UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Sistema Automático de Contagem de Audiência com Uso de Aprendizagem Profunda

Caio Souza Florentino

JOÃO PESSOA - PB

Abril - 2020

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Sistema Automático de Contagem de Audiência com Uso de Aprendizagem Profunda

Caio Souza Florentino

Dissertação submetida ao Centro de Informática da Universidade Federal da Paraíba como parte dos requisitos necessários para obtenção do grau de Mestre em Informática.

Orientador: Prof. Dr. Rostand Edson Oliveira Costa

João Pessoa
2020

Catalogação na publicação
Seção de Catalogação e Classificação

F633s Florentino, Caio Souza.
Sistema Automático de Contagem de Audiência com Uso de Aprendizagem Profunda / Caio Souza Florentino. - João Pessoa, 2020.
79 f. : il.

Orientação: Rostand Edson Oliveira Costa.
Dissertação (Mestrado) - UFPB/CI.

1. contagem de audiência. 2. aferição de bilheteria. 3. visão computacional. 4. aprendizagem de máquina. I. Costa, Rostand Edson Oliveira. II. Título.

UFPB/BC

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de Caio Souza Florentino, candidato ao título de Mestre em Informática na Área de Sistemas de Computação, realizada em 27 de julho de 2020.

Aos vinte e sete dias do mês de julho do ano de dois mil e vinte, às quatorze horas e trinta minutos, por meio de videoconferência, reuniram-se os membros da Banca Examinadora constituída para julgar o trabalho do sr. Caio Souza Florentino, vinculado a esta Universidade sob a matrícula nº 20181000902, candidato ao grau de Mestre em Informática, na área de "Sistemas de Computação", na linha de pesquisa "Computação Distribuída", do Programa de Pós-Graduação em Informática, da Universidade Federal da Paraíba. A comissão examinadora foi composta pelos professores: Rostand Edson Oliveira Costa (PPGI-UFPB) Orientador e Presidente da Banca, Guido Lemos de Souza Filho (PPGI-UFPB), Examinador Interno, Tiago Maritan Ugulino de Araújo (PPGI-UFPB), Examinador Interno, Denio Mariz Timoteo de Sousa (IFPB), Examinador Externo à Instituição. Dando início aos trabalhos, o Presidente da Banca cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que o mesmo fizesse a exposição oral do trabalho de dissertação intitulado: "Sistema Automático de Contagem de Audiência com Uso de Aprendizagem Profunda". Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: "***aprovado***". Do ocorrido, eu, Ruy Alberto Pisani Altafim, Coordenador do Programa de Pós-Graduação em Informática, lavrei a presente ata que vai assinada por mim e pelos membros da banca examinadora. João Pessoa, 27 de julho de 2020.

Prof. Dr. Ruy Alberto Pisani Altafim

Prof. Rostand Edson Oliveira Costa
Orientador (PPGI-UFPB)

Prof. Guido Lemos de Souza Filho
Examinador Interno (PPGI-UFPB)

Prof. Tiago Maritan Ugulino de Araújo
Examinador Interno (PPGI-UFPB)

Prof. Denio Mariz Timoteo de Sousa
Examinador Externo à Instituição (IFPB)

Agradecimentos

Agradeço à toda minha família pelo apoio e suporte não só durante o período do Mestrado, mas em todos os momentos.

Agradeço aos meus colegas de turmas, aos meus colegas e coordenadores dos Laboratórios de Pesquisa de qual participei e aos meu colegas de trabalho, sempre contribuindo com o compartilhamento de conhecimento.

Agradeço ao meu orientador, professor Dr. Rostand Edson Oliveira Costa, pela paciência, sabedoria e disponibilidade durante todo esse período de orientação.

Agradeço à todos os integrantes do PPGI, professores, servidores e coordenadores, pela contribuição constante e significativa.

Meus sinceros agradecimentos à todos que colaboraram de alguma forma.

Resumo

Contar objetos ou seres vivos é uma necessidade comum em muitas áreas da indústria, comércio e serviços. Automatizar essa atividade pode promover uma otimização do processo envolvido e, consequentemente, a redução de tempo e custos. Com isso em mente, a visão computacional é uma abordagem que oferece novas possibilidades para o processamento digital de imagens, dando ao computador uma capacidade de interpretação cada vez mais semelhante aos humanos. Este trabalho compara a eficiência das técnicas de contagem volumétrica, tanto na visão computacional tradicional quanto na aprendizagem profunda, na contagem de audiências em eventos presenciais. Como estudo de caso, a investigação concentrou-se na contagem de audiência de sessões de cinema e / ou teatro a partir das fotos da audiência. Medir o público real de forma automática, precisa e transparente é uma necessidade recorrente da indústria do entretenimento. Para a realização dos experimentos, foi necessário o desenvolvimento de uma base de imagens com exemplos de audiências e a quantidade de pessoas presente. A partir dos resultados foi possível observar a eficiência da aplicação da aprendizagem profunda nesse contexto. Quando comparada a várias técnicas automáticas de contagem volumétrica disponíveis, a aprendizagem profunda foi a estratégia que apresentou os melhores resultados, atingindo sensibilidade e precisão acima de 96%. É proposto um Sistema Automático de Contagem de Audiência que contém os módulos de classificação/contagem (uso de aprendizagem profunda para contagem de audiência), captura (monitoramento contínuo das imagens para melhor captura) e controle (integração, administração e operação do sistema). O trabalho realiza contribuições com o compartilhamento de conhecimento na seleção de redes neurais que melhor realizam a tarefa de contagem de um grande número de objetos em uma imagem, no desenvolvimento de duas bases de teste de detecção de pessoas em audiência, na especificação de critérios para realizar a tarefa de contagem com êxito e na viabilização e desenvolvimento de um sistema de contagem automática de audiências.

Palavras-chave: contagem de audiência; aferição de bilheteria; visão computacional; aprendizagem de máquina.

Abstract

Counting objects or living things is a common necessity in many areas of industry, commerce and services. Automating this activity can promote an optimization of the process involved and, consequently, the reduction of time and costs. With this in mind, computer vision is an approach that offers new possibilities for digital image processing, giving the computer an increasingly similar interpretive capability to humans. This work compares the efficiency of volumetric counting techniques, both in traditional computational view and in deep learning, in the counting of audiences in face-to-face events. As a case study, the investigation focused on the audience count of movie and / or theater sessions from the audience photos. Measuring billing automatically, accurately and transparently is a recurring need in the entertainment industry. For the accomplishment of the experiments, it was necessary to develop an image base with examples of audiences and the amount of people present. From the results it was possible to observe the great potential of the application of deep learning in this context. When compared to several automatic volumetric counting techniques available, deep learning was the strategy that presented the best results, reaching sensitivity and precision above 96%. It is proposed an Automatic Audience Counting System that contains the classification / counting modules (it uses deep learning for audience count), capture (continuous monitoring of images for better capture) and control (integrates, manages and operates the system). This work contributes with knowledge sharing in the following aspects: selection of neural networks that best perform the task of counting a large number of objects in images, development of two test image bases of detection of people in audience, specification of requirements to perform the counting task successfully and in enabling and developing an automatic audience counting system.

Keywords: audience couting; computer vision; machine learning; box office records.

Conteúdo

Lista de Símbolos

HTML : HyperText Markup Language

YOLO : You Only Look Once

SCAA : Sistema de Contagem Automática de Audiência

MCla : Módulo Classificador

MCap : Módulo de Captura

MCon : Módulo de Controle

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O aumento da população e da extensão territorial das cidades são desafios complexos e que precisam de planejamento efetivo. A urbanização traz consigo problemas do ponto de vista estrutural que agrava questões como transporte, poluição do ar, consumo de energia, qualidade de vida, administração do espaço público, entre outros. Áreas emergentes de estudo se preocupam com o aumento da eficiência dos processos dentro do espaço urbano utilizando-se de estratégias e ferramentas da computação. Atualmente já é possível a obtenção, sistematização e inferência sobre uma grande quantidade de dados variados obtidos por diversas fontes espalhadas pelas cidades, como sensores, câmeras e interações de seres humanos. A análise de tais dados gera conhecimento e diagnósticos sobre desafios presentes e futuros, indicando as melhores formas de resolvê-los [I.S. SILVA (2014)]. No caso das câmeras de imagem, por exemplo, o uso de tecnologias adicionais pode ampliar a função desse tipo de dispositivo. Ferramentas que analisem e extraiam informações das imagens capturadas podem proporcionar uma mudança significativa na capacidade de monitoramento de ambientes.

No contexto multidisciplinar das cidades, a otimização do uso da infraestrutura urbana não é apenas um desafio constante, mas uma necessidade crescente, e a busca pela automação de processos é uma das estratégias mais adotadas para isso. Sempre presente nas pesquisas e na indústria, a automatização tenta reproduzir, por meio de máquinas e/ou computadores, as decisões e ações humanas nos processos. Entretanto, o ser humano tem uma capacidade significativa de processar informações e aprender sobre elas, o que demanda o desenvolvimento de algoritmos que possam ser baseados nessa capacidade. A contagem de objetos ou seres vivos, por exemplo, é uma necessidade comum em diversas áreas da indústria [Uma

e Yuvarani (2017)], comércio [Del-Blanco et al. (2012)] e serviços em geral. Automatizar essa atividade pode promover uma otimização do processo envolvido e, por consequência, a redução de tempo e custos.

A visão pode ser compreendida como o sentido que mais fornece informações ao ser humano sobre o ambiente ao seu redor. Dentro deste cenário, a *Visão Computacional* busca emular a visão humana através de um conjunto de técnicas que extraem informações de uma imagem [Ballard (1982)]. A área de *Visão Computacional* abrange tarefas que podem ser divididas, de forma geral, em aquisição, processamento e interpretação de imagens. O processamento de imagens, em particular, é um processo onde a entrada do sistema é uma imagem e a saída é um conjunto de valores numéricos, que podem ou não compor uma outra imagem [Molz (2001)].

A técnica de *Contagem Volumétrica Automática*, uma das aplicações de Visão Computacional, permite detectar, reconhecer e quantificar objetos em uma imagem sem a necessidade da intervenção humana. Assim, é possível fazer a contagem de objetos de maneira automática. Dentro de cada etapa da contagem volumétrica vários métodos podem ser utilizados. O pré-processamento pode ser baseado na escala de cinza, operadores morfológicos ou limiarização, por exemplo [Fathy e Siyal (1995)]. A segmentação pode ser baseada em pontos, linhas, bordas ou orientada a regiões. O reconhecimento pode usar redes neurais ou métodos exatos [I.S. SILVA (2014)].

1.1 Justificativa e Motivação

Com o desenvolvimento da tecnologia e a redução dos custos computacionais, as técnicas de Aprendizagem Profunda, ou *Deep Learning*, que serão discutidas no Capítulo 2, tem se popularizado cada vez mais e ramificado para diversas áreas de atuação. Saúde, robótica, visão computacional, sistemas de recomendação, são algumas áreas onde a Inteligência Artificial, especificamente o *Deep Learning*, tem avançado constantemente. Na Visão Computacional, o uso de Aprendizagem Profunda, pode ser visto em carros autônomos, diagnóstico por imagens, reconhecimento facial, monitoramento de câmeras de vigilância, entre outros.

Outra aplicação de destaque é a Contagem Volumétrica, no caso deste trabalho, a quantificação de pessoas em um ambiente. Em ambientes públicos, essa aplicação pode servir

como suporte para o controle e segurança de espaços que tenham limitações na capacidade de pessoas presentes. A lotação acima da capacidade de ambientes confinados pode trazer diversos riscos a todos os presentes. Em locais comerciais, a contagem de pessoas em um ambiente pode fornecer informações estratégicas para a empresa como também aferir os dados de comparecimento de eventos. Este segundo caso reflete a situação de teatros e cinemas, onde o número de pessoas presentes em cada sessão é, em vários aspectos, a informação mais relevante do empreendimento. Na grande maioria das redes de cinema, principalmente, a única fonte de informação é a bilheteria. Em outros casos, existem de forma adicional aos dados da bilheteria, dados das catracas que ficam na entrada dos cinemas, mas que não possuem informações individualizadas de cada sala. De toda forma, nesses cenários, não existem maneiras eficientes e seguras de aferir, em uma data futura, os valores de público real de cada evento. A importância de que esses valores reflitam fielmente a realidade valem tanto para as redes de cinemas quanto para as produtoras dos filmes, visto que, toda a cadeia da indústria cinematográfica é baseada na quantidade de pagantes.

1.2 Definição do Problema

O problema que se deseja abordar é ineficiência na realização da contagem automática de audiências presencias. As técnicas já adotadas não permitem uma contagem eficiente que se adapte ao contexto de eventos com audiências presenciais.

No cenário de contagem de pessoas em um ambiente, alguns trabalhos apresentam métodos e ferramentas que realizam essa contagem através de câmeras ou sensores na entrada desses recintos, quantificando o número de pessoas que entram e saem do local. Esses métodos demonstram-se eficientes e simples na tarefa de contar pessoas em um determinado local. Entretanto, a capacidade de auditar os valores apurados após a contagem é primordial em determinadas aplicações. Nestes casos, a aferição através de uma única imagem se mostra mais prática e rápida quando comparado a aferição através de um vídeo onde pessoas, individualmente, passam por uma entrada.

O problema abordado possui características específicas e fundamentais que a difere de outros estudos. Existe uma necessidade, explanada anteriormente, de realizar a contagem de audiência com o uso de uma única imagem, ao invés de múltiplos *frames* ou vídeos (tanto da

audiência como da porta de acesso) para simplificar o processo de auditoria.

Outro ponto que a contagem por marcação de rostos soluciona, é a privacidade das pessoas presentes no ambiente. As câmeras são posicionadas de frente para audiência e as pessoas tem seus rostos expostos. Esse cenário requer que as marcações da detecção de pessoas esconda os rostos, entrando assim no mesmo caso anterior onde a detecção das faces representam a situação ótima. A marcação do rosto permite que ele seja coberto no momento da contagem.

Por fim, outra necessidade primordial é que a solução seja genérica a vários ambientes. Isso inclui a exigência por abranger ambientes com iluminações diferentes. Esta circunstancia requer, principalmente, que todo o pre-processamento realizado das imagens sejam feita de maneira automatizada. Neste caso, o pré-processamento inclui correção de brilho, contraste, nitidez e remoção de *background*. Todas essas circunstâncias demonstram a especifidade da solução abordada neste trabalho.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral é investigar e obter um conjunto de métodos para realizar a contagem de audiência em eventos presenciais de maneira eficiente, auditável e compatível com a privacidade dos envolvidos.

1.3.2 Objetivos Específicos

Os objetivos específicos são:

1. levantar o estado da arte de técnicas para contagem volumétrica em imagens;
2. prospectar e testar técnicas de visão computacional clássica na contagem de pessoas;
3. prospectar e testar técnicas de *deep learning* na contagem de pessoas;
4. propor uma metodologia de processamento de imagem para auxiliar na contagem de pessoas.

5. propor um fluxo de controle e classificação para contagem de audiência;

6. validar o sistema proposto através de experimentos em cenários reais;

7. avaliar e documentar os resultados obtidos.

1.4 Metodologia

Neste trabalho foi analisada a eficiência de técnicas de contagem volumétrica, tanto usando visão computacional tradicional quanto *deep learning*, na contagem de audiências em eventos presenciais, com o objetivo de propor um Sistema Automático de Contagem de Audiência. O sistema proposto é composto por três módulos:

- **módulo de classificação/contagem** (uso de aprendizagem profunda para contagem de audiência);
- **módulo de captura** (monitoramento contínuo da câmera para melhor captura das imagens);
- **módulo de controle** (integra os módulos do sistema).

Como estudo de caso, os experimentos se concentraram na contagem do público de sessões de cinema, teatro e eventos esportivos de pequeno porte a partir de fotos da plateia. A partir dos resultados obtidos foi possível observar o grande potencial de aplicação de *deep learning* nesta circunstância quando comparada com outras técnicas de contagem volumétrica automática disponíveis.

Para atender aos objetivos propostos neste trabalho, primeiramente foi utilizado uma metodologia de Revisão da Literatura com a finalidade de encontrar técnicas potenciais para solução do problema definido e analisar trabalhos que se separaram com cenários semelhantes. Em um segundo momento foram realizados experimentos com as técnicas investidas e desenvolvido um protótipo usando uma metodologia de Construção (*Build*) para validar a solução encontrada.

O trabalho realiza contribuições com compartilhamento de conhecimento em alguns aspectos ao longo de seu desenvolvimento. Chega-se a conclusão, após a realização de testes, a arquitetura de rede neural que melhor realiza a tarefa de contagem de um grande número

de pessoas em uma imagem, e com expectativa para desempenho semelhante com outros objetos. Também foram desenvolvidas, ao longo do trabalho, duas bases de imagens para realização de testes de detecção de pessoas em audiência, uma base com imagens colhidas da internet e outra com imagens próprias. Outro fator é a realização de uma especificação de requisitos para que se possa realizar a tarefa de contagem com êxito. E por fim, é viabilizado o desenvolvimento de um sistema de contagem automática de audiências.

1.5 Escopo

Durante o desenvolvimento do trabalho percebeu-se a necessidade do desenvolvimento dos Módulos de Captura e Controle para tornar possível a solução proposta. Sem esses módulos, haveria a necessidade de realizar muitas configurações e adaptações no módulo principal de Classificação.

O escopo de investigação deste trabalho é desenvolver e validar técnicas e estratégias que compões o Módulo de Classificação do sistema proposto. O Módulo de Classificação é o núcleo do sistema, onde a contagem de pessoas de audiência presencias é realizado.

Os módulos de Captura e Controle foram desenvolvidos para possibilitar o teste e validação do sistema como um todo. Sendo assim, fica como indicação de trabalhos futuros a investigação aprofundada desses dois módulos.

1.6 Estrutura da Dissertação

Esta dissertação está organizada como descrito a seguir.

No Capítulo 2 é apresentada uma fundamentação teórica sobre Contagem de Audiência, realizada uma contextualização da técnica de *deep learning* e apresentadas ferramentas para a sua aplicação. São apresentados técnicas de Visão Computacional Clássica e descrita três arquiteturas de redes neurais que serão utilizadas nos testes realizados neste trabalho.

No Capítulo 3 são descritos alguns trabalhos relacionados que também usam técnicas de inteligência artificial para detectar, classificar e contabilizar pessoas em ambientes e situações diversas. É discutido a importância destes trabalhos para respaldar as soluções que seriam propostas mas, também, é ressaltado a diferença entre esses trabalhos relacionados e

o problema enfrentado nesta pesquisa.

No Capítulo 4 é feita a descrição da arquitetura proposta para o Sistema Automático de Contagem de Audiência. São detalhados as premissas e requisitos para o desenvolvimento do sistema e apresentado os módulos e a arquitetura da solução apresentada.

No Capítulo 5 é realizado o detalhamento das estratégias e heurísticas adotadas para a realização dos experimentos, a descrição de como foram realizados os treinamentos e validação dos modelos gerados pelas redes neurais. Também são detalhadas a base de imagens usada para o treinamento das redes neurais e a base que foi desenvolvida para a realização dos testes. E por fim, é realizado um experimento fatorial 2K para verificação dos impactos de variações no cenário para a rede neural selecionada para o protótipo da solução.

No Capítulo 6 é apresentado os resultados dos experimentos descritos no Capítulo 5 e, posteriormente, é feita uma discussão e análise. São apresentados os resultados de sensibilidade e acurácia das redes neurais analisadas e os resultados do experimento fatorial 2K.

Finalmente, o Capítulo 7 traz as considerações finais do trabalho, algumas propostas de próximos passos da pesquisa e as contribuições que foram realizadas. É avaliado todo o processo realizado neste trabalho, discutido a capacidade do sistema proposto de realizar as tarefas que solucionem o sistema proposto e os pontos em que se pode avançar para resultados mais robustos.

Capítulo 2

Fundamentação Teórica

Com o objetivo de se obter uma completa compreensão do problema abordado, fez-se necessário uma revisão sobre alguns dos conceitos e tecnologias utilizados na implementação da solução. Assim, neste Capítulo são discutidos a Contagem Automática de Audiências Presenciais, conceitos e técnicas que podem ser aplicadas neste contexto.

São vistos entendimentos sobre a Visão Computacional Clássica e técnicas derivadas dessa área. São explorados o uso de detecção de bordas e classificadores para a contagem de objetos. E, posteriormente, debatidos conceitos de *Deep Learining* e apresentados exemplos de redes próprias para detecção de objetos em imagens.

2.1 Contagem Automática de Audiências Presenciais

A contagem automática de pessoas pode ser uma ferramenta importante na tomada de decisão na administração de espaços públicos. Isso possibilita a resolução de problemas onde a contagem tradicional pode ter um custo alto, demandar muito trabalho ou até ser imprecisa. O diagnóstico e controle da quantidade de pessoas em um espaço público, por exemplo, pode ser uma ferramenta importante de segurança e prevenção de incidentes.

Outra área onde esse cenário se encaixa é a indústria de entretenimento, que lida com audiências como: esportes, teatros e cinemas. No caso da indústria de cinema, as produtoras e distribuidoras cinematográficas precisam contabilizar de forma confiável a quantidade de pessoas que foram assistir a um filme. Aferir a bilhetagem de forma automática, precisa e transparente é uma necessidade recorrente deste segmento. Normalmente, esta contagem

fica a cargo das redes de exibição e, na maioria dos casos, não passa por nenhum tipo de verificação. Baseado no fato que o faturamento de um filme vem, na maior parte, da arrecadação com bilheteria, isso pode causar um desbalanceamento na indústria cinematográfica prejudicando a produção cultural. Também no contexto dos cinemas, já se tornou costume a divulgação de propagandas comerciais antes de cada sessão. A transparência e precisão sobre a presença de público também interessa a agências de *marketing* e empresas que querem divulgar seus produtos.

Na próxima seção, será feita uma a revisão da aplicação de técnicas de visão computacional e *deep learning* para a contagem automática de pessoas.

2.2 Visão Computacional

A visão pode ser considerada o sentido mais importante, que capta mais informações, para os seres humanos. A complexidade desse sentido passa despercebido, mas trata-se de um processo que lida com cerca de 60 imagens por segundo com milhões de *pixels* em cada imagem. De fato, mais da metade do cérebro humano está envolvido no processamento da informação visual, indicação de que esta é uma tarefa muito complexa. [Poppe (2010)]

A visão computacional é uma área de atuação da informática que apresenta uma diversidade de possibilidades e ganha cada vez mais atenção. As câmeras tem, progressivamente, se tornado mais complexas. Em vários casos, elas estão sendo integradas em dispositivos como *smartphones* e óculos inteligentes ou se comunicam através da internet. E isso, por sua vez, está estimulando o desenvolvimento de sistemas de visão computacional cada vez mais complexos. A existência de todas essas câmeras, em diferentes aplicações, significa que muito mais dados de vídeo estão sendo produzidos e que existe uma necessidade de extrair informações desses vídeos de maneira mais inteligente. Além disso, a Visão Computacional vem sofrendo mudanças significativas desde o surgimento e avanço da *Deep Learning*, que permite o desenvolvimento de sistemas que executam tarefas de reconhecimento de padrões cada vez mais complexa. [LeCun et al. (2010)]

2.2.1 Visão Computacional Clássica

A técnica de *Contagem Volumétrica Automática*, uma das aplicações de Visão Computacional, permite detectar, reconhecer e quantificar objetos em uma imagem sem a necessidade da intervenção humana. Assim, é possível fazer a contagem de objetos de maneira automática. Dentro de cada etapa da contagem volumétrica vários métodos podem ser utilizados. O pré-processamento, ações realizadas anteriormente a aplicação da técnica em questão, pode ser baseado na escala de cinza, operadores morfológicos ou limiarização, por exemplo. A segmentação, processo que tem como objetivo dividir uma imagem em regiões homogêneas ou de interesse, pode ser baseada em pontos, linhas, bordas ou orientada a regiões. O reconhecimento - detecção do padrão em análise - pode usar redes neurais ou métodos exatos [I.S. SILVA (2014).]

Contagem Volumétrica por Detecção de Bordas

Uma das formas de realizar a contagem volumétrica é através da detecção de bordas [Davis (1975)], que usa uma abordagem onde os pontos de maiores variações de intensidade são classificados como bordas. Envolve uma série de métodos usados para identificar os pontos em uma imagem onde existem mudanças claras e definidas na intensidade. Esses métodos servem para extrair as informações relacionadas à imagem, como por exemplo nitidez de imagem, melhorias e localização de objetos.

Na prática, a imagem é convertida para escala de cinza, e observa-se as maiores variações de tom. Essas variações ocorrem com mais frequência entre um objeto em primeiro plano e o fundo da imagem. Assim é possível destacar e detectar bordas.

Na Figura 2.1 é mostrado um exemplo da aplicação da detecção de bordas na detecção de pessoas. A Figura mostra diferentes resultados para variações dos valores de parâmetros inerentes a essa técnica. A detecção de bordas tem baixa complexidade de implementação, baixo custo computacional mas tem como objetivo a detecção, e posterior segmentação, de contornos em uma imagem. Não tem como característica buscar e detectar um determinado objeto.

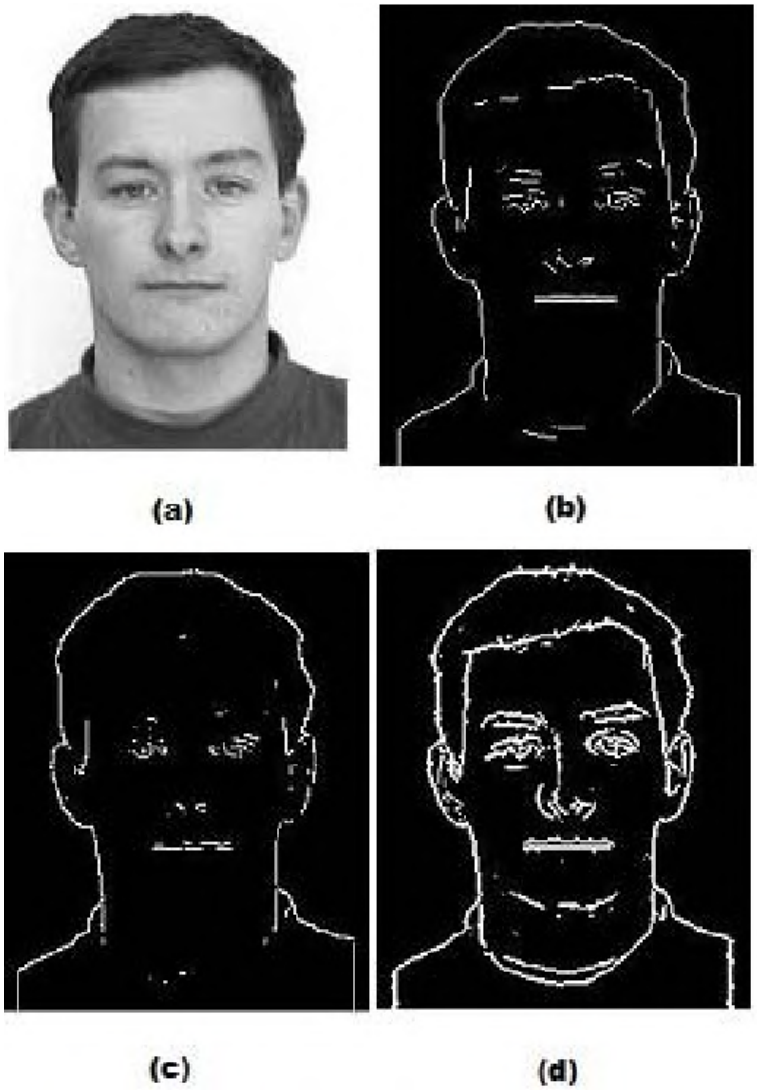


Figura 2.1: Exemplo da aplicação da técnica de detecção de bordas para contagem volumétrica. Fonte: Sudarshan, M et al.©

Contagem Volumétrica por Reconhecimento Facial por Classificadores

Outra forma de detecção de objeto é baseada em classificadores [Viola e Jones (2001)]. Existem conjuntos de classificadores que agrupados conseguem extrair padrões de imagens semelhantes. O classificador é treinado com amostras de um objeto específico, chamados exemplos positivos e exemplos negativos. Posteriormente, é gerado um vetor de características que pode ser aplicado a uma região de interesse em uma imagem de entrada. Essa técnica já é usada com frequência em redes sociais e câmeras digitais [Lienhart e Maydt (2002)].

Quando é necessário fazer a contagem volumétrica em imagens com mais detalhes e

mais pessoas, os classificadores atualmente estabelecidos Haar (baseado na ondulação Haar) e LBP (*Local Binary Patterns*) têm seu rendimento afetado negativamente.

O uso de Classificadores tem uma complexidade de implementação relativamente baixa e baixo custo computacional tanto na geração dos vetores de características quanto na aplicação dos classificadores. Mas, mesmo podendo ser treinado para se especializar na detecção de um objeto, essa técnica fica aquém de outras técnicas mais complexas em imagens com muitos detalhes.

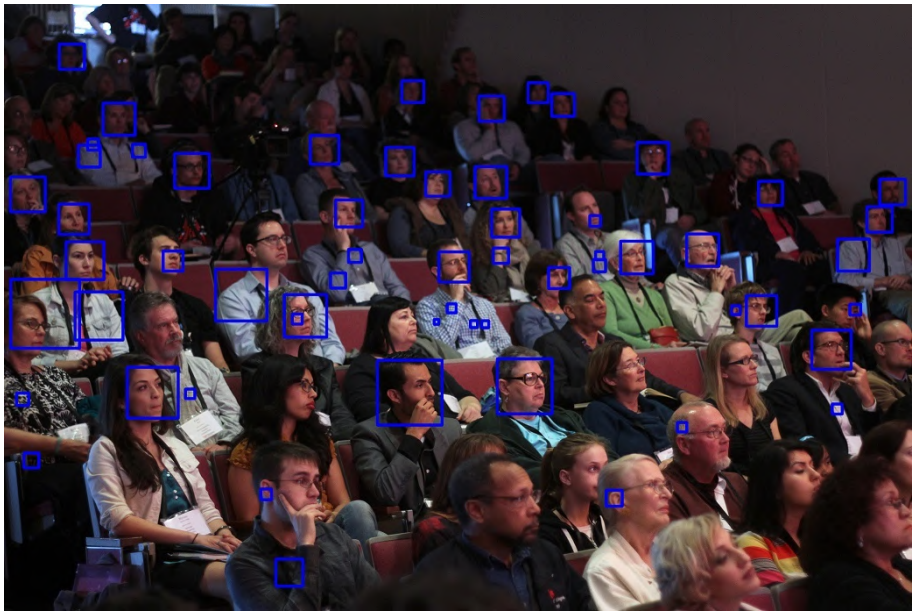


Figura 2.2: Resultado da detecção de objetos com classificadores Haar e LBP. Foto original por TEDx Monterey©

Outra técnica de contagem volumétrica utiliza-se classificadores HOG (*Histogram of Oriented Gradients*) [Tan et al. (2013)]. O método HOG tem como objetivo extrair informações referentes à orientação das arestas existentes em uma imagem, com o intuito de reconhecer padrões que possam caracterizar o objeto alvo da detecção. Os resultados se assemelharam aos dos classificadores Haar e LBP, também com baixo desempenho em imagens com muitos detalhes e número elevado de pessoas.

2.2.2 *Deep Learning* - DL

Sistemas baseados em aprendizado de máquina [Lempitsky e Zisserman (2010)], dentre várias outras aplicações [Ferreira et al. (2014)] [Messias et al. (2015)], são usados para identi-

ficar objetos em imagens, transcrever discursos em texto, áudio ou vídeo, agrupar notícias, avaliar postagens, oferecer produtos baseado no interesse dos usuários e selecionar os resultados mais relevantes de pesquisas.

Métodos de aprendizado profundo, ou *Deep Learning* [LeCun et al. (2015)], se treinados com um conjunto de dados rotulados suficientemente grande, podem atingir uma precisão que, às vezes, excedem o desempenho humano [Taigman et al. (2014)]. O termo "profundo"geralmente se refere ao número de camadas ocultas em uma rede neural [Schmidhuber (2015)]. Cada camada exerce uma função específica no processo de classificação de objetos.

Uma imagem, por exemplo, pode ser representada por uma matriz de *pixels*, e os recursos aprendidos na primeira camada de representação geralmente retratam a presença ou a ausência de bordas em locais particulares na imagem. As camadas seguintes, de forma geral, reconhecem o surgimento de padrões que em conjuntos caracterizam a presença de um objeto já conhecido. Um aspecto fundamental do aprendizado profundo é que essas camadas de recursos não são projetadas por especialistas humanos: são aprendidas a partir de dados usando um procedimento de aprendizado de propósito geral. Por tal característica, *deep learning* tem permitido grandes avanços na resolução de problemas que resistiram às melhores tentativas da comunidade de inteligência artificial por muitos anos [LeCun et al. (2015)].

Um dos tipos mais populares de redes neurais profundas é a *rede neural convolucional* (CNN ou ConvNet) [Razavian et al. (2014)]. As CNNs eliminam a necessidade de extração manual de recursos dos dados. Os recursos relevantes não são pré-treinados - eles são aprendidos enquanto a rede treina em uma coleção de imagens. Essa extração automatizada de recursos faz modelos de aprendizado profundo altamente precisos para tarefas de visão computacional, como a classificação de objetos [Bengio et al. (2012)].

Para o contexto de contagem volumétrica, foi realizada uma revisão da literatura e foram identificadas e selecionadas três abordagens DL baseadas em CNN com potencial de aplicação: YOLO [Redmon et al. (2016)], MT-CNN [Zhang et al. (2016a)] e ResNet [He et al. (2016)].

YOLO - You Only Look Once

YOLO [Redmon et al. (2016)] é uma abordagem para detecção de objetos em imagens. Essa técnica lida com a detecção de objetos como um problema de regressão para regiões de interesse espacialmente separadas e probabilidades de pertencimento de classe associadas. Uma única rede neural prediz regiões de interesse e probabilidades de classe diretamente de imagens completas em uma única avaliação.

Os sistemas de detecção atuais tomam um classificador para esse objeto e passam a avaliá-lo em vários locais e escalas em uma imagem de teste ou usam uma abordagem de janela deslizante onde o classificador corre em locais uniformemente espaçados em toda a imagem. Abordagens mais recentes, como R-CNN (Regional-CNN) [Ren et al. (2015)], usam métodos de região de interesse para gerar primeiro áreas delimitadas em potencial que possam conter um objeto em uma imagem e, em seguida, executar um classificador nas regiões propostas. Após a classificação, o pós-processamento é usado para refinar caixas de delimitação e eliminar detecções duplicadas. Cada classe de objeto deve ser treinado separadamente.

YOLO, por sua vez, aborda a detecção de objetos como um único problema de regressão, diretamente dos *pixels* da imagem. Desta forma, o sistema só analisa uma vez toda a imagem para prever quais objetos estão presentes e onde estão. Na Figura 2.3 é apresentada a arquitetura da rede YOLO. Atualmente, dentro do contexto de redes *deep learning*, a arquitetura da YOLO é considerada simplificada quando comparadas a outras redes.

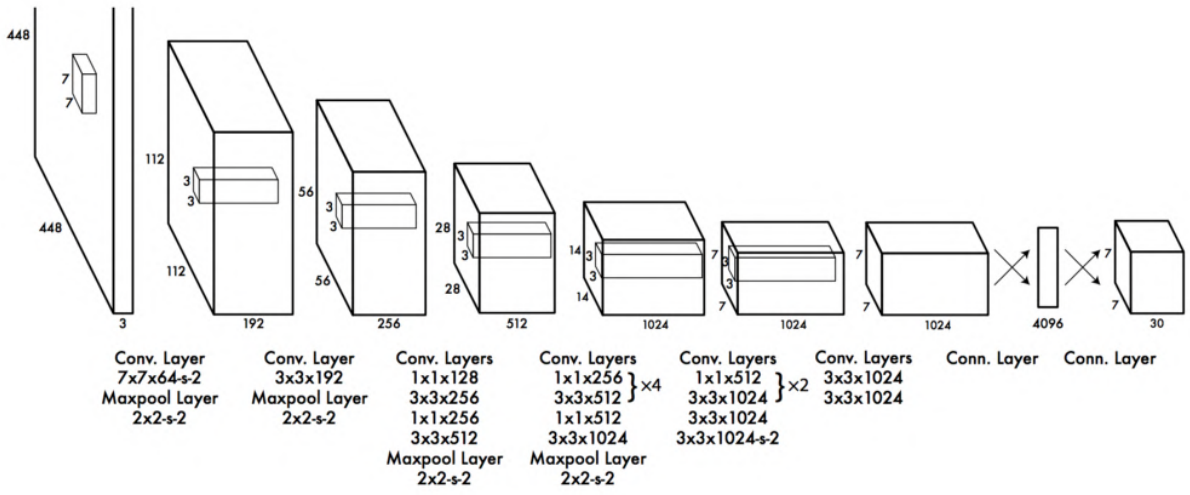


Figura 2.3: Detalhe das camadas de redes neurais que compõem a arquitetura da YOLO. Fonte: Ammar et al. (2019)

MT-CNN - Multi-Task Convolutional Neural Network

Pesquisas na área de detecção e classificação de objetos em imagens podem ser divididas em duas categorias: i) métodos baseados em regressão [Cao et al. (2014)] e ii) abordagens de ajuste de modelos [Yu et al. (2013)]. No entanto, a maioria dos métodos não correlacionam essas duas tarefas. Zhang *et al* [Zhang et al. (2016a)] propõe uma nova estrutura para integrar essas duas tarefas usando CNNs unificadas em cascata por meio de aprendizado multi-tarefa.

As CNNs propostas consistem em três etapas. No primeiro estágio, são detectadas áreas de interesse na imagem através de uma CNN superficial. Em seguida, refina-se essas áreas rejeitando um grande número que não são faces por meio de uma CNN mais complexa. Finalmente, usa-se uma CNN ainda mais complexa para refinar o resultado novamente e produzir cinco posições de fronteiras faciais. Por causa desta estrutura de aprendizagem multi-tarefa, o desempenho do algoritmo é notavelmente melhorado. Na Figura 2.4 é apresentada a arquitetura da rede Multi-Task CNN. Nesse caso, a arquitetura da Multi-Task CNN, com processos sendo realizados em paralelo, é considerada mais complexa que a arquitetura da YOLO, por exemplo.

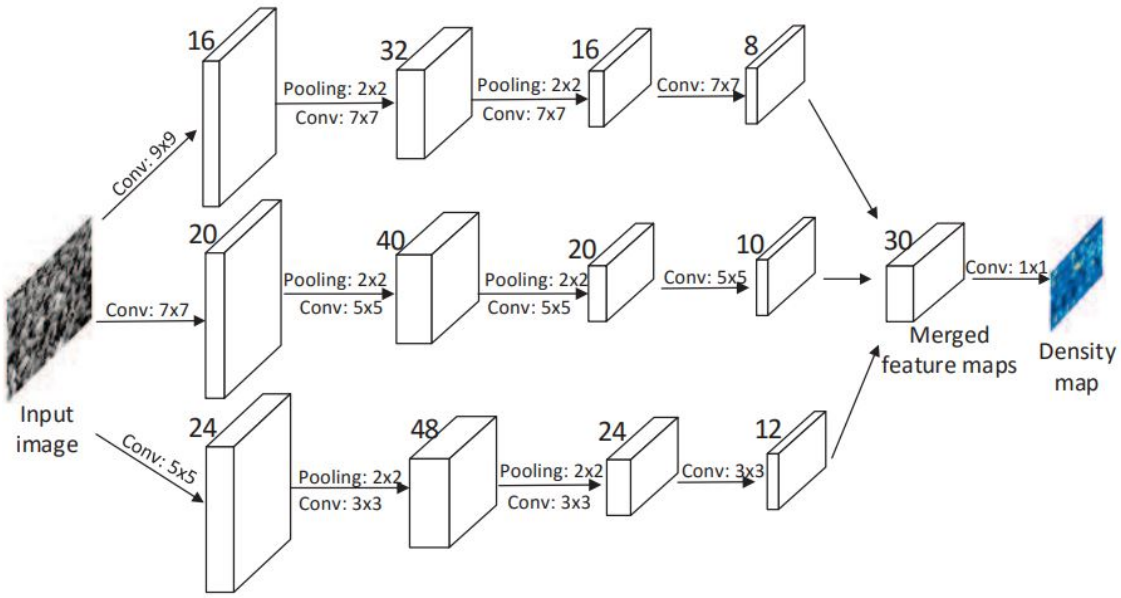


Figura 2.4: Detalhe das camadas de redes neurais que compõem a arquitetura da Multi-Task CNN. Fonte: Zhang et al. (2016a)

ResNet - Residual Network

Redes neurais convolucionais profundas levaram a uma série de avanços na classificação de imagens. Muitas outras tarefas de reconhecimento visual também se beneficiaram muito de modelos mais profundos. Assim, ao longo dos anos, há uma tendência a ir mais fundo, adicionando mais camadas as redes, na tentativa de aumentar a precisão e complexidade das tarefas que podem ser resolvidas. Mas, à medida que nos aprofundamos, o treinamento da rede neural torna-se mais difícil, a precisão começa a saturar e posteriormente se degrada. O *aprendizado residual* tenta resolver esses dois problemas.

Em geral, em uma rede neural convolucional profunda, várias camadas são empilhadas e treinadas para a tarefa em questão. A rede aprende vários padrões de nível baixo, médio e alto. Na aprendizagem residual, em vez de tentar aprender algumas características, tentamos aprender alguns resíduos. Resíduos podem ser simplesmente entendidos como subtração da características aprendidas da entrada dessa camada. A **ResNet** faz isso usando conexões de atalho (conectando diretamente a entrada de uma camada, com a entrada da camada seguinte). He *et al* [He et al. (2016)] demonstra que treinar esta forma de rede é mais fácil

do que treinar redes neurais convolucionais profundas simples e, também, que o problema de degradação da precisão é resolvido. Na Figura 2.5 é apresentada a arquitetura da rede ResNet com seu elevado número de camadas que torna necessário estratégias para contornar o problema de degradação, como discutido anteriormente. Nos experimentos realizados foi utilizado a arquitetura da ResNet101, que recebe esse nome por ser uma versão da ResNet com 101 camadas.

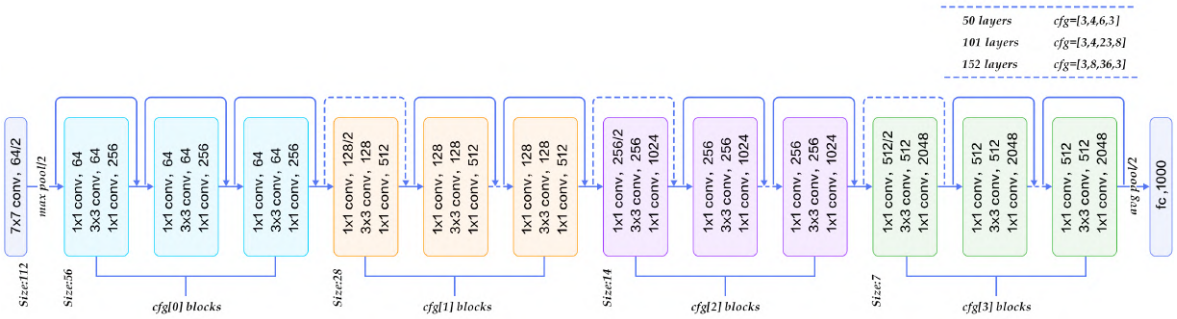


Figura 2.5: Detalhe das camadas de redes neurais que compõem a arquitetura da ResNet. Fonte: Li et al. (2017)

No Capítulo 5 serão apresentados alguns experimentos que foram realizados para avaliar a eficiência de alguns dos métodos apresentados neste Capítulo para os cenários de uso de interesse deste trabalho. No Capítulo 3 serão apresentados trabalhos que lidam com a tarefa de realizar detecção de pessoas e quais técnicas e estratégias são utilizadas.

Capítulo 3

Trabalhos Relacionados

Com o objetivo de realizar uma revisão da literatura em busca de trabalhos que se relacionem com o objetivo desta pesquisa, foi realizado uma pesquisa exploratória. Este tipo de pesquisa tem como objetivo proporcionar maior familiaridade com o problema, com intenção de torná-lo mais explícito e construir hipóteses [Gasque (2007)].

A pesquisa exploratória foi realizada em repositórios de trabalhos científicos como *ACM*, *Springer*, *IEEE Xplorer*, *Google Scholar*, entre outras bases disponíveis. Foram usadas como palavras-chave para busca os termos: *audience counting*, *people counting*, *object counting*, *public counting*, *people detection*, e também variações dos mesmos.

Não foram encontrados na literatura outros trabalhos que tenham como alvo, especificamente, a contagem automática de audiências em eventos culturais, artísticos ou esportivos como o buscado nesta pesquisa. Entretanto, uma série de iniciativas que possuem alguma relação com a investigação feita aqui, estão listados a seguir.

O trabalho de Barandiaran *et al* [Barandiaran et al. (2008)] destaca que a contagem de pessoas é um campo de pesquisa que ganhou muita atenção nos últimos anos. Há muitas câmeras de vigilância já instaladas ao nosso redor, mas, nem sempre, há meios para monitorá-las continuamente. Uma alternativa para isso é desenvolver tecnologias baseadas em visão computacional que processem essas imagens para detectar situações problemáticas ou comportamentos incomuns. O trabalho realça ainda que reconhecimento de padrões e objetos, análise de comportamentos, detecção de situações de emergência e contagem de conjuntos de elementos, assim como é o objetivo deste trabalho, possibilitam que a quantidade de dados no formato de vídeo gerados atualmente possam receber tratamento adequado.

Conte *et al* [Conte et al. (2010)] concluem que a estimativa do número de pessoas presentes em uma área pode ser uma informação extremamente útil por razões de segurança (por exemplo, uma alteração anômala no número de pessoas pode ser a causa ou o efeito de um evento perigoso) e para fins econômicos (otimizando o cronograma do sistema de transporte público com base no número de passageiros). Assim, vários trabalhos nos campos de análise de vídeo e vigilância inteligente abordam esta tarefa [Rahmalan et al. (2006)]. O trabalho de Conte *et al* propõe uma abordagem de contagem indireta. Neste caso, é utilizado um método que estima a quantidade de pessoas em uma cena baseada na quantidade de regiões de interesse na busca por pessoas em movimento. Apesar da robustez do método, a abordagem usando estimativa não preenche os requisitos como a privacidade das pessoas da audiência e a facilidade de auditora da contagem, além de não ser possível marcar uma pessoa individualmente, outro requisito deste trabalho.

Para Hou *et al* [Hou e Pang (2011)], a contagem de multidões tem sido um componente importante nos sistemas de videovigilância. Por exemplo, diferentes níveis de atenção podem ser disparados pela ocorrência, na multidão, de diferentes densidades de concentração de pessoas. Existem diferentes abordagens para contar o número de pessoas em aglomerações [Sindagi e Patel (2019)]. Há uma grande quantidade de pesquisas feitas na área de detecção humana e uma das abordagens mais utilizadas é detectar (e contar) diretamente seres humanos a partir de uma imagem ou vídeo. Porém é ressaltado que, muitas vezes, os sistemas precisam que alguns requisitos sejam obedecidos para que os resultados esperados sejam alcançados. Requisitos como: as pessoas devem estar se movendo, o fundo da imagem deve ser simples e a imagem precisa ter resolução alta. Normalmente, tais técnicas apresentam bons resultados em aglomerações de baixa e média densidades [Zhang et al. (2016b)].

No trabalho de Ryan *et al.* [Ryan et al. (2010)], os autores mencionam a contagem da quantidade de pessoas em um evento ou local como indicador chave de segurança e estabilidade. É proposto um algoritmo que utiliza rastreamento de pessoas e características do local para contabilizar a quantidade de pessoas. O método divide a cena em várias regiões através da segmentação e soma a quantidade de pessoas detectadas em cada região para chegar ao valor total. Como em outros casos, o sistema requer que as pessoas, na cena, estejam em movimento e que todo o corpo apareça. Além disso, como usa o método da remoção de

background, o fundo da imagem precisa ser simples. Apesar de alcançar bons resultados, os requisitos necessários não atende os critérios discutidos neste trabalho.

Subburaman *et al* [Subburaman et al. (2012)] utilizam um método de detecção de cabeças para quantificar pedestres em áreas públicas. O método consiste em remover o plano de fundo, representar a imagem apenas com tons de cinza, utilizar a orientação de gradiente que aponta para a região de maior interesse na imagem e usar um classificador em cascata reforçado para a detecção das cabeças. O método necessita de um pré-processamento para remoção do *background*, o que dificulta na generalização da solução e apresenta uma alta taxa de falso positivos.

Em [Visala (2014)], Visala *et al* investigam a detecção de pessoas através do dispositivo *Kinect*, da *Microsoft*. É utilizada a técnica de histograma, que faz um levantamento estatístico da ocorrência de cada valor na escala de cinza e detecta a silhueta dos objetos presentes nas imagens. Após essa etapa, o sistema usa a detecção de cabeça e ombros para rastrear as pessoas na cena. O dispositivo pode ainda ser utilizado para monitorar os movimentos das pessoas. Entretanto, o escopo da aplicação não abrange contagem de grande quantidades de pessoas.

O uso de uma Rede Neural Convolucional (CCN) profunda é proposto por Zhang *et al* [Zhang et al. (2015)] para a contagem de pessoas em grandes grupos. O treinamento da rede é feito para determinar a densidade da multidão e a quantidade de pessoas. Com esses dois dados, existe o auxílio mútuo que resulta na diminuição do erro. Adicionalmente, é implementado um método que classifica e seleciona as amostras que serão treinadas, funcionando como um pré-ajuste da rede neural. A abordagem não faz a contagem individual de pessoas e, apesar de apresentar bons resultados, seria difícil a validação e auditoria do sistema de contagem de audiências.

Como visto, uma boa parte do trabalhos de contagem de pessoas concentram-se na caracterização e rastreamento de pedestres por questões de segurança. Outras linhas de pesquisa focam na estimativa da quantidade de pessoas baseada no cálculo da densidade e usam técnicas de detecção de cabeças ou *deep learning* [Sindagi e Patel (2018)]. Os trabalhos são justificados, em sua maioria, com o objetivo de detectar padrões incomuns na movimentação de grandes grupos de pessoas, que sejam capazes de indicar algum tipo de acontecimento anômalo, o qual possa trazer riscos para as pessoas.

Outro aspecto abordado é sobre as maneiras de capturar imagens e os dispositivos que realizam essa tarefa. Esses dispositivos se tornam cada vez mais complexos e existe uma necessidade crescente de analisar estes dados. Existem diferentes métodos para aumentar a precisão dos sistemas e, para cada aplicação, diferentes técnicas podem ser usadas como *data augmentation*, combinação de mais de um técnica simultaneamente e o uso de estimativa de densidade para a contagem de objetos.

Apesar dos trabalhos apresentarem objetivos diferentes para a aplicação de contagem de pessoas, confirma-se o interesse crescente por técnicas de visão computacional para caracterização, detecção e monitoramento de objetos em imagens e vídeos. No melhor do conhecimento absorvido neste trabalho, esta é a única pesquisa que tem como foco a contagem automática de audiências presenciais.

No Capítulo 4 será apresentada a proposta de um sistema de Contagem Automática de Audiências Presencias. Também serão apresentadas a arquitetura proposta e as estratégias utilizadas em cada módulo.

Capítulo 4

Contagem de Audiências Presenciais

Neste Capítulo será apresentado o detalhamento da solução proposta para realizar Contagem de Audiências Presenciais. Inicialmente, são explicitados requisitos e premissas que a solução deve atender. Esses requisitos, baseados nos estudos apresentados na Fundamentação Teórica, são o arcabouço para que a solução tenha relevância.

Posterior a apresentação da solução proposta será detalhada a arquitetura deste sistema. O sistema possui três módulos: Módulo de Controle, Módulo de Captura e Módulo de Classificação. O detalhamento de cada módulo, os requisitos e premissas da solução e a abordagem proposta serão apresentados nas seções 4.1, 4.2, 4.3 e 4.4.

4.1 Requisitos e Premissas

Os requisitos contêm o comportamento, atributos e propriedades do sistema futuro. Portanto, o principal objetivo é garantir que eles sejam entendidos e seguidos durante o processo. O trabalho com os requisitos envolve algumas etapas, como por exemplo a identificação, análise, verificação e, finalmente, gestão. Os escopo da aplicação discutida neste trabalho possui requisitos e premissas específicos e fundamentais para relevância da solução.

Uma das premissas é que o sistema a ser desenvolvido permita simplificar o processo de auditoria. Uma das formas de atender este requisito é realizar a contagem a partir de uma única imagem, ao invés de múltiplos *frames* ou vídeos (tanto da audiência como da porta de acesso). Em caso de necessidade de uma recontagem realizada por um humano, o fato da contagem ser feita em um único *frame* facilitaria esse processo. Ainda neste sentido, as

soluções que marcam as pessoas detectadas apenas selecionando os rostos também contribuem para facilitar um possível auditoria. Para garantir que imagens e contagens geradas pelo sistema não foram adulteradas, é necessário o uso de estratégias extras, como registro em *blockchain* ou assinatura digital, que estão fora do escopo deste trabalho.

Outro requisito é a privacidade das pessoas presentes no ambiente, as câmeras são posicionadas de frente para audiência e as pessoas tem seus rostos expostos. Esse fato requer que as marcações da detecção de pessoas escondam os rostos, entrando assim no mesmo caso anterior onde a detecção das faces representa a situação ótima. A marcação das faces possibilita que os rostos das pessoas sejam cobertos nas imagens.

Por fim, um requisito primordial é que a solução seja genérica a vários ambientes. Esta condição demanda, principalmente, que todo o pré-processamento realizado nas imagens seja feito de maneira automatizada. Neste caso, o pré-processamento inclui correção de brilho, contraste, nitidez, remoção de *background*, etc. Todas essas circunstâncias demonstram a especifidade da solução abordada neste trabalho.

4.2 Sistema Automático de Contagem de Audiências Presenciais

Para atender os objetivos estabelecidos neste trabalho, é proposto um sistema para realizar a contagem de audiências presenciais. O Sistema de Contagem Automática de Audiências ou (SCAA) proposto tem como objetivo auxiliar na aferição do público real de eventos artísticos, culturais e esportivos que se caracterizem pela presença de pessoas na condição de audiência. O mecanismo de contagem proposto é baseado em três premissas básicas: i) ser auditável em qualquer tempo, ii) permitir a anonimização da audiência para fins de privacidade e iii) garantir o máximo de generalização em ambientes diversos.

O resultado de uma contagem realizada por esse sistema será uma única imagem onde cada pessoa detectada será marcada com um quadrado no rosto e um número associado a contagem. Como abordado anteriormente, cobrir o rosto das pessoas detectadas é um requisito para preservar a privacidade das pessoas presentes da audiência em questão. Já o resultado ser uma única imagem, e a utilização de números associados a contagem, são tratados desta maneira para facilitar o processo de uma possível auditagem da contagem.

Baseado nos resultados apresentados nos experimentos realizados durante este trabalho, foi escolhida como rede neural utilizado na detecção de pessoas a *Resnet 101*. A Resnet possui um tempo de resposta mais elevado quando comparado com as outras redes experimentadas, mas esse não é um critério crucial para a aplicação abordada neste trabalho. Em compensação esta rede possui resultados de acurácia e sensibilidade superiores as outras.

O sistema tem a capacidade de selecionar qual imagem será utilizada para realizar a contagem e quais imagens serão descartadas. A seleção é feita baseado na iluminação de cada imagem. Em salas de cinemas, por exemplo, por causa da iluminação da cena, *frames* em sequência podem ter uma diferença de iluminação significante.

4.3 Arquitetura do Sistema Proposto

De maneira prática, o funcionamento do sistema é baseado na captura de várias imagens durante a realização do evento, no cálculo da média, piso e teto do público presente e na comparação de tais resultados com o público real aferido. Tal comparação pode ser configurada para disparar alertas quando determinados limiares de diferença forem encontrados, permitindo uma intervenção direta do auditor apenas nos casos com potencial desvio relevante. Para apoiar a análise do auditor, o SCAA recupera tanto as informações de público real quanto as imagens do evento em pauta. O protótipo do sistema foi desenvolvido tendo como alvo o contexto de cinemas, embora a plataforma possa ser usada, com pequenas adaptações, em outros cenários como teatros, ginásios, auditórios etc.

Na imagem 4.1 é apresentada a arquitetura proposta para o sistema discutido neste trabalho. A arquitetura mostra como funciona a interação entre os módulos.

A arquitetura do piloto do SCAA desenvolvido para cinemas possui três componentes principais, como descrito a seguir.

4.3.1 Módulo Classificador (MCla)

A *engine* de contagem, propriamente dita, recebe como entrada imagens de público em eventos e faz a identificação e contagem das pessoas presentes no evento. O MCla usa um algoritmo baseado em inteligência artificial e foi treinado usando um conjunto de centenas de milhares de imagens. A sua execução é na retaguarda, podendo ser compartilhado por múl-

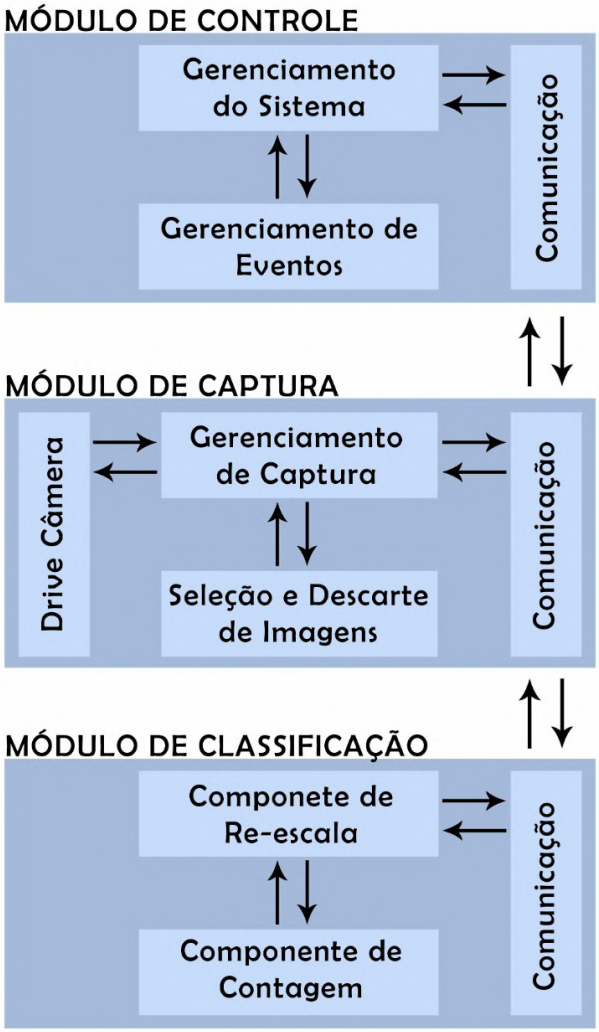


Figura 4.1: Arquitetura do Sistema de Contagem com interação entre os Módulos de Controle, Captura e Classificação.

tiplos eventos. A centralização desse módulo, deve-se ao fato de ser recomendado o uso de placas de vídeo dedicadas para aplicações com *Deep Learning*. Tais *hardwares* representam a maior parte do custo envolvido no desenvolvimento do sistema.

Quando acionado pelo Módulo de Controle (MCon), o Módulo Classificador (MCla) recebe uma imagem obtida pelo Módulo de Captura (MCap) para realizar a contagem. Essa imagem pode ser redimensionada caso seja configurado, essa etapa pode ser realizada para padronizar as imagens que serão classificadas. Com os pesos da rede neural profunda anteriormente treinados, o Módulo Classificador realiza a detecção e posteriormente a contagem de pessoas. A saída do Classificador é um conjunto de *bounding boxes*, área quadrada que

delimita o objeto detectado.

As imagens capturadas nas situações abordadas neste trabalho, possuem uma característica em comum. Com a câmera posicionada na frente da audiência, as pessoas localizadas próximas a tela ou palco aparacerão na imagem com um tamanho maior (mais *pixels*) que as pessoas localizadas mais distantes da câmeras. Esse padrão pode ser observado na Figura 4.2 e possibilita uma primeira etapa de refinamento dos resultados. É possível identificar e remover as ocorrências de falso positivos mais grosseiros, que fogem mais desse padrão. As estratégias para tratar esses casos serão discutidas no Capitulo 5.

Figura 4.2: Exemplo de imagem característica de cinemas, teatros e auditórios. Fonte: Ognian Mladenov©

4.3.2 Módulo de Captura (MCap)

O módulo de captura é executado no local do evento e é responsável por controlar a coleta de imagens pelas câmeras, em uma frequência configurável para cada evento. O MCap também garante a integridade da origem, localização, data e hora das imagens até a sua entrega ao Módulo de Controle.

O processo de captura das imagens é uma etapa crítica do sistema, como será discutido no Capítulo 5. Em ambientes e momentos com menor luminosidade, a eficiência de todas as

abordagens verificadas sofrem um impacto relevante. Em sessões de cinema, por exemplo, o sistema propõe monitorar a quantidade de pessoas durante todo o evento. Isso implica invariavelmente, na necessidade de processar imagens quando a única luminosidade do ambiente é proveniente do reflexo da tela. Nessa conjuntura, é relevante obter a imagem de quando a tela reflete a maior intensidade de luz possível (a depender do cena do filme). Em conformidade com os parâmetros de captura estabelecidos pelo Módulo de Controle, o Módulo de Captura monitora constantemente a luminosidade da tela em busca de imagens boas.

4.3.3 Módulo de Controle (MCon)

O Módulo de Controle é o componente de integração, administração e operação do sistema, interagindo com o Módulo de Captura e com o Módulo Classificador. O MCon possui configurações para serem usadas pelas instâncias do MCla para obter parâmetros de configuração para operação (frequência de coleta em cada evento, frequência de envio etc), para entrega das imagens capturadas e sinalização de atividade (sondas). Ele também possui configuração de parâmetros para pré-processamento e armazena as imagens e contagens realizadas.

São funções deste módulo iniciar um processo de contagem em um evento, passar as configurações de parâmetros para o módulo de captura e o módulo de classificação, finalizar o processo de contagem e informar os resultados obtidos. Este módulo também será responsável, quando necessário, em se comunicar com agentes externos de controle de auditoria dos eventos inspecionados.

4.4 Prototipação

O protótipo desenvolvido foi implementado utilizando a linguagem de programação *Python*. Essa escolha deve-se ao fato de existir ferramentas para essa linguagem que auxiliam no desenvolvimento do sistema. Foram usados bibliotecas como *OpenCV* (para aplicação de métodos de processamento de imagem e visão computacional) e *Keras* para a classificação utilizando rede neural.

No sistema desenvolvido, são configurados no Módulo Classificador um *id* para o evento que vai se iniciar, uma *tag* para mais detalhes sobre o evento, parâmetros sobre o número de contagens que serão realizados naquele evento e o número de *frames* capturados por

frequência. Pode-se ainda configurar parâmetros de captura que serão executado pelo Módulo de Captura. O Módulo Classificador recebe a informação de contagem e a imagem final e armazena em sua base de dados.

No incio determinado do evento, o Módulo de Captura é inicializado automaticamente e começa a executar sua rotina pré-determinada. O MCap aciona uma câmera posicionada dentro da sala onde a evento acontece, que captura as imagens e retorna para o módulo. O MCap então repassa as imagens para o componente de seleção de descarte, que baseados em métricas como nitidez e iluminação seleciona a melhor imagem para classificação.

O Módulo Classificador recebe uma imagem do Módulo de Captura para que seja realizada a contagem de pessoas na cena. Antes de passar pelo Classificador/ Detector em si, a imagem recebida passa por um componente de re-escala que altera as dimensões da imagem para um tamanho que facilite a detecção de pessoas pelo componente de contagem. O componente de contagem é o classificador *Deep Learning* que é composto por uma rede neural profunda, com os pesos previamente treinados, que recebe a imagem, faz detecção de pessoas e retorna a imagem com a marcação de *bounding boxes* em cada uma delas.

A arquitetura do sistema proposto, com é composta por 3 módulos, foi definida com o objetivo de atender as premissas e requisitos apresentada neste Capítulo. No Capitulo 5 serão apresentados os experimentos realizados para avaliar as estratégias de contagem investigadas e os experimentos realizados para validar o sistema proposto como um todo.

Capítulo 5

Metodologia e Planejamento Experimental

Para alcançar os resultados esperados no desenvolvimento da solução proposta foi utilizada a **Metodologia de Construção**. Uma metodologia de pesquisa desse tipo consiste na construção de uma solução - um artefato ou sistema de software - para demonstrar que é possível alcançar os objetivos. Para ser considerado pesquisa, a construção da solução deve ser nova ou incluir novos recursos que não foram demonstrados anteriormente em outros artefatos. [Elio et al. (2011)]

Uma etapa importante no processo de construção são as fases de *Verificação* e *Validação*. Nesse contexto, os testes dos artefatos construídos constituem a verificação se o sistema consegue realizar todo o processo de contagem de forma correta e sem que ocorram erros no processo. Já a etapa de Validação consiste no planejamento e execução de experimentos e análise dos resultados obtidos.

O primeiro componente do protótipo construído foi o Módulo de Classificação que corresponde ao núcleo principal, o *core* da solução. A escolha se deve ao fato de que este módulo pode ser testado e avaliado independentemente dos outros[1]. Posteriormente, foi desenvolvido o Módulo de Captura e, por fim, o Módulo de Controle.

Como visto anteriormente, a prospecção de tecnologias aplicáveis indicou que o Módulo de Classificação poderia ser implementado usando uma abordagem mais tradicional, utilizando visão computacional, ou através de recursos de inteligência artificial, sobretudo

[1]Neste caso, os testes foram realizados sobre imagens estáticas coletadas previamente.

técnicas de *deep leaning*. Neste sentido, foi realizada inicialmente uma avaliação independente da aplicabilidade de tais estratégias no contexto de contagem automática de audiências presenciais.

5.1 Avaliação de Técnicas de Visão Computacional

Com o objetivo de fazer uma investigação exploratória onde buscava-se métodos que realizassem a contagem automática de audiências presenciais, foram realizados experimentos com técnicas de Visão Computacional Clássica. A escolha inicial por essas técnicas menos complexas justifica-se pelo fato de que o contexto da contagem automática de audiência em cinemas e teatros, normalmente, se refere à cenários estáticos (pessoas sentam em cadeiras com lugares fixos) e delimitado (as dimensões e condições de iluminação são fixas e conhecidas previamente).

5.1.1 Contagem por Detecção de Bordas

Para avaliar a contagem de pessoas utilizando o método de Detecção de Bordas, foi usado o algoritmo de *Canny Detection* [Qian e Huang (1996)]. Entre outros parâmetros, é possível estabelecer o limiar desejado (*thresholding*) de detecção do contraste entre o fundo da imagem e o objeto em primeiro plano.

Na Figura 5.1 é apresentada a implementação do algoritmo de detecção de bordas utilizando os valores de limiar que melhor se aplicaram as imagem. A técnica se mostra eficiente para imagens com pouco detalhamento, porém para imagens com muitos objetos existem limitações claras. Como pode ser observado na Figura 5.1, o resultado do processo de detecção de bordas não torna possível a contagem, não sendo possível identificar cada pessoa presente na imagem.

Os resultados se mostram impróprios para os objetivos deste trabalho uma vez que não foi possível realizar a segmentação e contagem de pessoas. A técnica também não mostrou potencial para a melhora dos resultados com estratégias auxiliares.

Figura 5.1: Resultado da técnica de detecção de bordas para contagem volumétrica. Foto original por TEDx Monterey©

5.1.2 Contagem por Reconhecimento Facial com Classificadores

Também foram realizados experimentos com outra técnica de Visão Computacional Clássica utilizando classificadores Haar e LBP.

Quando é necessário fazer a contagem volumétrica em imagens com mais detalhes e mais pessoas, esses classificadores têm seu rendimento afetado negativamente. Na figura 5.2 é mostrado o resultado da aplicação desse método para detectar faces de pessoas, e é possível perceber a presença de vários falsos positivos.

Nesse contexto, um falso positivo ocorre quando o classificador identifica um objeto que não existe naquela posição. Por isso, foi usada uma relação entre o tamanho do objeto detectado e a posição do objeto na imagem para tentar eliminar parte das ocorrências de falsos positivos, mas poucas ocorrências foram descartadas. Por fim, é possível perceber que uma quantidade significativa de rostos não foi detectada. Em termos gerais essa técnica apresentou acurácia de 54% e sensibilidade de 29%. Acurácia e sensibilidade são métricas de avaliação da precisão e eficiência das técnicas avaliadas e serão discutidas nas próximas seções. A diferença principal entre elas é que acurácia não leva em consideração erros por falso positivo e a sensibilidade as leva em consideração.

Outra análise das técnicas de contagem volumétrica através de Visão Computacional Clássica foi feita usando classificadores HOG Tan et al. (2013) com resultados igualmente

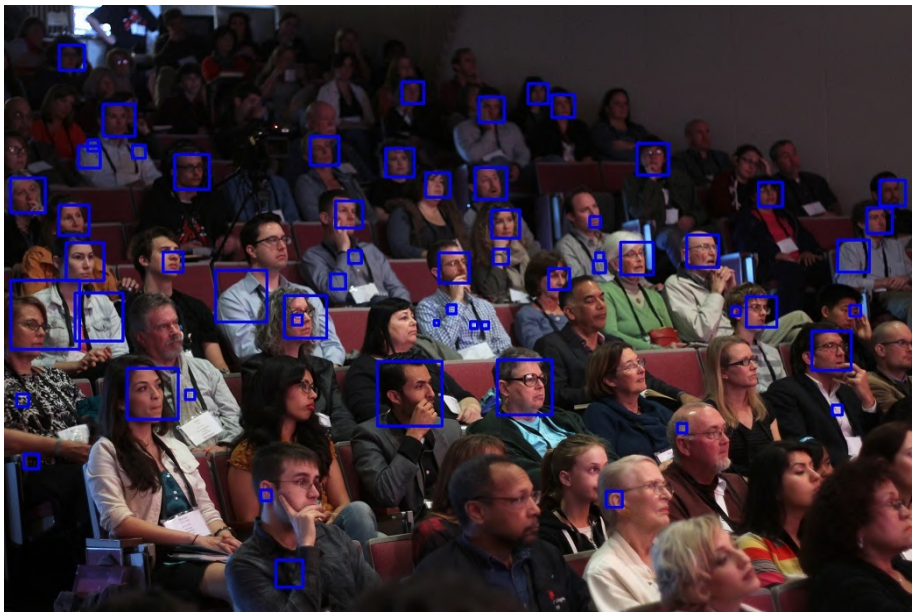


Figura 5.2: Resultado da detecção de objetos com classificadores Haar e LBP. Foto original por TEDx Monterey©

insatisfatórios. Os dados de acurácia e sensibilidade foram nos mesmos patamares da contagem feita com classificadores Haar e LBP. A técnica com uso de classificadores se mostra mais propícia para imagens com menos detalhes, com pessoas em um mesmo plano e em menor número, não sendo compatível com o objetivo deste trabalho.

Após essa prospecção inicial das estratégias tradicionais de visão computacional, ficou evidente que, pelos resultados apresentados, uma contagem automática de audiência eficiente e que identifique individualmente cada pessoa em uma única imagem, demandava abordagens diferentes.

Isso talvez se explique por que, embora o valor de acurácia das técnicas de Visão Computacional Clássica pode, em alguns casos, ser semelhante ou melhor que algumas arquiteturas de *deep learning*, os valores de sensibilidade (crítico nessa classe de aplicação) são abaixo do esperado.

Neste sentido, os desafios impostos pelos ambientes com baixa iluminação, visada obstruída e rostos fragmentados, presentes nos cenários objeto do estudo (como plateias de cinemas e teatros), levaram à busca por técnicas que pudessem se adaptar em tais condições. O novo caminho escolhido neste ponto da investigação foi avaliar a aplicação de técnicas de aprendizado de máquina no problema em pauta.

5.2 Avaliação de Técnicas de *Deep Learning*

O surgimento de *Deep Learning* trouxe avanços importantes na tarefa de reconhecimento de padrões. Cada vez mais é possível resolver problemas complexos sem a necessidade da intervenção humana. Entretanto, pra alcançar bons resultados, o processo requer uma quantidade relevante de dados.

No contexto em pauta, existem diversas bases de dados de imagens disponíveis para o treinamento das redes neurais usadas em *deep learning*. A **Pascal VOC** [Everingham et al. (2015)], por exemplo, possui registro de 20 classes e 27.450 amostras de faces distribuídas em 11.530 imagens. Na classe 'pessoas' desta base, as anotações são feitas demarcando toda área em que cada pessoa aparece na imagem. Essa característica ajuda na contextualização dos objetos, mas torna a classificação de faces mais complexa.

Outra base bastante relevante, a **WIDER-FACE** [Everingham et al. (2015)], possui 393.703 amostras de faces em 32.203 imagens divididas em 61 classes. Cada classe descreve ações/situações em que as pessoas aparecem, mas para essa aplicação apenas interessa as anotações das posições das pessoas na imagem. Nesse banco de imagens, ao contrário de outras bases, as amostras consideram apenas as faces das pessoas, como é mostrado na Figura 5.3. Essa característica torna a WIDER-FACE bastante interessante para a aplicação neste estudo, quando consideramos que, em cenários de audiência, na maioria das vezes as pessoas estão apenas com a face e parte do torso visíveis e com o restante do corpo encoberto.

5.3 Estratégias e Heurísticas Adotadas

Para a avaliação das redes neurais listadas no Capítulo 2: *YOLO, MT-CNN e ResNet*, e de outras arquiteturas de reconhecimento de objetos em imagens, foi usado a técnica de aprendizado supervisionado. Essa metodologia requer que cada amostra do banco de dados usada no treinamento seja rotulada com a classificação esperada.

Na contagem automática de audiência, os objetos a serem classificados possuem características particulares. As pessoas presentes em cinemas, teatros, e outras formas de audiência, na maioria dos casos, estão apenas com a parte superior do corpo à mostra. Além disso, devido a inclinação das pessoas nas imagens capturadas nesse tipo de ambiente, também há

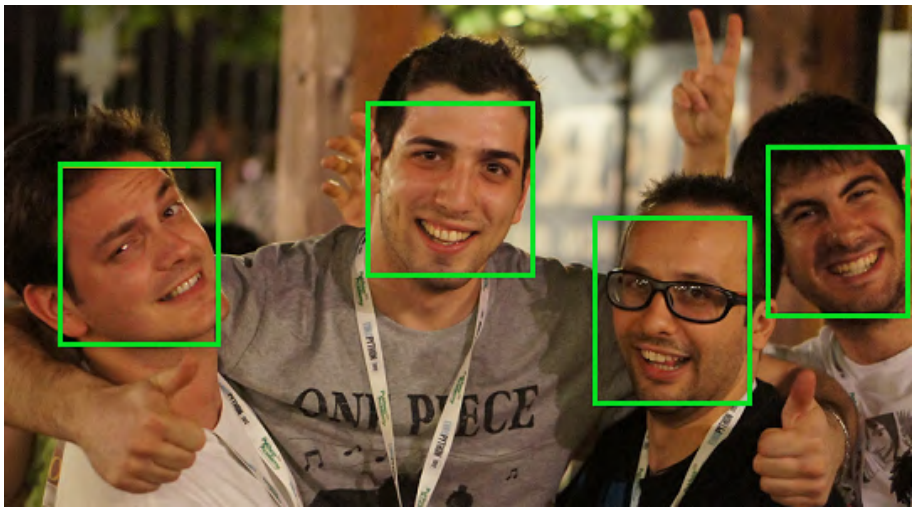


Figura 5.3: Exemplo de imagem semelhante as usadas no banco de imagens WIDER-FACE. Foto original por Michael Fötsch©

grande variação na dimensão das amostras, que são maiores no primeiro plano e menores no fundo. Isso já aponta para uma melhor adequação de *deep learning* para o tipo de problema em pauta, pois essas variações de características não precisam ser extraídas manualmente. Cada camada de rede procura por padrões que caracterizem o objeto e vai ajustando seus parâmetros para melhorar a classificação.

Quando monta-se uma base de imagens que vai ser usada em algum treinamento supervisionado, as características das amostras devem ser levadas em consideração. Para banco de imagens usados para detecção e classificação de objetos, a posição do objeto e o nível de iluminação da cena são características relevantes [LeCun et al. (2004)].

No caso específico de contagem volumétrica, a quantidade de objetos na mesma cena também é um parâmetro importante [Rodriguez et al. (2011)]. A densidade de objetos pode ser um elemento dificultador para alguns modelos de aprendizagem mais que para outros e, em imagens com alta densidade, existe a probabilidade de objetos muito pequenos dificultarem a detecção.

5.3.1 Treinamento e Validação

As três redes selecionadas (*YOLO, MT-CNN e ResNet*) foram treinadas usando a base de dados WIDER-FACE. Para os treinamentos dos modelos, a base foi separada randomicamente

em duas partes[2]: 70% das imagens da WIDER-FACE foram utilizadas para treinamento, enquanto que as outras 30% foram reservadas para validação. Cada abordagem redimensiona a imagem de entrada para o tamanho que melhor se encaixa com sua arquitetura, alguns usam dimensões fixas, outros variam durante o treinamento.

O processo de validação (testes realizados durante o treinamento) é um etapa importante para prevenir o problema de *overfitting*. Esse problema ocorre quando o modelo se especializa demais na base de treinamento mas a taxa de acertos, para amostras sobre as quais ele não treinou, começa a diminuir [Hawkins (2004)]. Como estratégia para prevenir o *overfitting*, os treinamentos foram interrompidos quando o erro de validação não diminuiu após duas iterações completas sobre a base de treinamento (épocas). Como consequência, cada treinamento de cada rede teve um número de épocas diferente.

Os treinamentos e testes foram feitos com o auxílio de uma *GPU NVidia GeForce 940mx* e ferramentas de computação em nuvem. Os modelos de *deep learning* foram implementados na ferramenta **Keras** [Chollet et al. (2015)], uma API de alto nível usada para desenvolvimento de redes neurais artificiais.

Foi cogitado o uso de *data augmentation* e estratégias de pré-processamento, como a equalização de histograma, para serem usados como métodos de aumentar a generalidade do base e diminuir a presença de ruídos nas imagens, respectivamente. O *data augmentation* realizou mudanças de escala, brilho, cisalhamento e rotação nas imagens. Os primeiros testes não apresentaram melhorias significativas nos resultados de acurácia durante o treinamento das redes. Na base de validação, a melhoria de acurácia não ultrapassou a casa do 1%. Não foi possível verificar o impacto na base de testes. No fim, optou-se pelo não uso dessas técnicas levando em consideração o custo benefício. O fato da base Wider-Face ter um número de imagens consideravelmente elevado e já incluindo uma variedade significativa nas imagens, contribuíram para esse cenário.

[2]Para bases muito grandes, o ponto onde essa divisão é feita não é um elemento crítico, sendo respeitada normalmente a premissa de que a base de treinamento seja maior do que a base de validação Yang et al. (2016).

5.4 Planejamento dos Experimentos

5.4.1 Métricas de Interesse

No processo de desenvolvimento de um sistema baseado em *deep learning*, assim como em outros tipos de sistemas, é necessário mensurar a qualidade de acordo com o objetivo da tarefa. Existem funções matemáticas que auxiliam a avaliar a capacidade que o sistema possui de cometer erros e acertos. Essas métricas são importantes na avaliação, diagnósticos e comparação entre sistemas.

Entre as métricas de performance de sistemas de classificação, uma comumente usada é **acurácia** [Williams et al. (2006)]. De forma simples, acurácia é a proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo, ou seja, o número de acertos (positivos) dividido pelo número total de exemplos. No caso do estudo descrito neste trabalho, a acurácia foi calculada pela razão da quantidade de pessoas que o modelo *deep learning* detectou em cada imagem dividido pela quantidade real total de pessoas (contado de forma manual anteriormente). A quantidade real de pessoas é baseado na contagem realizada por humanos, esse tipo de contagem possui um erro aproximado de 5% [Russakovsky et al. (2015)]. Essa métrica é uma indicação do resultado que o sistema deve apresentar quando estiver sendo usado em campo. A equação para cálculo da acurácia é apresentada a seguir:

$$Acurcia = \frac{VerdadeiroPositivo+VerdadeiroNegativo}{VerdadeiroPositivo+VerdadeiroNegativo+FalsoNegativo+FalsoPositivo}$$

Usar a acurácia como única métrica de avaliação pode gerar uma falsa sensação de precisão, já que esse tipo de cálculo pode contabilizar, em algumas situações, um falso positivo como acerto. Uma métrica que resolve esse problema, também usado com frequência, é a **sensibilidade**.

Sensibilidade [Garcia et al. (2008)] pode ser definida como percentual de classificações realizadas corretamente pelo modelo. No caso deste trabalho, é a probabilidade de um objeto ser classificado como pessoa, dado que realmente é uma pessoa. Ou seja, ela representa a capacidade do sistema em predizer corretamente uma condição para os casos em que realmente ela existe. Essa métrica fornece uma possibilidade de diagnóstico mais detalhado que pode ser útil para ajustes e correções no sistema. A equação para cálculo da sensibilidade é apresentada a seguir:

$$Sensibilidade = \frac{VerdadeiroPositivo}{VerdadeiroPositivo+FalsoNegativo}$$

5.4.2 Execução dos Testes

Para a realização dos testes de acurácia e sensibilidade das três redes na contagem automática de audiências, uma base específica para testes[3] foi montada de forma que melhor representasse o contexto alvo.

Neste sentido, foi construído um banco de imagens com o propósito específico de realizar a contagem automática de pessoas em uma audiência. As imagens selecionadas foram obtidas através de mecanismos de busca na internet e, em sua maioria, mostram um grupo de pessoas sentadas em um ambiente fechado como, por exemplo, a imagem mostrada na Figura 5.4.

Figura 5.4: Exemplo de imagem usada no banco de imagens desenvolvido para este trabalho. Foto original por Bartek Barczyk©

O banco de dados de testes conta com 75 imagens e 3.206 amostras de faces e foi usado para realizar a aferição da acurácia e sensibilidade da contagem. A contagem de pessoas foi feita manualmente para cada imagem selecionada.

Para uma melhor avaliação do desempenho das três redes treinadas, cada imagem da base de testes foi classificada com relação à posição das faces em relação ao ângulo da câmera, com relação ao tipo de iluminação da cena e também a quantidade de pessoas presentes.

[3]Diferentemente das bases de treinamento e validação, a base de testes não é usada na construção dos modelos apenas na sua aferição.

Dentro do possível, para o agrupamento das imagens, os intervalos categóricos foram definidos com o intuito de balancear o número de amostras em cada classe.

Para a posição das faces, foram considerados três classes: i) **frontal**, quando todas (ou a maioria) as pessoas estão de frente para o ângulo da imagem; ii) **lateral**, quando todas (ou a maioria) as pessoas estão de lado para o ângulo da imagem; e iii) **híbrida**, quando não existe uma maioria significativa de pessoas de frente ou de lado para o ângulo da imagem. A classe de imagens classificada como frontal possui 54 imagens, a lateral possui 15 e a híbrida possui 6.

Para a classificação da iluminação da imagem, foram usados apenas dois valores[4]: **clara**, quando todas (ou a maioria) as pessoas estão com os rostos iluminados) ou **mista**, quando não existe uma maioria significativa de pessoas com rostos iluminados. A classe de imagens classificada como clara possui 64 imagens e a mista possui 11.

A última classificação foi a quantidade de pessoas, para a qual foram usados três intervalos categóricos: de 0 a 25 pessoas, de 26 a 50 pessoas e de 51 a 200 pessoas. O grande intervalo da última classe (de 51 a 200 pessoas) deve-se a dificuldade de encontrar imagens de qualidade mínima aceitável com mais de 100 pessoas. A classe de imagens com 0 a 25 pessoas possui 27 imagens, a classe com 26 a 50 pessoas possui 23 e a classe com 51 a 100 pessoas possui 25.

5.4.3 Experimento Controlado

Na busca por mais material e dados para análise, foi realizada a coleta de imagens em um experimento realizado no auditório do Centro de Tecnologia da Universidade Federal da Paraíba. O objetivo era obter um conjunto de imagens de pessoas formando uma audiência em um ambiente controlado. Os documentos autorizando o uso do local para realização do experimento e a autorização dos participantes no uso do material colhido para fins acadêmicos são apresentados no Apêndice A.

Esse experimento tem o objetivo de consolidar o resultado dos experimentos anteriores

[4]Nas imagens selecionadas para o banco de imagens em questão, nenhuma apresenta a totalidade, ou maioria, das pessoas com rosto com pouca iluminação. Isso deve-se ao fato de que as pessoas localizadas mais à frente, geralmente, tem seu rosto iluminado pelo *flash* da câmera, pela iluminação da tela (em cinemas), ou pela iluminação regular do local.

para rede ResNet. A escolha da rede ResNet para o a solução apresentada neste trabalho se dá pelo desempenho desta rede apresentada nos experimentos anteriores. Essa seção apresentará detalhes desse experimento e seus resultados.

Experimento Fatorial 2^k

Na estatística, um experimento fatorial é um experimento no qual o planejamento possui dois ou mais fatores, cada um com valores discretos ou representado em níveis, e cujo o conjunto de todos os testes assumem todas as combinações possíveis desses níveis em todos esses fatores. Esse tipo de experimento permite que se estude o efeito de cada fator na resposta, assim como os efeitos das interações entre os fatores. Para a grande maioria dos experimentos fatoriais, cada fator possui apenas dois níveis. Com isso o número de unidades experimentais é calculado por 2^k, onde K é o número de fatores analisados. [Calado (2003)]

O número de fatores de um experimento precisa ser cuidadosamente decidido. Se esse número for pequeno pode não existir representatividade do experimento com o cenário real. E caso esse número seja grande, pode tornar a realização do experimento inviável. Como padrão são designados valores de -1 e 1 para cada nível de cada fator.

Figura 5.5: Exemplo de imagem capturada durante experimento no auditório do Centro de Tecnologia da UFPB

Para este experimento foram selecionados 5 fatores que, diante de experimentos anteriores, mostraram demandar mais atenção. Segue detalhamentos destes fatores:

1. Fator X1 - Posição da Câmera: sendo o nível -1 a câmera da posição lateral e o nível 1 a câmera na posição frontal;

2. Fator X2 - Nível de Aglomeração das pessoas: sendo o nível -1 as pessoas sentadas agrupadas (perto umas das outras) e nível 1 as pessoas sentadas espalhadas (distante umas das outras);

3. Fator X3 - Distância das pessoas para a Câmera: sendo o nível -1 as pessoas sentadas distantes da câmera (no fundo do auditório) e o nível 1 as pessoas sentadas próximas a câmera (na frente do auditório);

4. Fator X4 - Iluminação da Imagem: sendo o nível -1 imagens com baixa iluminação (o brilho da imagem foi reduzido em 65%) e o nível 1 a imagem com sua iluminação total;

5. Fator X5 - Resolução da Imagem: sendo nível -1 imagens com 853 x 480 *pixels* e o nível 1 imagens com 1280 x 720 *pixels*;

Outro fator que havia interesse em ser investigado seria o número de pessoas presentes na imagem, também classificado em dois níveis. Porém por questões logísticas, disponibilidade do auditório e número de pessoas que se dispuseram a realizar o experimento, não foi possível obter material para analisar esse fator.

Na Figura 5.5 é mostrada uma imagem exemplo, obtida durante o experimento. A imagem é um exemplo de captura com câmera frontal, com as pessoas sentadas espalhadas, próximas a câmera, com iluminação total e resolução de 1280 x 720 *pixels*.

Na Figura 5.6 é mostrada um exemplo de uma contagem realizada na imagem capturada durante experimento. A imagem é um exemplo de uma captura com câmera lateral, com as pessoas sentadas espalhadas, distante da câmera, com iluminação total e resolução de 1280 x 720 *pixels*.

Na Tabelas 5.1 e 5.2 são apresentados os dados do Experimento Fatorial 2^k. O campo teste indica o número do experimento na ordem em que foi executado. O campo Fator de

Figura 5.6: Exemplo de resultado da contagem na imagem capturada durante experimento no auditório do Centro de Tecnologia da UFPB

Controle indica qual fator a coluna se refere e o nível do fator naquele experimento. O campo Resultado indica o resultado da sensibilidade da contagem de pessoas realizada na imagem em uma escada de 0 à 1. As tabelas foram divididas em 2 para melhor apresentação. A tabela 5.1 apresenta os dados dos experimentos 1 à 16 e a tabela 5.2 apresenta os dados dos experimentos 17 à 32.

O detalhamento da Metologia, heurísticas e estratégias de treinamento são necessários para melhor compreensão dos resultados encontrados. No Capítulo 6 serão apresentados e discutidos os resultados nos experimentos realizado neste capítulo.

Tabela 5.1: Tabela Experimento Fatorial 2^k (Testes do 1-16)

Teste	**Fator de Controle**					**Sensibilidade**
	X1	**X2**	**X3**	**X3**	**X5**	
1	-1	-1	-1	-1	-1	0,676
2	1	-1	-1	-1	-1	0,471
3	-1	1	-1	-1	-1	0,706
4	1	1	-1	-1	-1	0,618
5	-1	-1	1	-1	-1	0,971
6	1	-1	1	-1	-1	0,971
7	-1	1	1	-1	-1	0,971
8	1	1	1	-1	-1	1,000
9	-1	-1	-1	1	-1	0,794
10	1	-1	-1	1	-1	0,647
11	-1	1	-1	1	-1	0,735
12	1	1	-1	1	-1	0,735
13	-1	-1	1	1	-1	0,971
14	1	-1	1	1	-1	0,971
15	-1	1	1	1	-1	1,000
16	1	1	1	1	-1	1,000

Tabela 5.2: Tabela Experimento Fatorial 2^k (Testes do 17-32)

Teste	**Fator de Controle**					**Sensibilidade**
	X1	**X2**	**X3**	**X3**	**X5**	
17	-1	-1	-1	-1	1	0,971
18	1	-1	-1	-1	1	0,794
19	-1	1	-1	-1	1	0,882
20	1	1	-1	-1	1	0,912
21	-1	-1	1	-1	1	1,000
22	1	-1	1	-1	1	0,971
23	-1	1	1	-1	1	1,000
24	1	1	1	-1	1	1,000
25	-1	-1	-1	1	1	0,971
26	1	-1	-1	1	1	0,794
27	-1	1	-1	1	1	0,941
28	1	1	-1	1	1	0,971
29	-1	-1	1	1	1	1,000
30	1	-1	1	1	1	1,000
31	-1	1	1	1	1	0,971
32	1	1	1	1	1	1,000

Capítulo 6

Resultados e Análise

Neste Capítulo serão apresentados e analisados os resultados dos experimentos realizados e descritos no Capítulo anterior além de um experimento realizado em campo. Os resultados são baseados nas métricas de sensibilidade e acurácia já mencionados.

Os experimentos foram separados em três conjuntos de cenários para análise do impacto desses casos no resultado apresentado pelo sistema proposto: i) diferentes posições dos rostos da audiência em relação a câmera de aquisição das imagens; ii) com diferentes intensidades de iluminação; e iii) com variação da quantidade de pessoas nas imagens. Os resultados deste experimento são detalhados na Seção 6.1. Na Seção 6.2, serão apresentados e analisados os resultados do experimento controlado no modelo Fatorial 2^k. Na Seção 6.3 será apresentado o Teste de Campo que foi realizado em um sessão de cinema real. Por fim, na Seção 6.4 é realizado uma discussão sobre os resultados obtidos.

6.1 Apresentação dos Resultados

Os resultados dos experimentos realizados nos três cenários previstos e também de forma geral são apresentados a seguir.

Na Tabela 6.1 é apresentado resultado total de acurácia e sensibilidade das redes YOLO, MT-CNN e ResNeT para todas as imagens do banco de imagens de teste com imagens retiradas da *internet*. Como pode ser observado, o modelo baseado em ResNet obteve percentuais médios de acurácia e sensibilidade relevantes, com valores acima de 96%. Estes valores superam todas as demais técnicas avaliadas neste estudo para a contagem automática, in-

Tabela 6.1: Acurácia e sensibilidade das redes YOLO, MT-CNN e ResNet.

Resultado Total de Acurácia e Sensibilidade das Redes		
	Acurácia	Sensibilidade
YOLO	56,6%	49,4%
MT-CNN	55,5%	55,3%
ResNet	98,2%	96,5%

Tabela 6.2: Acurácia das redes por posição das pessoas em relação ao ângulo da imagem.

Resultado da Acurácia por Posição das Faces			
	Frontal	Lateral	Híbrida
YOLO	64,9%	34,9%	62,7%
MT-CNN	71%	21,3%	42,7%
ResNet	99,3%	94,2%	100%

cluindo as tradicionais, baseadas em técnicas clássicas de visão computacional e também as outras duas baseadas em *deep learning*. Quanto ao desempenho, com os testes executados no *hardware* de referência apresentado anteriormente, a YOLO apresentou o menor tempo de processamento para fazer a contagem com uma média de 0,329 segundos, seguido pela MT-CNN com 1,112 segundos e ResNet com 24,1 segundos. O fato que justifica a grande diferença de tempo de processamento entre a ResNet e as outras é que a ResNet possui um número de camadas significativamente maior.

Nas Tabelas 6.2 e 6.3 são apresentados os resultados de acurácia e sensibilidade, respectivamente, para as imagens classificadas por posição dos rostos. As redes YOLO, MT-CNN e ResNet foram avaliadas em três cenários com pessoas com os rostos predominantemente frontais em relação a câmera, com rostos predominantemente laterais e imagens com ambos os casos.

Nas Tabelas 6.5 e 6.4 são apresentados os resultados de acurácia e sensibilidade para as imagens classificadas por iluminação da cena. As redes foram avaliadas em imagens claras (maioria dos rostos iluminados) e mista (número semelhante entre rostos iluminados e não iluminados).

Tabela 6.3: Sensibilidade das redes por posição das pessoas em relação ao ângulo da imagem.

Resultado da Sensibilidade por Posição das Faces			
	Frontal	Lateral	Híbrida
YOLO	56,4%	30%	59,1%
MT-CNN	70,7%	21,1%	42,7%
ResNet	97,8%	92,5%	100%

Tabela 6.4: Sensibilidade das redes por tipo de iluminação da cena.

Resultado da Sensibilidade por Iluminação da Cena		
	Clara	Mista
YOLO	49,8%	47,1%
MT-CNN	57,4%	43,7%
ResNet	96,3%	98,1%

Tabela 6.5: Acurácia das redes por tipo de iluminação da cena.

Resultado da Acurácia por Iluminação da Cena		
	Clara	Mista
YOLO	57,1%	54%
MT-CNN	57,6%	43,9%
ResNet	97,6%	100%

Já nas Tabelas 6.7 e 6.6 são apresentados os resultados de acurácia e sensibilidade para as imagens classificadas por quantidade de pessoas na cena. Os casos em que as redes foram avaliadas foram em conjunto de imagens com 0 a 25 pessoas presentes, de 26 a 50 e de 51 a 200.

Tabela 6.6: Sensibilidade das redes em relação a quantidade de pessoas na cena.

Resultado da Sensibilidade por Quantidade de Pessoas			
	0 a 25	26 a 50	51 a 200
YOLO	86,3%	72,5%	31,9%
MT-CNN	61,3%	66,8%	48,8%
ResNet	99,1%	99,2%	94,7%

Tabela 6.7: Acurácia das redes em relação a quantidade de pessoas na cena.

Resultado da Acurácia por Quantidade de Pessoas			
	0 a 25	26 a 50	51 a 200
YOLO	100%	78,9%	32,3%
MT-CNN	62%	67,1%	48,9%
ResNet	100%	100%	96%

6.2 Experimento Controlado

Para análise dos resultados obtidos com a rede ResNet em experimentos do tipo Fatorial 2^k, descrito na Seção 5.4.3, é calculado o Efeito Principal de cada fator em questão. O Efeito Principal analisa o impacto da variação de nível daquele fator no resultado obtido quando os fatores restantes não se alteram. O Efeito Principal pode ser calculado de mais de uma forma. Utilizando a estratégia representar os níveis dos fatores com -1 e 1, o efeito principal de cada fator pode ser calculado pela seguinte equação simplificada [Calado (2003)]:

$$EfeitoPrincipal(X_j) = \frac{\sum |(N_i * Y_i)|}{2^k/2}$$

Onde *j* é o índice do fator analisado. *Ni* é o valor do nível (-1 ou 1) do experimento de índice *i*. *Yi* é o resultado da sensibilidade do experimento de índice *i*. E k é o número de fatores analisados no experimento. Na Tabela 6.8 são apresentados os valores dos Efeitos Principais de cada fator para o Experimento Controlado descrito na Seção 5.4.3.

Tabela 6.8: Resultado do Efeito Principal para os 5 fatores do experimento

Efeito Principal		
Índice	**Fator**	**Valor**
X1	Posição da Câmera	-0,0460
X2	Nível de Aglomeração	0,0313
X3	Distância das Pessoas para a Câmera	0,2004
X4	Iluminação da Imagem	0,0386
X5	Resolução da Imagem	0,1232

Para o Efeito Principal, quanto maior o valor absoluto maior é o impacto da variação de nível daquele fator no resultado do experimento. Para este experimento os maiores valores de Efeito Principal foram para os fatores de afastamento das pessoas da câmera e a resolução da imagem. Nos dois casos o tamanho do rosto das pessoas em *pixels* é o fator que causa o impacto. Quanto menor esse tamanho, mais dificuldade a rede tem de detectar os rostos. O resultado apresentado é significativo o bastante para indicar a necessidade de mais de uma câmera em ambientes muito grandes.

Outra análise que pode ser feita com um experimento do tipo Fatorial 2^k é sobre o Efeito de Interação. O resultado de Efeito de Interação mostra o grau de dependência e correlação entre os fatores analisados[Calado (2003)]. Corroborando com a análise do Efeito Principal, os fatores X3 e X5 apresentaram o maior valor de Efeito de Interação. Esse resultado era esperado e, como comentado anteriormente, a maior distância para a câmera e a menor resolução fazer com que os rostos presentes na imagem sejam representado com menos *pixels*, menos detalhes, dificultando a detecção.

6.2.1 Curva ROC

Uma maneira adicional de avaliar a performance de um problema de classificação é utilizando a Curva ROC (*Receiver Operating Characteristics*). Essa avaliação foi realizada com a mesma base utilizada na Seção 6.2. A curva ROC é uma medida de desempenho para tarefas de classificação em várias configurações de *thresholds*, nesse caso, o limiar de confiança para tomada de decisão de uma classificação. O gráfico possui dois eixos: Taxa de verdadeiro positivo (*True Positive Rate - TPR*) e Taxa de falso positivo (*False Positive Rate -*

FPR) para diferente limiares de classificação. A análise da curva diz quanto modelo é capaz de distinguir entre classes [Prati et al. (2008)]. Quanto mais próximo de 1 for a Área Sob a Curva (AUC) melhor o modelo sabe distinguir entre classes.

Na Figura 6.1 é apresentada a Curva ROC para os resultados do experimento realizado no CT da UFPB com a rede ResNet para limiares de confiança variando de 0% a 100%. O AUC calculado é de 0,97, aproximadamente, o que consolida o modelo de classificação como capaz de realizar a tarefa com eficiência. Para se construir o gráfico ROC plota-se FPR no eixo dos ordenadas (eixo x) e TPR no eixo das abscissas (eixo y) para cada valor de confiança de 0% a 100%.

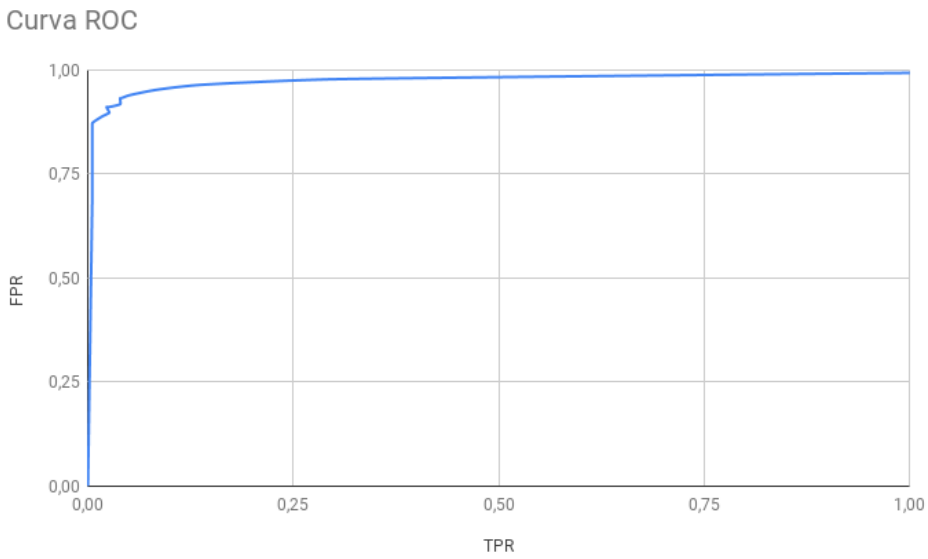


Figura 6.1: Curva ROC para os resultados do experimento realizado no CT da UFPB

6.3 Teste de Campo

O Módulo de Classificação apresentou resultados esperados, estando pronto para fazer os primeiros testes de campo. Esses testes possibilitariam avaliar a comunicação com as câmeras e as estratégias de captura e assim fazer o levantamento de requisitos para o desenvolvimento dos Módulos de Captura e de Controle.

Para este fim, foi realizado no Cinesystem Paulista um experimento onde foram realizados dois tipos de coleta de imagens: uma para calibragem, apenas com a equipe, e outro com

público real em sessões reais. Durante este primeiro teste de campo, foram capturadas centenas de imagens nas duas categorias e foi feito o processamento das mesmas, tanto *in loco* quanto a *posteriori*, que são as duas formas de operação da contagem. Nos testes com essas últimas imagens, foi obtido uma acurácia da ordem de 95% nas imagens de calibragem e de cerca de 87% nas imagens de sessões reais. Na Figuras 6.2 e 6.3 são mostradas exemplos de contagem de audiência realizada nas imagens capturadas no CineSystem Paulista.

Figura 6.2: Exemplo de saída do sistema de detecção de pessoas em situação de audiência na imagem capturada no CineSystem Paulista.

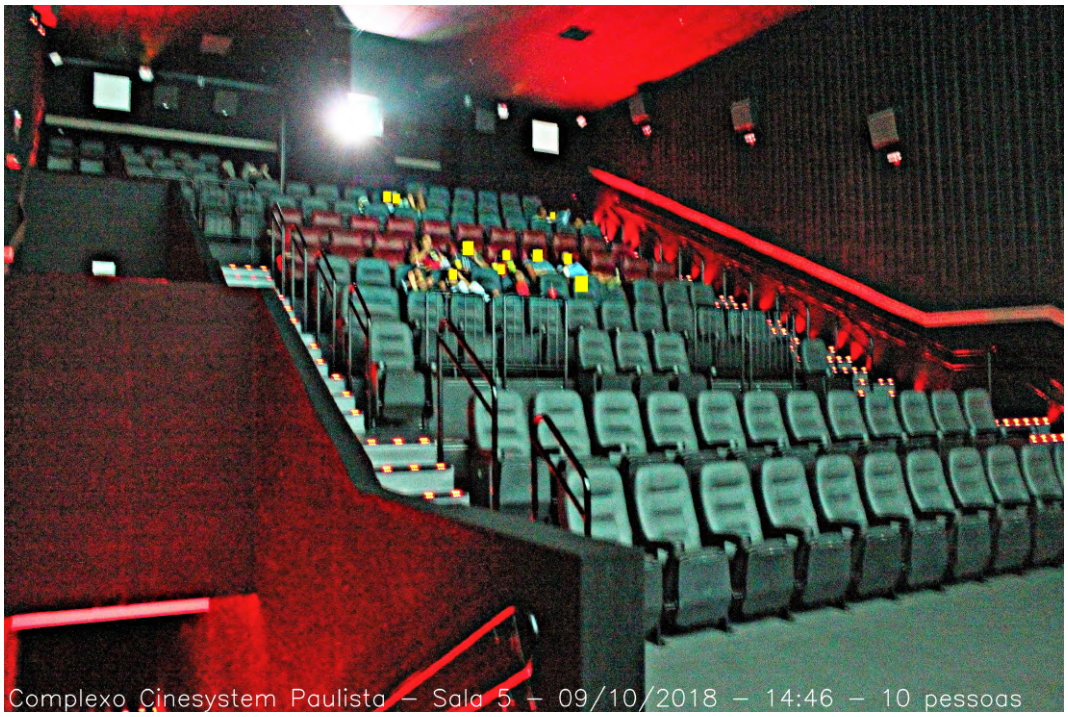


Figura 6.3: Exemplo de saída do sistema de detecção de pessoas em situação de audiência na imagem capturada no CineSystem Paulista.

Existe uma expectativa que esses valores possam ser melhorados, sobretudo porque não foi possível repetir o mesmo posicionamento da câmera usado na calibragem durante as sessões reais, nas quais as imagens foram capturadas em ângulo diagonal com o público. Além disso, como a iluminação durante as sessões depende da cena dos filmes e varia bastante de intensidade, detectou-se que será necessário incluir, no Módulo de Captura, um processo de seleção e descarte das imagens coletadas com iluminação insuficiente. Outro aspecto relevante é que ainda não ficou estabelecido que os modelos de câmeras usados nos testes (SLR, Go Pro e celular) sejam os mais adequados, sendo importante incluir outras alternativas, como câmeras de vigilância e câmeras com infravermelho. Finalmente, o fato de ser necessário haver uma pessoa a frente da audiência tirando fotos durante as sessões teve um efeito inibidor da captura, quando o ideal é ter uma câmera fixa, bem posicionada, operando de forma autônoma ou sendo controlada remotamente.

6.4 Discussão

Sobre a precisão da contagem, a eficácia de um sistema de contagem automática depende de alguns fatores, parte deles relacionado com a precisão da classificação e parte deles com a qualidade da imagem.

No primeiro caso, o algoritmo de contagem automática precisa ser treinado e calibrado para realizar a contagem dentro de um limiar aceitável de erro. Isso é feito usando grandes bancos de dados de imagens de contextos similares e ajustando os parâmetros e a sua sensibilidade para se equilibrar falsos positivos e falsos negativos nos resultados.

Garantir uma qualidade adequada das imagens capturadas, por sua vez, requer o ajuste de uma série de aspectos ambientais, tanto gerais quanto específicos de cada ambiente, como ângulo de posicionamento, iluminação, brilho, nível de oclusão, cobertura etc. Algumas destas dimensões podem ser trabalhados após a captura, em uma fase de pré-processamento das imagens e na sensibilidade da contagem, outros já precisam estar corretos no momento da coleta.

Como pode ser observado na Seção 6.1, o desempenho da ResNet chega a atingir 100% em alguns cenários. O pior resultado encontrado, de 92,5% e obtido na métrica de sensibilidade em cenários onde a posição da face é majoritariamente lateral, ainda foi três vezes melhor do que a média das outras técnicas aplicadas.

Nas figuras 6.4, 6.5 e 6.6 são apresentados exemplos do resultado de detecção de pessoas em situação de audiência.

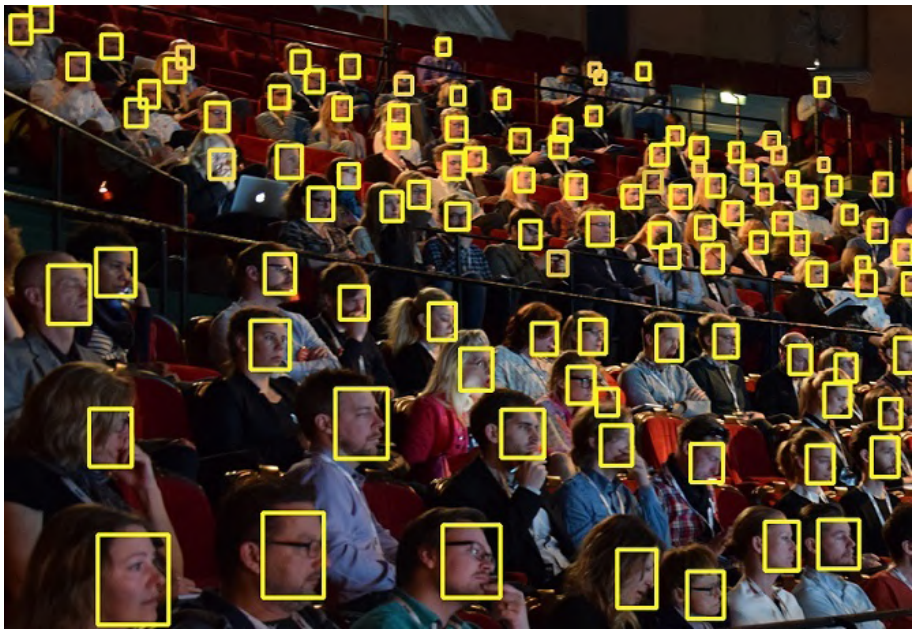


Figura 6.4: Exemplo de saída do sistema de detecção de pessoas em situação de audiência. Foto original por inUse Experience©

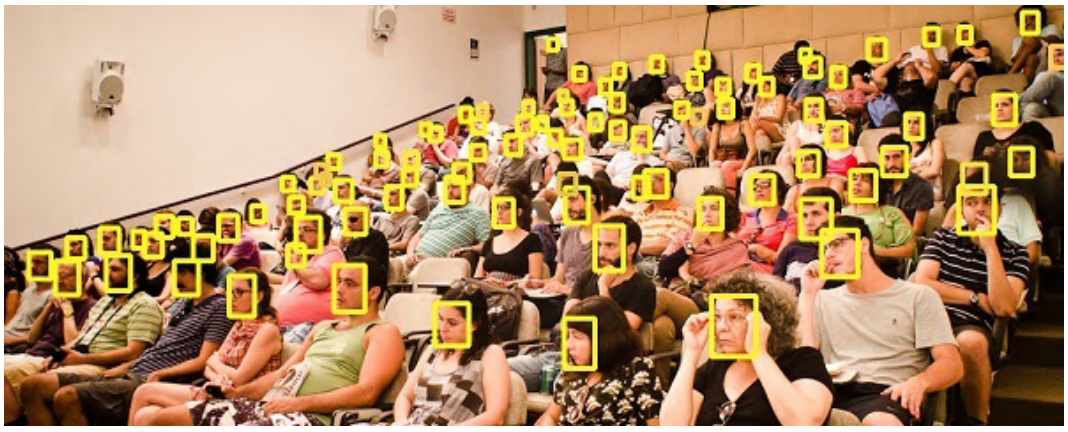


Figura 6.5: Exemplo de saída do sistema de detecção de pessoas em situação de audiência. Foto original por overmundo©

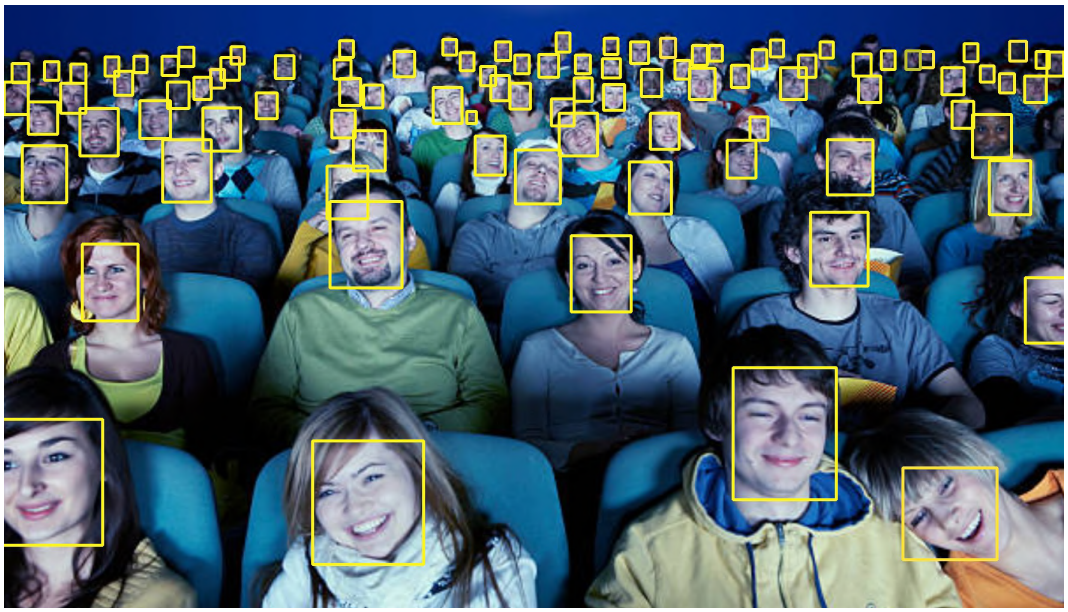


Figura 6.6: Exemplo de saída do sistema de detecção de pessoas em situação de audiência. Foto original: Er Creatives Services Ltd©

No Capítulo 7 são realizadas as conclusões deste resultados e do trabalho como um topo. Também é comentado as contribuições realizadas e as indicações de trabalhos futuros.

Capítulo 7

Conclusões e Trabalhos Futuros

Este capítulo apresenta a conclusão deste trabalho e as considerações finais sobre o projeto de pesquisa, como também deixa claro as limitações da proposta que não foram consideradas no escopo deste trabalho.

7.1 Considerações Finais

No estudo realizado no âmbito deste trabalho, foi avaliado o uso de técnicas de visão computacional clássicas e *deep learning* para realizar a contagem automática de pessoas em uma audiência. O intuito é realizar uma contagem segura, rápida, precisa, e não invasiva sem a necessidade da intervenção humana. Para alcançar este propósito, foram avaliadas técnicas tradicionais de contagem volumétrica e também três arquiteturas para o treinamento de redes neurais (YOLO, MT-CNN e ResNet). Adicionalmente, também foi desenvolvida uma base de imagens com amostras de pessoas em audiências para os testes de acurácia e sensibilidade das soluções investigadas.

Os resultados demonstram o potencial da técnica de *deep learning* para este tipo de aplicação. Em todos os cenários avaliados, em imagens separadas por posição das pessoas, iluminação e quantidade de pessoas, a ResNet obteve os melhores resultados para as métricas de interesse, com médias acima de 96%. A estratégia de aprendizado residual permite que redes mais profundas tenham a mesma complexidade de redes com menos camadas. Mas, apesar da mesma complexidade e sem o aumento da degradação do gradiente, a rede com mais camadas tem a capacidade de aprender mais características e reconhecer mais padrões. Isso

em uma aplicação onde o mesmo objeto pode aparecer de maneiras tão diferentes, devido ao tamanho que eles aparecem na imagem, pode ser fundamental para resultados melhores.

A YOLO, por sua vez, apresentou o menor tempo de execução e demonstra, como descrito em outros trabalhos [Guo et al. (2016)], a relevância em aplicações de detecção e classificação em tempo real. Como possui a abordagem de analisar a imagem apenas uma vez, para garantir a velocidade de execução, a baixa acurácia nesse tipo de aplicação, onde existe uma alta densidade de objetos a serem detectados e classificados, já era esperada.

Figura 7.1: Exemplo de imagem capturada no experimento realizado no Auditório do Centro de Tecnologia da UFPB.

Após os resultados dos primeiros experimentos, foi desenvolvido um protótipo utilizando a ResNet como rede neural para detecção de pessoas. A ResNet é a rede que precisa de mais tempo para fazer a detecção, mas na aplicação desenvolvida neste trabalho, esse requisito não é o mais significativo. A ResNet apresentou resultados de acurácia e sensibilidade significativamente melhor que as demais redes, por isso sua escolha para o protótipo.

O protótipo desenvolvido foi submetido a um experimento adicional com imagens capturadas no auditório do Centro de Tecnologia da Universidade Federal da Paraíba. Con-

siderando uma resolução de imagem mínima como requisito do sistema, o resultado final alcançado foi satisfatório e reafirma a capacidade do sistema de resolver a problemática descrita neste trabalho.

7.2 Trabalhos Futuros

Ainda que os objetivos desta pesquisa tenham sido atingidos e uma versão inicial de um sistema de referência tenha sido validado em um escopo delimitado para este trabalho, questões importantes foram percebidas e podem ser tratadas posteriormente em uma segunda etapa da pesquisa. Tais questões são apresentadas e discutidas nesta seção.

Os módulos de Captura e Controle foram desenvolvidos para possibilitar o teste e validação do sistema como um todo. Eles facilitam e auxiliam a execução e operação do Módulo de Classificação, núcleo do sistema proposto de contagem de audiências presenciais. Sendo assim, fica como indicação de trabalhos futuros a investigação aprofundada e validação das técnicas e estratégias desses dois módulos.

Apesar da dificuldade de se encontrar uma base de imagens com essas características, é promissor o uso de imagens capturadas com câmeras infravermelhas. Esse tipo de câmera pode mitigar o problema de imagens com pouca iluminação em ambientes escuros. Para a avaliação e melhor desempenho dessa técnica, faz-se necessário imagens desse tipo para treinamento e teste da rede neural.

Outro ponto seria o uso de técnicas de *data augmentation* focada no cenário da aplicação devem ser adotadas em experimentos futuros para permitir adicionar mais valores as base de imagens utilizadas. Após os resultados obtidos no experimento Fatorial 2K fica evidente a obtenção ou desenvolvimento de exemplos de imagens de treinamento de faces de baixa resolução ou *tiny faces*.

7.3 Contribuições

Durante o desenvolvimento deste projeto de pesquisa, foi publicado um artigo nos anais do Simpósio Brasileiro De Sistemas Multimídia E Web (Webmedia), conforme referência a seguir.

Florentino, C. S. e Costa, R. (2018). *A study on the use of deep learning for automatic audience counting*. Em *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pgs. 221–228. ACM. [Florentino e Costa (2018)]

Como outra contribuição, uma base específica para testes, diferentemente das bases de treinamento e validação, foi montada de forma que melhor representasse o contexto alvo. Neste sentido, foi construído um banco de imagens com o propósito específico de realizar a contagem automática de pessoas em uma audiência. As imagens selecionadas foram obtidas através de mecanismos de busca na internet e, em sua maioria, mostram um grupo de pessoas sentadas em um ambiente fechado como, por exemplo, a imagem mostrada na Figura 5.4.

O banco de dados de testes conta com 75 imagens e 3.206 amostras e foi usado para realizar a aferição da acurácia e sensibilidade da contagem. A contagem de pessoas foi feita manualmente para cada imagem selecionada.

Para uma melhor avaliação do desempenho das três redes treinadas, cada imagem da base de testes foi classificada com relação à posição das faces em relação ao ângulo da câmera, com relação ao tipo de iluminação da cena e também a quantidade de pessoas presentes. Dentro do possível, para o agrupamento das imagens, os intervalos categóricos foram definidos com o intuito de balancear o números de amostras em cada classe.

Uma segunda base de imagens, com imagens próprias, foi montada durante o desenvolvimento deste trabalho. Em um ambiente controlado, foram capturadas 62 imagens de pessoas em situação de audiência com uma combinação de seis fatores para análise. Cada um desses fatores possuem dois níveis de valores. Esse formato de base é ideal para realização de experimentos o tipo Fatorial 2K. As bases poderão ser acessadas no *link*: `https://drive.google.com/drive/folders/1OIv4TLaDYXWpfnv7Jz57ORtqzqD1nV_F?usp=sharing`.

Bibliografia

Ammar, A., Koubaa, A., Ahmed, M., e Saad, A. (2019). Aerial images processing for car detection using convolutional neural networks: Comparison between faster r-cnn and yolov3. *arXiv preprint arXiv:1910.07234*.

Ballard, D. H. (1982). Cm brown computer vision. *NY: Prentice Hill*.

Barandiaran, J., Murguia, B., e Boto, F. (2008). Real-time people counting using multiple lines. Em *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pgs. 159–162. IEEE.

Bengio, Y., Courville, A. C., e Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538.

Calado, V. (2003). *Planejamento de Experimentos usando o Statistica*. Editora E-papers.

Cao, X., Wei, Y., Wen, F., e Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Conte, D., Foggia, P., Percannella, G., Tufano, F., e Vento, M. (2010). A method for counting people in crowded scenes. Em *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pgs. 225–232. IEEE.

Davis, L. S. (1975). A survey of edge detection techniques. *Computer graphics and image processing*, 4(3):248–270.

Del-Blanco, C. R., Jaureguizar, F., e García, N. (2012). An efficient multiple object detection

and tracking framework for automatic counting and video surveillance applications. *IEEE Transactions on Consumer Electronics*, 58(3):857–862.

Elio, R., Hoover, J., Nikolaidis, I., Salavatipour, M., Stewart, L., e Wong, K. (2011). About computing science research methodology.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., e Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Fathy, M. e Siyal, M. Y. (1995). An image detection technique based on morphological edge detection and background differencing for real-time traffic analysis. *Pattern Recognition Letters*, 16(12):1321–1330.

Ferreira, J. Z., Rodrigues, J., Cristo, M., e de Oliveira, D. F. (2014). Multi-entity polarity analysis in financial documents. Em *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pgs. 115–122, New York, NY, USA. ACM.

Florentino, C. S. e Costa, R. (2018). A study on the use of deep learning for automatic audience counting. Em *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pgs. 221–228. ACM.

Garcia, S., Fernandez, A., Luengo, J., e Herrera, F. (2008). A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13(10):959.

Gasque, K. C. G. D. (2007). Teoria fundamentada: nova perspectiva à pesquisa exploratória.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., e Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

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

Hou, Y.-L. e Pang, G. K. (2011). People counting and human detection in a challenging situation. *IEEE transactions on systems, man, and cybernetics-part a: systems and humans*, 41(1):24–33.

I.S. SILVA, B. R. d. A. (2014). Sistema de contagem automática de objetos utilizando processamento digital de imagens em dispositivos móveis. Master's thesis, Universidade do Estado do Rio Grande do Norte.

LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

LeCun, Y., Huang, F. J., e Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. Em *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pgs. II–104. IEEE.

LeCun, Y., Kavukcuoglu, K., e Farabet, C. (2010). Convolutional networks and applications in vision. Em *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pgs. 253–256. IEEE.

Lempitsky, V. e Zisserman, A. (2010). Learning to count objects in images. Em *Advances in neural information processing systems*, pgs. 1324–1332.

Li, Y., Zeng, J., Zhang, J., Dai, A., Kan, M., Shan, S., e Chen, X. (2017). Kinnet: Fine-to-coarse deep metric learning for kinship verification. Em *Proceedings of the 2017 Workshop on Recognizing Families In the Wild*, pgs. 13–20.

Lienhart, R. e Maydt, J. (2002). An extended set of haar-like features for rapid object detection. Em *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pgs. I–I. IEEE.

Messias, J., Magno, G., Benevenuto, F., Veloso, A., e Almeida, V. (2015). Brazil around the world: Characterizing and detecting brazilian emigrants using google+. Em *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web*, WebMedia '15, pgs. 85–91, New York, NY, USA. ACM.

Molz, R. F. (2001). Uma metodologia para o desenvolvimento de aplicações de visão computacional utilizando um projeto conjunto de hardware e software. Master's thesis, Universidade Federal do Rio Grande do Sul.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990.

Prati, R., Batista, G., e Monard, M. (2008). Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, 6(2):215–222.

Qian, R. J. e Huang, T. S. (1996). Optimal edge detection in two-dimensional images. *IEEE Transactions on Image Processing*, 5(7):1215–1220.

Rahmalan, H., Nixon, M. S., e Carter, J. N. (2006). On crowd density estimation for surveillance.

Razavian, A. S., Azizpour, H., Sullivan, J., e Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. Em *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pgs. 512–519. IEEE.

Redmon, J., Divvala, S., Girshick, R., e Farhadi, A. (2016). You only look once: Unified, real-time object detection. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, pgs. 779–788.

Ren, S., He, K., Girshick, R., e Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Em *Advances in neural information processing systems*, pgs. 91–99.

Rodriguez, M., Laptev, I., Sivic, J., e Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. Em *Computer Vision (ICCV), 2011 IEEE International Conference on*, pgs. 2423–2430. IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Ryan, D., Denman, S., Fookes, C., e Sridharan, S. (2010). Crowd counting using group tracking and local features. Em *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pgs. 218–224. IEEE.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Sindagi, V. A. e Patel, V. M. (2018). A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16.

Sindagi, V. A. e Patel, V. M. (2019). Inverse attention guided deep crowd counting network. *arXiv preprint arXiv:1907.01193*.

Subburaman, V. B., Descamps, A., e Carincotte, C. (2012). Counting people in the crowd using a generic head detector. Em *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pgs. 470–475. IEEE.

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

Tan, H., Yang, B., e Ma, Z. (2013). Face recognition based on the fusion of global and local hog features of face images. *IET computer vision*, 8(3):224–234.

Uma, J. e Yuvarani, P. (2017). Detection of shapes and counting in toy manufacturing industry with help of phython. Em *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*, pgs. 1–5. IEEE.

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

Visala, A. (2014). People detection and tracking using a network of low-cost depth cameras.

Williams, N., Zander, S., e Armitage, G. (2006). A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. *ACM SIGCOMM Computer Communication Review*, 36(5):5–16.

Yang, S., Luo, P., Loy, C. C., e Tang, X. (2016). Wider face: A face detection benchmark. Em *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, X., Huang, J., Zhang, S., Yan, W., e Metaxas, D. N. (2013). Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. Em *Computer Vision (ICCV), 2013 IEEE International Conference on*, pgs. 1944–1951. IEEE.

Zhang, C., Li, H., Wang, X., e Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. Em *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pgs. 833–841.

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

Zhang, Y., Zhou, D., Chen, S., Gao, S., e Ma, Y. (2016b). Single-image crowd counting via multi-column convolutional neural network. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, pgs. 589–597.

Apêndice A

Registro Experimento Controlado

Universidade Federal da Paraíba
Centro de Informática
Departamento de Informática
Programa de Pós-Graduação em Informática

Ao Diretor do Diretor do Centro de Tecnologia da UFPB,
Prof. Dr. Antônio de Mello Villar,

Solicitamos, em nome do aluno Caio Souza Florentino, regularmente matriculado no Programa de Pós-Graduação em Informática, matrícula nº 20181000902, sob a orientação do Profº Rostand Costa, o uso do **Auditório da Área Administrativa do Centro de Tecnologia** para a realização de sessão de fotos que serão usadas como material de estudo de Mestrado. O evento será realizado uma única vez, das 17h às 18h, em uma das datas a seguir que melhor se adeque à agenda do Auditório do Centro de Tecnologia.

Opções de datas:
20/09, 25/09, 27/09, 02/10, 04/10, 09/10.

Em caso de mais de uma data possível, solicitamos a data mais próxima.

Saudações cordiais,

João Pessoa, 18 / setembro / 2018

Prof. Dr. Clauirton de Albuquerque Siebra
(Coordenador do Programa de Pós-Graduação)

Clauirton de Albuquerque Siebra
Coordenador do Programa de Pós-Graduação em Informática
SIAPE 1723491

Prof. Hamilton Soares da Silva
(Diretor de Centro)

Prof. Dr. Hamilton Soares da Silva
Diretor de Centro - CI/UFPB
Mat. Siape 003367271

RECEBIDO 19SET18

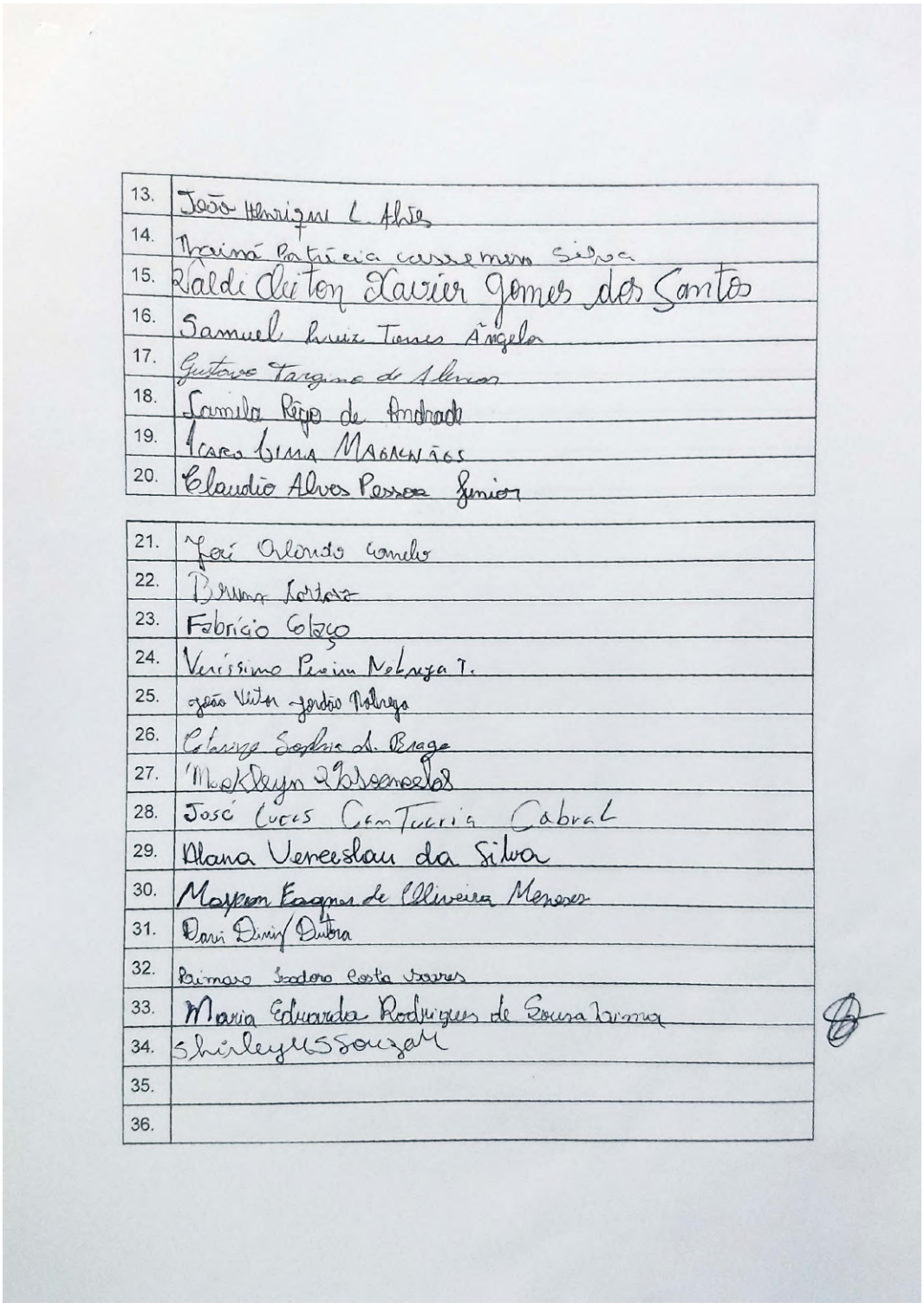
Universidade Federal da Paraíba
Centro de Informática
Departamento de Informática
Programa de Pós-Graduação em Informática

No dia 25 de Setembro de 2018 foi realizado pelo aluno Caio Souza Florentino do Programa de Pós-Graduação em Informática, no Auditório da Área Administrativa do Centro de Tecnologia da Universidade Federal da Paraíba, um experimento que visava simular e registrar em vídeo uma platéia.

As pessoas que subscrevem participaram do experimento e autorizam o uso do material para fins acadêmicos.

João Pessoa, 25 / Setembro / 2018

1.	Aron da Silva Fragoso
2.	Isabel Oliveira Araújo
3.	Vinícius Pinagé A. Lima
4.	Jhopson de V. Brito Braga
5.	Diego da Silva Amorim
6.	Carlos Henrique de O. Pereira
7.	Rubens Matheus Brasil da Silva Lima.
8.	Daniel dias dos Santos
9.	Roxmon Gomes
10.	Elvis Almeida
11.	Juliana Gonzaga Silva Costa
12.	Marcelo Henrique Ferro Mendes de Souza



13.	João Henrique L. Alves
14.	Thaina Patrícia carneiro Silva
15.	Raldi Cleiton Xavier Gomes dos Santos
16.	Samuel Luiz Torres Ângelo
17.	Gustavo Targino de Alencar
18.	Camila Rego de Andrade
19.	Icaro Lima Magalhães
20.	Claudio Alves Pessoa Junior

21.	José Orlando Coelho
22.	Bruna Santos
23.	Fabrício Colaço
24.	Veríssimo Pereira Nobrega T.
25.	João Victor Jordão Tobias
26.	Clarisse Sophia d. Braga
27.	Mackleyn Rosendo
28.	José Lucas Cantuária Cabral
29.	Diana Vereslau da Silva
30.	Maycon Eagnes de Oliveira Menezes
31.	Dani Dias Dutra
32.	Raimara Isadora Costa Soares
33.	Maria Eduarda Rodrigues de Sousa Lima
34.	Shirley Souza
35.	
36.	