores dos parâmetros solicitados. O resultado é composto por uma matriz contendo o tempo médio de execução, a quantidade de faces detectadas e a posição de cada face detectada para cada par de valores dos parâmetros na matriz. As faces detectadas são vizualizadas na imagem ao se selecionar um dos resultados, permitindo, assim, a verificação da qualidade da detecção e a presença de falsos-positivos.

Na figura a seguir temos um exemplo da interface com imagem e parâmetros selecionados, ainda sem exibição do resultado da análise.

<pending> fazer citação da imagem utilizada no exemplo</pending>

À esquerda da interface há um elemento de entrada que permite a seleção a imagem a ser utilizada no ensaio. Na imagem, exibida logo abaixo, serão delimitadas as faces detectadas.

Figura 4 – Interface e seus parâmetros.



Fonte: autor.

À direita há dois campos, "Host" e "Port", que permitem a seleção do dispositivo a ser testado, a partir do endereço IP e porta em que a aplicação estará "escutando".

Logo abaixo estão os parâmetros a serem definidos e enviados ao servidor para a execução do ensaio. Três deles, "Scale factor", "Min neighbors" e "Min size", são parâmetros passados na própria função *detectMultiScale* do OpenCV, e o significado de cada um foi descrito no início desta seção. O parâmetro "Number of Samples" determina quantas vezes o servidor executa cada combinação de parâmetros para obter um resultado médio.

Para o parâmetro "Min Neighbors", é possível definir apenas um valor para cada execução, que será utilizado em todas as combinações de parâmetros. Já para os parâmetros "Scale Factor" e "Min Size" é possível definir valores mínimos, máximos e as quantidades de valores ("Steps") a serem distribuídos linearmente nos intervalos definidos para cada um dos dois parâmetros. O servidor executará todas as combinações possíveis a partir dos conjuntos de valores delimitados, e retornará uma matriz com um resultado para cada combinação.

Por fim, ao se clicar no botão "Send", a imagem e os parâmetros são enviados para o servidor. O servidor executa o algoritmo de detecção com todas as combinações possíveis e retorna uma matriz de resultados contendo número de faces detectadas e tempo médio de execução.

Figura 5 – Exemplo de resultado retornado.

Resultado

All		Scale Factor				
Faces		1.050	1.087	1.125	1.163	1.200
Mean Time						
Min. Size (%)	0.5	F: 91 T: 3.157	F: 87 T: 1.675	F: 83 T: 1.254	F: 80 T: 0.963	F: 81 T: 0.831
	1.5	F: 90 T: 2.191	F: 87 T: 1.235	F: 83 T: 0.946	F: 79 T: 0.667	F: 81 T: 0.613
	2.5	F: 70 T: 1.249	F: 58 T: 0.656	F: 52 T: 0.457	F: 55 T: 0.41	F: 54 T: 0.342
	3.5	F: 23 T: 0.573	F: 15 T: 0.328	F: 14 T: 0.215	F: 20 T: 0.209	F: 13 T: 0.154
	4.5	F: 3 T: 0.347	F: 2 T: 0.179	F: 2 T: 0.132	F: 2 T: 0.111	F: 1 T: 0.11

Fonte: autor.

Como pode ser visto na figura 5, o resultado é exibido em forma de uma matriz. No eixo vertical têm-se a distribuição dos valores definidos para o parâmetros "Min. Size" e no eixo horizontal têm-se a distribuição dos valores definidos para o parâmetro "Scale Factor", de conforme os limites e quantidades de passos definidos para cada um.

Cada célula da matriz apresenta o resultado da detecção, utilizando a combinação de parâmetros corresponde, com base em duas métricas, a quantidade de faces detectadas, na parte superior da célula, e o tempos médio em segundos, na parte inferior da célula. Importante ressaltar que o tempo médio refere-se ao tempo que o algoritmo levou para a detecção de todas as faces detectadas, e não ao tempo médio de execução de cada face. A média se dá de acordo com a quantidade de vezes que o algoritmo de detecção foi executado para cada combinação de parâmetros, definido em "Number of Samples".

Em uma primeira análise, a partir resultado do exemplo dado na figura 5, é possível observar a tendência de se ter maior quantidade de faces detectadas, bem como maior tempo de execução, quanto mais ao topo e à esquerda está a célula na matriz. Essa

observação é um tanto óbvia tendo em vista que quanto menor o tamanho mínimo de faces a ser considerado pelo algoritmo (parâmetro "Min. Size"), mais faces candidatas tendem a ser encontradas e também mais iterações serão realizadas. O mesmo acontece para o parâmetro "Scale Factor", quanto menor o valor do parâmetro, menor o passo entre as escalas, maior a quantidade de imagens escalonadas avaliadas e, portanto, maior a chance de uma determinada face ser detectada, bem como maior quantidade de iterações realizadas.

A matriz de resultados por si só não é suficiente para se determinar que uma combinação de parâmetros é a mais adequada a se adotar. Isso se deve ao fato de que não se pode afirmar que o resultado com o maior número de faces detectadas é o melhor pois, dentro do conjunto de faces detectadas, alguma poucas ou até várias delas podem ser falsos positivos, de tal forma que a qualidade do resultado da detecção deva ser considerado ruim.

Outra questão a se considerar é que, dependendo do ambiente e do objetivo da aplicação, a detecção da maior quantidade de faces possível pode não ser a prioridade. Em determinado tipo de aplicação, pode ser mais interessante que o algoritmo detecte apenas faces que estejam mais próximas da câmera (e por conta disso relativamente maiores que as demais), com um tempo de resposta menor, do que detectar o máximo de faces possível, inclusive as mais distantes, que seriam descartadas, com um tempo de resposta maior.

Para tanto, é importante que haja uma análise qualitativa, ao se checar, para cada resultado analisado, quais foram as faces detectadas, resultantes da correspondente combinação de parâmetros. E, a partir da avaliação dos resultados disponíveis, comparando a presença de falsos positivos, a quantidade e tamanho das faces detectadas e o tempo médio de resposta, determinar uma combinação de parâmetros mais adequada a se adotar em uma aplicação de detecção no dispositivo testado ou concluir que o dispositivo não seria capaz de rodas a aplicação da forma desejada.

Para facilitar esse tipo de análise qualitativa, a matriz de resultados responde de forma interativa, de forma que, ao se clicar em uma célula da matriz, as faces detectadas a partir dos parâmetros correspondentes daquela célula são destacadas na

imagem original através de retângulos verdes, facilitando assim a verificação de falsos positivos e quais faces presentes na imagem o algoritmo conseguiu detectar com tais parâmetros.

A seguir são apresentados alguns exemplos de resultados, a partir da matriz de resultado apresentada na figura 5. Para facilitar a visualização, serão apresentados apenas os cortes da matriz com o resultado selecionado e a imagem com as faces destacadas.

No exemplo da figura 6, têm-se um resultado com muito poucas faces detectadas devido ao valor relativamente alto do parâmetro "Min. Size".

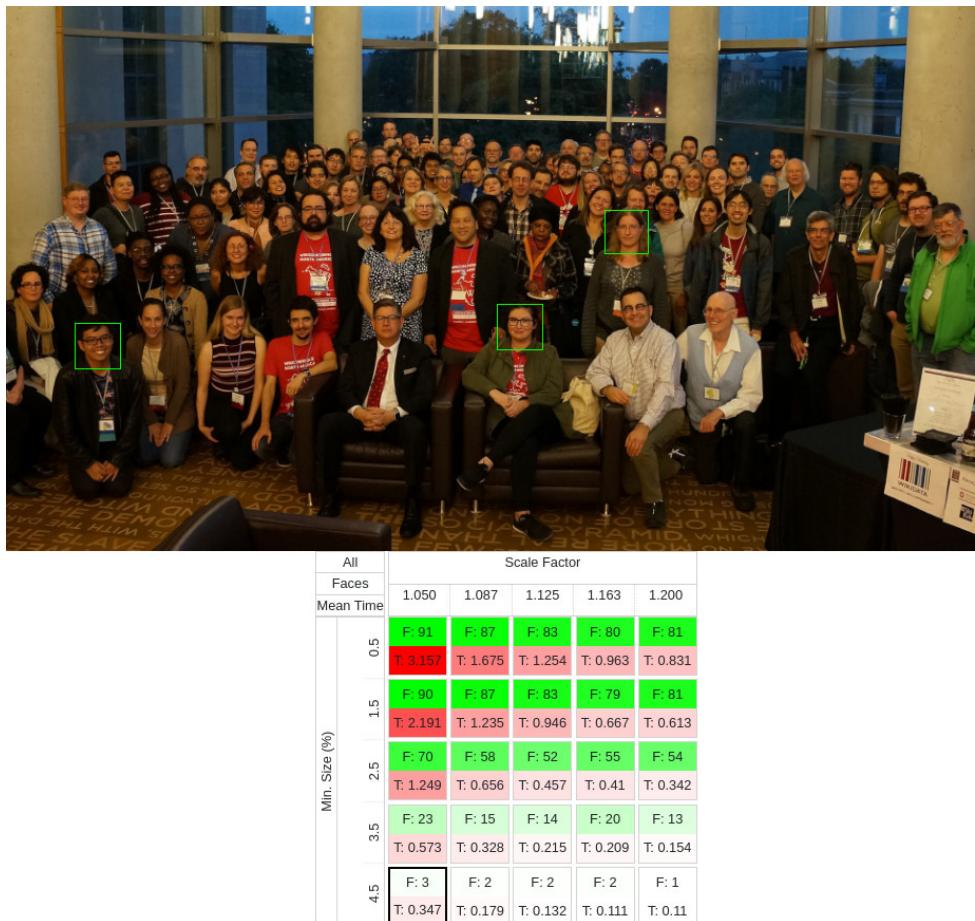
Já a figura 7 apresenta um resultado com várias faces detectadas, inclusive pequenas faces ao fundo. Porém, devido aos valores relativamente baixos dos parâmetros "Min. Size" e "Scale Factor", vê-se claramente a presença de três falsos positivos, além de um tempo médio de detecção consideravelmente alto, acima de 3 segundos.

Por fim, a figura 8 apresenta um resultado que possivelmente pode ser considerado satisfatório para determinados tipos de aplicação. Nesse caso, a maioria das faces presentes na imagem foi detectada e com um tempo de resposta razoavelmente baixo, pelo menos se comparado ao resultado apresentado na figura 7.

Outra facilidade que a ferramenta traz é quanto à flexibilidade na escolha dos limites e quantidades de valores a serem testados para cada parâmetros, permitindo, assim, ajustar os mesmos iterativamente, de forma a se ter cada vez melhores resultados e, consequentemente, melhores opções de otimização.

Ainda partindo da matriz de resultados da figura 5, alguns limites podem ser ajustados para uma próxima rodada de análise. Por exemplo, observa-se que nos resultados em que o valor de "Min. Size" é maior que 2.5, a quantidade de faces detectadas é muito baixa. Caso seja considerado insatisfatório, o novo valor máximo na distribuição desse parâmetro pode ser definido como 2.5 ou 3.0. Analisando o parâmetro "Scale Factor", pode-se observar que valores abaixo de 1.125 resultam em um tempo de resposta muito alto e apresenta muitos falsos positivos. Além disso, alguns resultados com o valor máximo apresentado de "Scale Factor", 1.200, apresentam uma quantidade

Figura 6 – Exemplo de resultado com pucas faces detectadas.



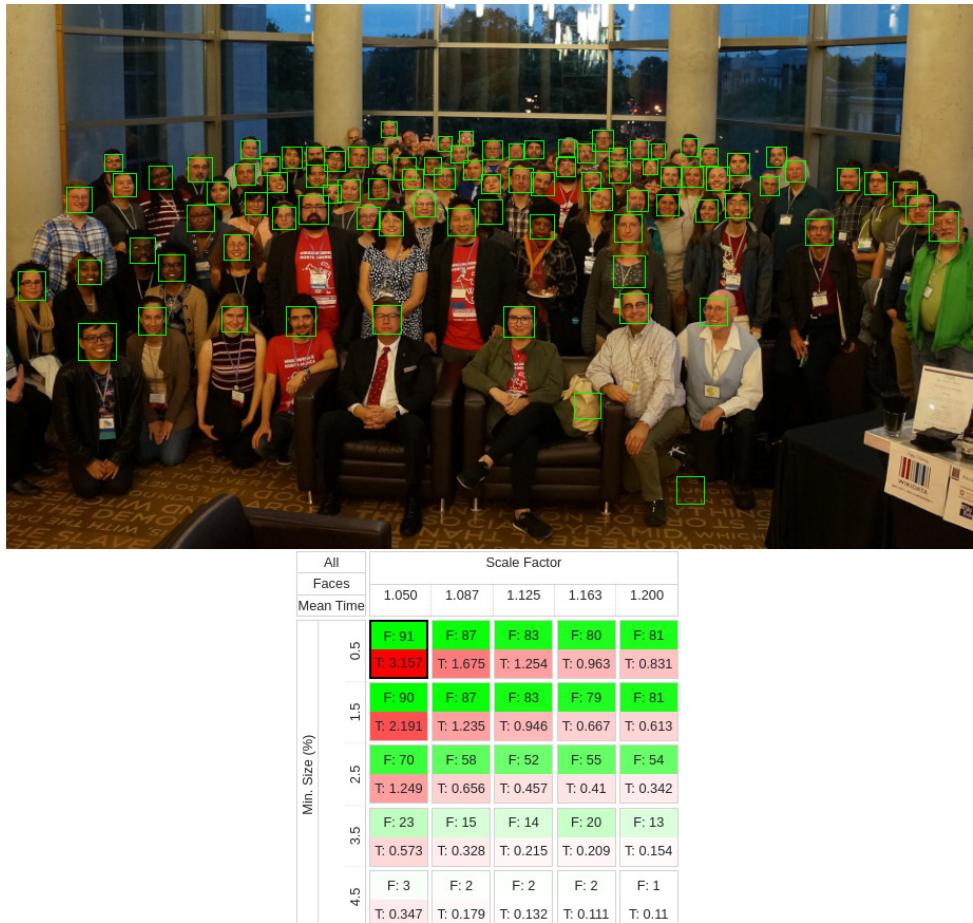
Fonte: autor.

alta de faces detectadas, sendo possível que um valor maior possa apresentar um resultado igualmente bom mas com um tempo de resposta menor. Os limites de "Scale Factor" poderiam ser ajustados, por exemplo, para 1.080 e 1.250. Assim, uma próxima rodada de análise com os limites ajustados irão retornar uma maior e melhor gama de resultados.

A título de exemplo, a figura 9 apresenta a matriz de resultado com os limites ajustados. Observa-se uma melhor distribuição, com números de faces detectadas e tempo de resposta mais próximos do que podem ser considerados como satisfatórios, possibilitando uma análise mais precisa para determinar os melhores parâmetros.

Há de se observar que, pelo fato de a análise estar sendo feita a partir de apenas uma imagem, não é de se esperar que o resultado quanto à presença, ou não, de falsos positivos, seja o mesmo para todos os frames que serão analisados em uma aplicação

Figura 7 – Exemplo de resultado com várias faces detectadas e alguns falsos positivos.



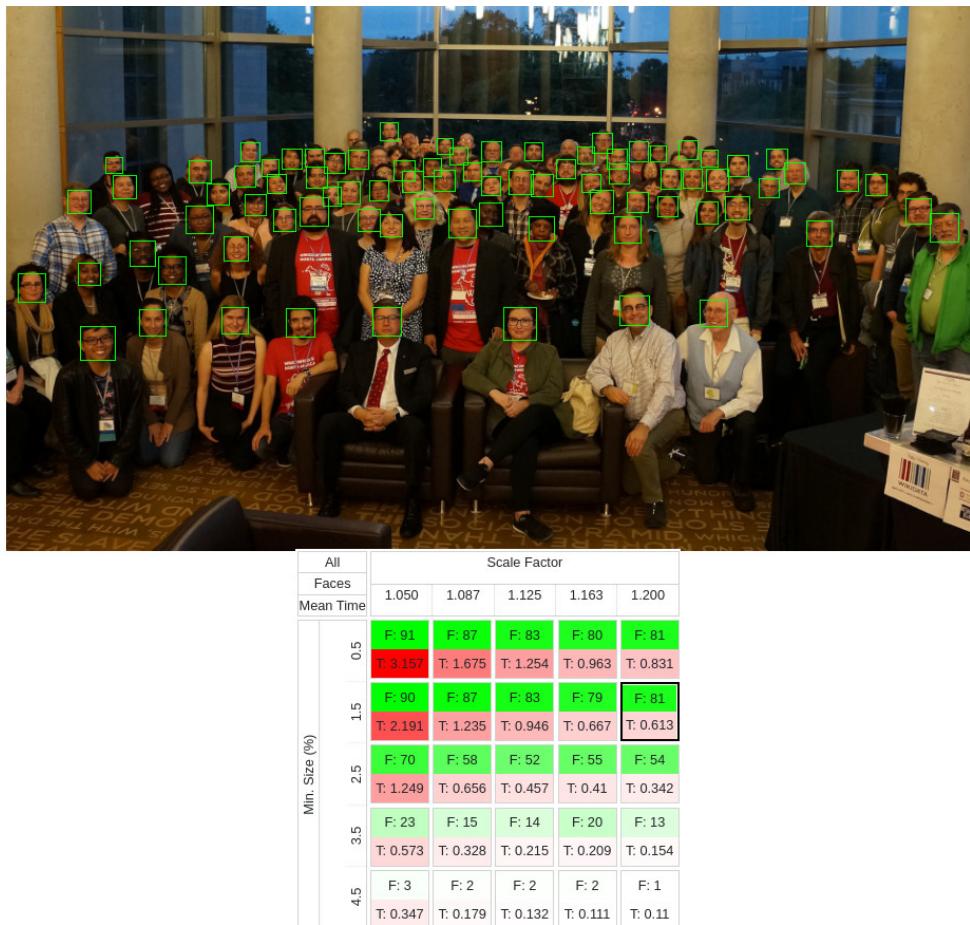
Fonte: autor.

real, mesmo a câmera estando fixa, capturando sempre o mesmo cenário.

Em uma aplicação real da ferramenta, é interessante que os valores escolhidos dos parâmetros sejam testados utilizando-se outras imagens da mesma cena com suas possíveis variações, como por exemplo a diferente quantidade de pessoas e objetos, iluminações diferentes, etc.

Tendo definido-se uma melhor combinação de valores dos parâmetros, pode-se então obter uma análise mais aprofundada, com métricas adicionais, que auxiliará na análise e comparação dos resultados entre os diferentes cenários e dispositivos testados. Os resultados serão analisados considerando não só qualidade de detecção e tempo de resposta, como também a demanda de banda em rede e quantidade de dados trafegados para se progragar o resultado a uma próxima etapa de processamento dentro de um sistema distribuído.

Figura 8 – Exemplo de resultado possivelmente satisfatório.



Fonte: autor.

Ao se clicar no botão "Get Metrics", no lado esquerdo da tela (vide figura 4), os parâmetros definidos de acordo com a célula selecionada na matriz de resultados são enviados ao servidor no dispositivo que está sendo testado. Este, por sua vez, executa novamente o algoritmo de detecção, logando também o tempo de execução de outras etapas preparatórias, bem como outras informações referentes aos tamanhos das imagens.

Na figura 10, pode-se observar como os dados retorados são exibidos na interface do usuário. A seguir, uma breve explicação de cada item:

- 1.0 - Params: os parâmetros utilizados no algoritmos de detecção, conforme célula selecionada na matriz de resultados na etapa anterior;
- 1.1 - Full Image Resolution: a resolução da imagem original, em pixels;

Figura 9 – Exemplo de matriz de resultado com limites ajustados.



Fonte: autor.

- 1.2 - Full Image Size (bytes): o tamanho da imagem original, em bytes;
- 1.3 - Full Image Encoded Size (bytes): o tamanho da imagem original codificada em bitmap e base64, em bytes;
- 1.4 - Number of detected faces: o número de faces detectadas;
- 1.5 - Cropped Faces Images Total Size (bytes) / (% from full image size): o tamanho total das imagens de faces detectadas, recortadas da imagem original, em bytes, e o seu percentual com relação ao tamanho da imagem original;
- 1.6 - Encoded Faces Images Total Size (bytes) / (% from full image size): o tamanho total das imagens de faces detectadas, codificadas em bitmap e base64, em bytes, e o seu percentual com relação ao tamanho da imagem original codificada;
- 2.1 - Loading Image (s): o tempo de carregamento da imagem original (leitura em disco), em segundos;
- 2.2 - Convert Image to Gray (s): o tempo de conversão da imagem original para escala de cinza, em segundos. Etapa anterior necessária para execução do algoritmo de detecção;

- 2.3 - Detection (s): o tempo de execução do algoritmo de detecção em si, em segundos;
- 2.4 - Build Encoded Faces Images (s): o tempo de execução da etapa de encodamento das imagens em base64, em segundos;
- 2.5 - Total execution time (s): o tempo total de execução e todas as etapas, desde o carregamento da imagem até o encodamento em base64, em segundos.
- 3.1 - Faces images: as imagens das faces detectadas, recortadas da imagem original em seu tamanho real. Permite-se ter uma ideia da qualidade de resolução individual das faces, bem como facilita a identificar mais facilmente a presença de falsos positivos que possam não terem sido identificados na etapa de análise anterior.

Figura 10 – Exemplo de resultado com as métricas.

Metric	Value
1.0 - Params	Min. Size Face: 1.5 / Scale Factor: 1.208 / Min. Neighbors: 3
1.1 - Full Image Resolution	1080 x 1920
1.2 - Full Image Size (bytes)	6.220.800
1.3 - Full Image Encoded Size (bytes)	8.294.472
1.4 - Number of detected faces	82
1.5 - Cropped Faces Images Total Size (bytes) / (% from full image size)	543.111 / 8.7%
1.6 - Encoded Faces Images Total Size (bytes) / (% from full image encoded size)	738.056 / 8.9%
2.1 - Loading Image (s)	0.274
2.2 - Convert Image to Gray (s)	0.004
2.3 - Detection (s)	0.654
2.4 - Build Encoded Faces Images (s)	0.008
2.5 - Total execution time (s)	0.94
3.1 - Faces images	

Fonte: autor.

É importante ressaltar que, como os testes são feitos a partir de imagens pré-selecionadas, a métrica 2.1 considera o tempo de carregamento em memória da

imagem salva em disco. O problema é que esse tempo é limitado apenas pela velocidade de leitura no sistema de arquivos, enquanto que em uma aplicação real a aquisição das imagens se darão a partir de uma fonte, como um sensor conectado à placa, uma câmera USB ou via streaming, por exemplo.

Para se ter no estudo um tempo de aquisição de imagem mais realista, foram feitos testes à parte para se obter o tempo médio de captura a partir de um módulo de câmera conectado diretamente ao dispositivo. Foi utilizado um módulo com sensor OmniVision OV5647, com capacidade de capturar imagens com resolução de até 2592x1944, e interface CSI, próprio para ser utilizado com um Raspberry Pi. Para captura de imagens, foi utilizado o pacote *picamera*, que provê uma interface simplificada em Python para o módulo da câmera.

Segundo a própria documentação (Picamera, 2022), o pacote *Picamera* oferece diferentes formas de se fazer a captura de imagens, chamados de *ports*. De acordo com o *port* utilizado, a câmera se comporta de maneira diferente durante a captura, o que irá influenciar na qualidade da imagem e tempo de aquisição dos frames. Como cada cena possui uma proposta diferente de aplicação, os testes de tempo médio de captura foram feitos de forma diferentes e mais adequada a cada cena e serão melhores detalhados nas próximas subseções da seção a seguir.

3.3 Cenários de testes

Nesta seção são apresentados os dois cenários definidos para os testes. Cada cenário representa uma possível aplicação diferente e possui diferentes requisitos que servirão de balizas para a calibração e comparação dos resultados.

Para um estudo mais completo, serão testadas algumas variações de cada cenário, seja quanto à quantidade de faces presentes e/ou quanto à resolução da imagem testada, além, claro dos diferentes dispositivos que serão testados.

3.3.1 Cena 1

Nessa primeira cena deseja-se representar o monitoramento de um espaço aberto e amplo, onde é esperado um fluxo alto de pessoas e que estas possam estar a qualquer distância da câmera, sendo desejável que o dispositivo consiga detectar faces muito pequenas a ponto de maximizar quantidade de faces detectadas.

- **Variações possíveis** - variação de quantidade de faces e variação de resolução da imagem.
- **Requisitos mínimos** (para balizar a parametrização) - máximo 3 segundos de resposta.
- **Imagen para testes** - para representar esta cena, foi selecionada uma imagem (figura 11) com várias faces olhando na direção da câmera e em diferentes distâncias. Ter todas as faces olhando para a mesma direção, distancia-se de um cenário real, mas torna-se ideal para o estudo pois serve como um pior caso para a cena e facilita a obtenção de variações por quantidade de faces.

Figura 11 – Imagem selecionada para testes da cena 1.



Fonte: Wikimedia Hackathon Barcelona 2018, por Ckoerner, 2018¹.

¹ Disponível em: <https://commons.wikimedia.org/wiki/File:Wikimedia_Hackathon_Barcelona_2018_group_photo.jpg>

Arquivo de imagem sob a licença CC BY-SA 4.0: <<https://creativecommons.org/licenses/by-sa/4.0/deed.en>>

3.3.1.1 Variação de quantidade de faces

Nessa cena, como a aplicação tem por objetivo detectar um elevado número de faces simultaneamente, pretende-se verificar como a variação da quantidade de faces presente na imagem afeta tanto no tempo de resposta do processo de detecção quanto no tamanho do conjunto das imagens das faces detectadas, recortadas da imagem completa, representando economia na utilização de banda para transmissão do resultado da detecção.

Para a realização desses testes, a escolha de uma imagem com um grande número de faces foi importante para que se pudesse obter 5 variações da mesma, reduzindo iterativamente a quantidade de faces em cada uma, da forma mais linear possível.

As variações foram preparadas utilizando-se o software de edição de imagem GIMP (GNU Image Manipulation Program), aberto e gratuito. Partindo-se da imagem original, uma certa quantidade de faces foram borradadas de forma a se tornarem indetectáveis, como se aquela faces não estivessem presentes na imagem. A imagem resultante se tornou a primeira variação por quantidade de faces. A partir da imagem dessa primeira variação, foi feito o mesmo procedimento, borrando a mesma quantidade de faces que ainda não haviam sido borradadas. E assim foi feito sucessivamente até obter-se um total de 5 variações. Importante ressaltar que, durante a preparação das imagens, preocupou-se em borrar as faces de uma forma bem distribuída.

Na figura 12, um exemplo de variação com a redução de 40 faces. Ao se comparar o resultado das variações, é de se esperar que a diferença de faces detectadas seja igual à diferente da quantidade de faces borradadas, porém, não necessariamente será igual, pois pode ser que alguma face não seja detectada pelo algoritmo de qualquer forma.

Para obtenção dos dados nos testes de variação de quantidade de faces deve-se, a partir da imagem original, com todas as faces detectáveis, utilizar a ferramenta para encontrar uma combinação otimizada de parâmetros de detecção. Com os melhores parâmetros definidos, obtém-se as métricas da imagem original e também de todas a variações de quantidade de faces geradas a partir dessa mesma imagem para futura

Figura 12 – Exemplo de variação de cena com redução de 40 faces.



Fonte: autor.

comparação.

3.3.1.2 Variação de resolução da imagem

Outra variação que é possível obter dessa mesma imagem e que fará parte do estudo é no que tange à resolução da imagem. Quanto maior a resolução da imagem, maior tende a ser a capacidade e qualidade de detecção de faces menores, mas também maior é a demanda de processamento e maior tende a ser o tempo de espera de uma detecção, bem como a utilização de banda para passar adiante as imagens das faces detectadas.

Dependendo do cenário e dos possíveis tipos de aplicação, considerar trabalhar a detecção em imagens com resoluções menores que a capacidade do sensor pode ser suficiente para viabilizar o uso do dispositivo de borda para a tarefa de detecção, desde que cumprindo os requisitos de tempo de resposta e capacidade de detecção.

Para a realização dos testes com variação de resolução, foram obtidas das imagens originais (a original e suas respectivas variações por quantidade de faces) imagens com resoluções menores, através de interpolação, utilizando-se também o software GIMP para o processamento. As imagens originais de cada cena foram selecionadas já

com resolução em quad-hd (2560x1440), e foram obtidas as variações nas resoluções full-hd (1920x1080) e hd (1280x720). Para cada redução de resolução, são obtidas 5 novas imagens, uma a partir da imagem original e outras quatro a partir da variações de quantidade de faces.

Para a obtenção dos dados nos testes de variação de resolução, utilizou-se novamente a ferramenta para encontrar a melhor combinação de parâmetros de detecção para as imagens na nova resolução e, a partir desses novos parâmetros, obteve-se as métricas de todas as imagens com a referida resolução.

3.3.1.3 Tempo médio de captura

Como já adiantado no final da seção anterior, o tempo de aquisição da imagem que será considerado no estudo será o tempo médio de captura de frames a partir de um sensor OV5647, utilizando-se o pacote *Picamera*. Para a cena em questão, onde objetiva-se maximizar a quantidade de faces detectadas, podendo estas estarem bem distantes da câmera, a qualidade da imagem é um fator importante.

Um dos possíveis *ports* usados pelo Picamera é o *still port*, que força a captura da imagem utilizando-se toda a área do sensor, que no caso é de 2592x1944 pixels, mesmo que a resolução de saída da imagem seja menor. Além disso, usa um forte algoritmo para redução de ruído, proporcionando maior qualidade de imagem, sendo assim a forma mais adequada de considerar na captura de imagem no contexto dessa cena.

Para se obter o tempo médio de captura foi escrito um algoritmo simples, que faz a captura consecutiva de 10 frames utilizando-se o *still port* e retorna o tempo médio de captura. No algoritmo é configurado a resolução de saída dos frames para corresponder às resoluções testadas.

3.3.2 Cena 2

Nessa segunda cena, deseja-se representar o controle de acesso a um ambiente, onde o dispositivo com a câmera está posicionado estrategicamente próximo à abertura de acesso (porta, portão, etc) de forma a capturar de perto a face de quem estiver adentrando ao local. É importante uma resposta rápida na detecção pois a pessoa entrando no ambiente estará em movimento, permanecendo no campo de visão da câmera por pouco tempo.

- **Variações possíveis** - variação de resolução da imagem.
- **Requisitos mínimos** (para balizar a parametrização) - máximo 0.3 segundo de resposta com detecção a partir de 1s de distância.
- **Imagens para testes** - para representar esta cena, foram obtidos imagens a partir do sensor OV5647 conectado ao dispositivo, com o próprio autor simulando a entrada no ambiente.

3.3.2.1 Captura das imagens para teste

Para a captura dos frames para os testes, o dispositivo com a câmera foi posicionado estrategicamente ao lado de uma porta, a uma altura aproximada de 1,70, e levemente inclinado em direção à entrada, de forma que a câmera pudesse captar faces mais próximas e também a distâncias de até 2m. <pending> adicionar foto </pending>

Foram feitos alguns testes para obter a posição de uma pessoa mais próxima da entrada em que a câmera conseguisse capturar uma boa imagem do rosto. A partir dessa posição, mediu-se 1,6 m de distância, e capturou-se novos frames, com a pessoa posicionada a essa distância. A partir de uma distância de 1,7m uma pessoa caminhando rapidamente a no máximo 6 km/h, leva pelo menos 1s para chegar à posição inicial obtida ($6 \text{ km/h} * 1000 \text{ m/km} * 1\text{h}/3600\text{s} = 1,67\text{m/s}$). A uma presença de pelos menos 1s diante da câmera, com a face capturável, e a um tempo máximo de 0.3s de resposta definido na detecção, têm-se pelo menos 3 frames capturados e analizados enquanto uma mesma pessoa caminha até a entrada do ambiente. Esse

será considerado nosso pior caso e o frame obtido com a pessoa posicionada a 1.67m da posição inicial será utilizado nos testes para definição dos parâmetros e obtenção das métricas para comparação.

Explicar como será o uso da ferramenta dessa vez, no intuito de conseguir o limiar para reconhecimento da face àquela distância e no menos tempo

Script para obtenção dos frames

Utilização do port video (still já inviabiliza)

Importância da distância pra garantir pegar a partir de 1s de distância para menos (1,6 m, considerando velocidade média de corrida 6km/h) dando chance de capturar 3 vzs considerando máximo de 0.3 segundo de resposta - esse seria o pior caso

Definir 3 tamanhos de imagem. Explicar que será obtido o maior e pra amnter a mesma imagem a resolução reduzida no GIMP Fazer olhando para frente mas com a ressalva de que a pessoa pode estar olhando para outro lado.

3.3.2.2 Tempo médio de captura

3.3.2.3 Variações de resolução

3.3.3 Dispositivos testados

Serão testados dois dispositivos SBC *Single Board Computer* diferentes, cujos resultados serão comparados.

- SBC 1 - Raspberry Pi 4: quad-core Cortex-A72 com clock de 1,5 GHz e 4 GB de memória RAM.
- SBC 2 - Raspberry Pi Zero W: single-core ARM1176JZF-S com clock de 1,0 GHz e 512 MB de memória RAM.

4 EXPERIMENTOS REALIZADOS

A realização dos experimentos baseou-se na utilização da ferramenta de parametrização e obtenção de métricas, aplicado nas imagens tratadas de cada cena, conforme variações definidas no Capítulo anterior. Os dados obtidos foram organizados de forma a facilitar a análise e comparação entre as variações em cada cena e também entre os diferentes dispositivos sendo testados.

Para cada variação de resolução, que depende de uma nova seleção de parâmetros, será exibido a matriz de resultados final juntamente com as faces detectadas, após já feita a análise para se chegar a um melhor resultado, indicando quais foram os parâmetros definidos.

4.1 Cena 1

4.1.1 Otimização de parâmetros

Primeiramente, foi realizada a otimização dos parâmetros para cada variação de resolução, a partir das imagens com todas as faces disponíveis de cada resolução. No caso desta cena, que exige relativamente muito mais processamento, utilizou-se o Raspberry Pi 4B para a definição dos parâmetros, repetindo-os nos testes com o Raspberry Pi Zero W para base de comparação.

A definição dos parâmetros 'ótimos' não é objetiva. Para esta cena, buscou-se um melhor resultado em que houvesse a maior quantidade de faces detectadas sem a presença de falsos positivos e no menor tempo. Durante a análise, teve-se a razoabilidade de considerar na comparação entre os diferentes resultados da matriz que, um grande aumento no tempo de detecção não justifica um pequeno ganho relativo na quantidade de faces detectadas.

As próximas três subsubseções 4.1.1.1, 4.1.1.2 e 4.1.1.3 mostram através de

imagens, para cada resolução testada, a última iteração da matriz de resultados com os dados entrada, a célula destacada com os parâmetros escolhidos e a imagem com a faces detectadas.

Os parâmetros definidos de cada resolução foram usados para obter as métricas de cada variação de quantidade de faces detectáveis, e em ambos os dispositivos testados. Não será apresentado nesse documento cada resultado das métricas obtidos em tela. Todos os dados foram tabelados e utilizados para as comparações feitas nas subseções seguintes.

4.1.1.1 Resolução 1440p

Figura 13 – Otimização Cena1 resolução 1440p.



Fonte: autor.

4.1.1.2 Resolução 1080p

Figura 14 – Otimização Cena1 resolução 1080p.

Min. Neighbors: 3 Number of Samples: 6

Scale Factor* (x-axis) Min. Size in % from image width** (y-axis)

From	To	Steps	From	To	Steps
1.035	1.065	8	0.9	1.2	4

* Parâmetro que especifica o quanto a imagem é reduzida a cada escala.
** Tamanho mínimo da face a ser detectada, em razão da largura da imagem.

Send

Resultado

All Faces Mean Time	Scale Factor							
	1.035	1.039	1.044	1.048	1.052	1.056	1.061	1.065
	Min. Size (%)	0.9	1.0	1.1	1.2			
	F: 145 T: 3.59	F: 146 T: 3.278	F: 142 T: 2.88	F: 133 T: 2.707	F: 140 T: 2.503	F: 137 T: 2.347	F: 131 T: 2.152	F: 128 T: 1.98
	F: 145 T: 3.649	F: 146 T: 3.254	F: 142 T: 2.885	F: 133 T: 2.705	F: 140 T: 2.51	F: 137 T: 2.335	F: 131 T: 2.163	F: 128 T: 1.975
	F: 141 T: 3.468	F: 139 T: 3.11	F: 132 T: 2.75	F: 125 T: 2.548	F: 132 T: 2.354	F: 127 T: 2.213	F: 122 T: 2.019	F: 120 T: 1.825
	F: 117 T: 3.056	F: 113 T: 2.694	F: 114 T: 2.459	F: 111 T: 2.298	F: 103 T: 2.096	F: 95 T: 1.938	F: 108 T: 1.876	F: 100 T: 1.704



Fonte: autor.

4.1.1.3 Resolução 720p

Figura 15 – Otimização Cena1 resolução 720p.

Min. Neighbors: 4 Number of Samples: 2

Scale Factor* (x-axis) Min. Size in % from image width** (y-axis)

From	To	Steps	From	To	Steps
1.021	1.04	8	1.5	1.8	4

* Parâmetro que especifica o quanto a imagem é reduzida a cada escala.
** Tamanho mínimo da face a ser detectada, em razão da largura da imagem.

Send

Resultado

All	Scale Factor							
	1.021	1.024	1.026	1.029	1.032	1.035	1.037	1.040
Faces	F: 55	F: 52	F: 51	F: 45	F: 44	F: 44	F: 42	F: 43
Mean Time	T: 2.978	T: 2.387	T: 2.142	T: 2.003	T: 1.769	T: 1.649	T: 1.572	T: 1.461
1.5	F: 55	F: 52	F: 51	F: 45	F: 44	F: 44	F: 42	F: 43
1.6	T: 2.749	T: 2.344	T: 2.184	T: 1.972	T: 1.781	T: 1.632	T: 1.558	T: 1.46
1.7	F: 43	F: 38	F: 42	F: 36	F: 35	F: 32	F: 33	F: 34
1.8	T: 2.59	T: 2.221	T: 2.087	T: 1.9	T: 1.688	T: 1.543	T: 1.45	T: 1.387
	F: 27	F: 25	F: 24	F: 18	F: 16	F: 17	F: 13	F: 11
	T: 2.328	T: 2.067	T: 1.92	T: 1.674	T: 1.559	T: 1.408	T: 1.288	T: 1.228



Fonte: autor.

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

- AAZAM, M.; HUH, E. N. Fog computing and smart gateway based communication for cloud of things. In: *Proceedings - 2014 International Conference on Future Internet of Things and Cloud, FiCloud 2014*. [S.I.: s.n.], 2014. ISBN 9781479943586. 10
- AGGARWAL, C. C. *Neural Networks and Deep Learning*. [S.I.: s.n.], 2018. 9, 10, 16, 17
- ANDERSON, D. A. The aggregate burden of crime. *Journal of Law and Economics*, 1999. ISSN 00222186. 13
- ATZORI, L.; IERA, A.; MORABITO, G. The internet of things: A survey. *Computer networks*, Elsevier, v. 54, n. 15, p. 2787–2805, 2010. 22
- Bellavista, P. et al. Convergence of manet and wsn in iot urban scenarios. *IEEE Sensors Journal*, v. 13, n. 10, p. 3558–3567, 2013. 22
- BROWNLEE, J. *What is Deep Learning?* 2019. Disponível em: <<https://machinelearningmastery.com/what-is-deep-learning/>>. 16
- BRUNELLI, R.; POGGIO, T. Face recognition through geometrical features. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [S.I.: s.n.], 1992. ISBN 9783540554264. ISSN 16113349. 21
- CENDRILLON, R.; LOVELL, B. Real-time face recognition using eigenfaces. In: *Visual Communications and Image Processing 2000*. [S.I.: s.n.], 2000. ISSN 0277786X. 21
- CENEDESE, A. et al. Padova smart City: An urban Internet of Things experimentation. In: *Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014, WoWMoM 2014*. [S.I.: s.n.], 2014. ISBN 9781479947867. 14
- CERQUEIRA, D. R. C. et al. ANÁLISE DOS CUSTOS e Consequencias da violência no Brasil. *Texto Para Discussão*, 2007. 9, 13
- CHEN, N. et al. Dynamic urban surveillance video stream processing using fog computing. In: *Proceedings - 2016 IEEE 2nd International Conference on Multimedia Big Data, BigMM 2016*. [S.I.: s.n.], 2016. ISBN 9781509021789. 9, 15, 16
- COX, I. J.; GHOSN, J.; YIANILOS, P. N. Feature-based face recognition using mixture-distance. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.I.: s.n.], 1996. ISSN 10636919. 21
- DLIB C++ Library. Disponível em: <<http://dlib.net/>>. 21
- DOLUI, K.; DATTA, S. K. Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing. In: *GloTS 2017 - Global Internet of Things Summit, Proceedings*. [S.I.: s.n.], 2017. ISBN 9781509058730. 10

- FISCHLER, M. A.; ELSCHLAGER, R. A. The Representation and Matching of Pictorial Structures Representation. *IEEE Transactions on Computers*, 1973. ISSN 00189340. 19
- G1. *Brasil tem a terceira maior taxa de roubos da América Latina, diz Pnud*. 2013. 9, 13
- HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. ISBN 9781467388504. ISSN 10636919. 17
- HELLER, M. *Melhores bibliotecas de Machine e Deep Learning*. 2019. Disponível em: <<https://cio.com.br/melhores-bibliotecas-de-machine-e-deep-learning/>>. 17
- HU, W. et al. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 2004. ISSN 10946977. 15
- JAFRI, R.; ARABNIA, H. R. A Survey of Face Recognition Techniques. *Journal of Information Processing Systems*, 2009. ISSN 1976-913X. 20
- JEBARA, T. S. 3D POSE ESTIMATION AND NORMALIZATION FOR FACE RECOGNITION. *Department of Electrical Engineering*, 1996. 21
- KANADE, T. *Computer recognition of human faces*. [S.l.: s.n.], 1977. 19, 20
- LABLED Faces in the Wild. Disponível em: <<http://vis-www.cs.umass.edu/lfw/>>. 21
- MERENDA, M.; PORCARO, C.; IERO, D. Edge machine learning for ai-enabled iot devices: A review. *Sensors (Switzerland)*, 2020. ISSN 14248220. 10
- NEAPOLITAN, R. E. Neural Networks and Deep Learning. *Artificial Intelligence*, p. 389–411, 2018. 17
- NIKOUEI, S. Y. et al. Smart surveillance as an edge network service: From harr-cascade, SVM to a Lightweight CNN. In: *Proceedings - 4th IEEE International Conference on Collaboration and Internet Computing, CIC 2018*. [S.l.: s.n.], 2018. ISBN 9781538695029. 9, 15, 16
- OLSON, T. J. *AUTOMATIC VIDEO MONITORING SYSTEM WHICH SELECTIVELY SAVES INFORMATION*. 2006. 16 p. Disponível em: <<https://patentimages.storage.googleapis.com/f0/7b/a2/58558376b25dca/US7023469.pdf>>. 14
- PACHECO, A. et al. A Smart Classroom Based on Deep Learning and Osmotic IoT Computing. In: *2018 Congreso Internacional de Innovacion y Tendencias en Ingenieria, CONIITI 2018 - Proceedings*. [S.l.: s.n.], 2018. ISBN 9781538681312. 9, 23, 24
- Picamera. 2022. [Online; acessado em 23-10-2022]. Disponível em: <<https://picamera.readthedocs.io/>>. 36
- PORTER, R.; FRASER, A. M.; HUSH, D. Wide-area motion imagery. *IEEE Signal Processing Magazine*, 2010. ISSN 10535888. 16
- PUVVADI, U. L. et al. Cost-effective security support in real-time video surveillance. *IEEE Transactions on Industrial Informatics*, 2015. ISSN 15513203. 14, 15

QUIRITA, V. H. A. ESTUDO DE MÉTODOS AUTOMÁTICOS DE RECONHECIMENTO FACIAL PARA VÍDEO MONITORAMENTO. 2014. 9, 18, 19, 20

Ramos Lima, G.; Marques Ciarelli, P. Sistema de Videomonitoramento com Identificação de Suspeitos Utilizando Biometria Facial. In: . [S.l.: s.n.], 2019. 14

SINGH, S. Optimize cloud computations using edge computing. In: IEEE. *2017 International Conference on Big Data, IoT and Data Science (BID)*. [S.I.], 2017. p. 49–53. 23

SONG, H. et al. *Smart Cities: Foundations, Principles, and Applications*. [S.l.: s.n.], 2017. ISBN 9781119226444. 9

Stan Z. Li, A. K. J. *Handbook of Face Recognition*. Second edi. New York: Springer, 2011. 19, 20

SZELISKI, R. Computer vision: algorithms and applications. *Choice Reviews Online*, 2011. ISSN 0009-4978. 10, 18, 19

TSAKANIKAS, V.; DAGIUKLAS, T. Video surveillance systems-current status and future trends. *Computers and Electrical Engineering*, 2018. ISSN 00457906. 15, 18

VERHELST, M.; MOONS, B. Embedded Deep Neural Network Processing: Algorithmic and Processor Techniques Bring Deep Learning to IoT and Edge Devices. *IEEE Solid-State Circuits Magazine*, 2017. ISSN 19430582. 10

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2001. ISSN 10636919. 18, 25

WILSON, J.; KELLING, G. Broken Windows: the police and neighborhood safety. *The Atlantic Monthly*, 1982. ISSN 0094-6575. 14

ZAFEIRIOU, S.; ZHANG, C.; ZHANG, Z. A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 2015. ISSN 1090235X. 18, 19