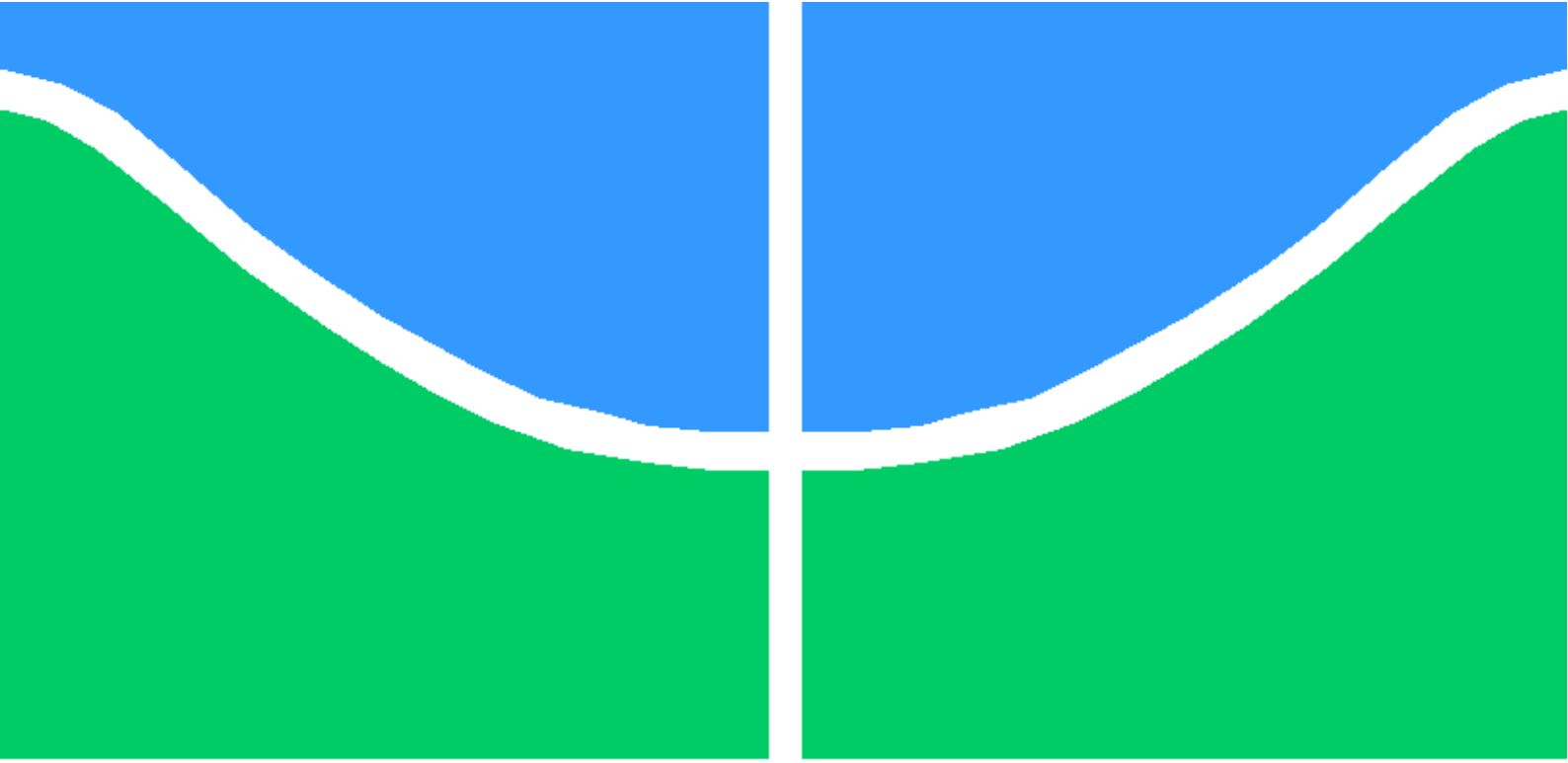


Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Elétrica

Estimação do número de pessoas em imagens de multidões

Luiz Henrique Padovani

Brasília
Dezembro de 2016

Luiz Henrique Padovani

Estimação do número de pessoas em imagens de multidões

Trabalho submetido ao Departamento de Engenharia Elétrica da Universidade de Brasília como requisito parcial para obtenção do Título de Bacharel em Engenharia Elétrica.

Universidade de Brasília
Faculdade de Tecnologia

Orientador: Prof. Eduardo Peixoto Fernandes da Silva, PhD.

Brasília
Dezembro de 2016

Luiz Henrique Padovani

Estimação do número de pessoas em imagens de multidões. Luiz Henrique Padovani. – Brasília, Dezembro de 2016.

54 p. : il. (algumas color.) ; 297 mm.

Orientador: Prof. Eduardo Peixoto Fernandes da Silva, PhD.

Trabalho de Conclusão de Curso – Universidade de Brasília
Faculdade de Tecnologia – Dezembro de 2016.

1. Processamento de Imagens. 2. Visão Computacional. 3. Estimação. 4. Multidões. Estimação do número de pessoas em imagens de multidões. ENE/FT – UnB/DF, Brasil.

Luiz Henrique Padovani

Estimação do número de pessoas em imagens de multidões

Trabalho submetido ao Departamento de Engenharia Elétrica da Universidade de Brasília como requisito parcial para obtenção do Título de Bacharel em Engenharia Elétrica.

Banca Examinadora

Prof. Eduardo Peixoto Fernandes da Silva, PhD. – UnB/DF
Orientador

Prof. Leonardo Rodrigues Araujo Xavier de Menezes, PhD. – UnB/DF
Examinador 1

Prof. Daniel Guerreiro e Silva, Dr. – UnB/DF
Examinador 2

Brasília, 12 de dezembro de 2016.

Dedico este trabalho aos meus pais, Rosana e Edinoel, aos meus irmãos, Luiz Gustavo e João Gabriel, aos meus tios e avós, à minha melhor amiga e eterna namorada Laís e aos meus amigos.

Agradecimentos

Aos meus pais, Rosana e Edinoel, pelo exemplo de conduta honesta e honrada, do qual eu me orgulho e sobre o qual eu me sustento; e por terem fornecido os subsídios, em todos os aspectos, que tornaram possível minha caminhada até aqui.

À minha companheira Laís Turra, pelo seu amor e apoio incondicionais, e pela força e motivação concedidas com tanto carinho.

Aos meus irmãos, Gu e João, pelos momentos e risadas, que tornaram a ladeira menos íngreme.

Ao professor Eduardo, pela valiosa orientação, e por toda a disposição dedicada a mim e a este trabalho.

A todos os professores, funcionários da Universidade, familiares e amigos que, direta ou indiretamente, contribuíram para minha formação e conclusão deste trabalho.

Resumo

O presente trabalho aborda o problema da estimação da contagem de pessoas em imagens de multidões empregando técnicas de processamento de imagens e visão computacional. Devido às dificuldades que algumas abordagens apresentam ao tratar de imagens que contêm uma alta densidade populacional — principalmente devido à baixa resolução dos objetos de interesse e ao elevado grau de oclusão encontrados em imagens dessa natureza; e por se basearem na detecção dos objetos ou na detecção de trajetórias coerentes — foi proposta uma abordagem baseada em características locais de *pixels* e em análise de textura para efetuar a estimação do número de pessoas nessas imagens. O método apresentado tem o propósito de complementar a atuação de outros métodos baseados em detecção, ainda que a combinação destes métodos não tenha sido objeto deste trabalho. O método proposto foi idealizado para trabalhar com técnicas que foram escolhidas visando a extração de informações texturais das imagens, o que reflete um ganho de representação das características inerentes a multidões densas devido ao surgimento de padrões periódicos em imagens dessa natureza. Desse modo, utilizou-se um modelo de regressão linear para relacionar as informações processadas da imagem com a contagem real de pessoas. Os resultados gerados pelo modelo revelam que, com os *features* utilizados, a saída do estimador parece acompanhar a contagem real de pessoas. Por fim, estabeleceu-se uma comparação com os resultados obtidos por outros trabalhos no intuito de mensurar se a acurácia obtida pelo método proposto se enquadra dentro das expectativas de desempenho almejadas. A conclusão é pela validade do método apresentado.

Palavras-chaves: estimação da contagem de pessoas; processamento de imagens; visão computacional; imagens de multidões.

Abstract

The present work deals with the problem of crowd counting in images using image processing and computer vision based techniques. Due to the difficulties shown by some approaches dealing with crowded images — mainly because of low resolution objects and occlusion, inherently found in images of this nature; and because it is based on object detection or coherent trajectories detection — a method based in image features like pixel and texture analysis was proposed to estimate the number of people in crowded images. The method aims to give complementary information for other detection methods. The proposal was designed to explore the periodic patterns that shows up in dense crowd images. In that sense, texture based techniques were chosen to extract the frequency and spatial information. This generates a gain in characteristics representation of dense crowds. Then, the processed information was used, combined with a linear regression model, to predict the crowd counting in a complex annotated image dataset. The results generated by the model reveal that, as expected, the features chosen seem to go along with the ground truth. Finally, a comparison was made with other studies to ascertain validation.

Key-words: crowd counting; image processing; computer vision; Crowds pictures.

Sumário

Lista de ilustrações

Lista de tabelas

1 Introdução

Este capítulo apresenta os objetivos que este trabalho visa atingir e introduz o contexto no qual está inserido, bem como a motivação para o seu desenvolvimento.

1.1 Contextualização

Um problema que sempre existiu na história humana é a contagem de grandes números, afinal, quem contaria, um a um, o número de grãos de areia em uma praia, ou o número de árvores em uma floresta? Quando se trata de grandes números, nós evitamos contá-los. A estimação é uma alternativa à contagem consideravelmente menos custosa. Entretanto, há um custo inerente: a imprecisão. Toda estimativa tem algum grau de imprecisão, do contrário, não seria uma estimativa.

A missão dos que estimam é encontrar um número para o valor de uma contagem que represente, com a fidedignidade necessária à aplicação, o próprio valor real da contagem, de modo que, para o propósito estipulado, contagem real e estimação possuam valores intercambiáveis. É com esse espírito que nasce a ideia de estimar o número de pessoas em multidões.

As subseções seguintes trazem um contexto amplo, incluindo os fatores não técnicos que estimularam a produção deste trabalho e tratando de casos mais gerais de contagem de pessoas. O Capítulo 2 trará um contexto mais específico do problema, abordando diversas técnicas que foram e são utilizadas para estimação de pessoas em imagens de multidões.

1.1.1 Motivação

A motivação para este trabalho surgiu, em meio a um período agitado da política nacional, da constatação da divergência para as estimativas divulgadas, por diferentes órgãos, do número de pessoas que compareciam a protestos e manifestações políticas. A reflexão quanto às possíveis implicações que estimativas enviesadas ou distorcidas poderiam causar, somada ao apreço pela verossimilhança das informações públicas, despertou a inquietação do autor.

Não raro, encontram-se relatos noticiados mostrando divergência excessiva entre órgãos divulgadores de estimativas do número de presentes em manifestações políticas (Portal de notícias G1, 2015, 2016). Eventos históricos também não escapam da imprecisão dos métodos globalmente difundidos. Um caso que ganhou notoriedade foi A Marcha de Um Milhão de Homens (*The Million Man March*), que aconteceu em Washington, D.C., na década de 90

(The Washinton Post, 2015). Uma entidade federal do governo dos Estados Unidos estimou a multidão presente em 400 mil participantes, equanto os organizadores alegavam que haviam mais de 1 milhão de pessoas. A controvérsia cessou quando um estudo da Universidade de Boston (BU Center for Remote Sensing, 1997) concluiu, utilizando uma metodologia que será mostrada no próximo capítulo, que haviam entre 650 mil e 1.1 milhão de presentes.

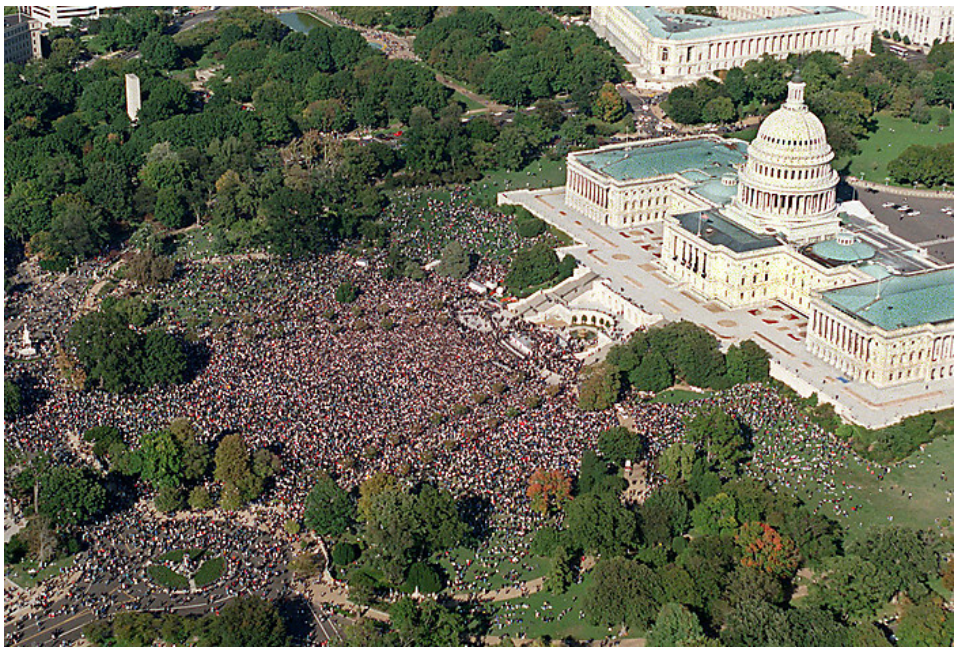


Figura 1 – A Marcha de Um Milhão de Homens, que aconteceu em outubro de 1995 nos EUA em um protesto por direitos civis. Fonte: (The Washinton Post, 2015).

1.1.2 Relevância

A contagem de pessoas é um elemento importante a ser considerado em diversas aplicações que envolvam a análise do comportamento humano. Alguns exemplos são: gerenciamento de multidões em grandes eventos, incluindo planejamento e controle em tempo real; projeto de espaços públicos e edificações; vigilância, controle e segurança de ambientes e análise de comportamento do consumidor em lojas de varejo (Tian et al., 2008; Store Smarts, 2016).

Diversos trabalhos já trataram do tema, utilizando abordagens baseadas em análise de *pixels*, análise de textura e análise de objetos. Métodos baseados em análise de *pixels* utilizam características locais da imagem como bordas para estimar a contagem (Chan & Vasconcelos, 2012; Lempitsky & Zisserman, 2010; Chen et al., 2012). Estes métodos visam realizar uma estimação da densidade de pessoas, ao invés de uma contagem propriamente dita. Métodos baseados em textura avaliam e extraem informações da imagem a partir do reconhecimento de padrões. Técnicas de análise de textura envolvem Matriz de Co-Ocorrência de Níveis de

Cinza, Análise de Fourier e Dimensões Fractais (Idrees et al., 2013; Marana et al., 1998; Chan & Vasconcelos, 2012). Métodos baseados em análise de objetos tentam detectar pessoas nas imagens e são utilizados para aplicações envolvendo imagens com baixa densidade de pessoas. A baixa resolução de imagens densas, assim como a presença de oclusão, prejudicam o desempenho desses métodos em imagens com alta densidade de pessoas (Rabaud & Belongie, 2006; Rodriguez et al., 2011).

1.2 Definição do Problema

Atualmente, os métodos utilizados por entidades governamentais e empresas para a estimação do número de pessoas em multidões não são automatizados, ou seja, ainda requerem que em alguma etapa do processo de estimação haja a participação de um humano. Em casos nos quais é necessária monitoração ininterrupta ou que envolva coleta de dados para análise, a contagem não automática pode se mostrar vulnerável, lenta ou até mesmo inviável. Dessa forma, métodos automatizados tornam-se úteis ou até mesmo fundamentais em algumas aplicações.

Pesquisas vêm sendo conduzidas para tentar automatizar os sistemas de contagem de pessoas, mas a grande maioria dos trabalhos ainda não são capazes de apresentar resultados satisfatórios quando a multidão é composta por centenas ou milhares de pessoas (Chan & Vasconcelos, 2012; Chen et al., 2012). Além disso, devido ao grande número de aplicações voltadas para a detecção, rastreamento e\ou análise de comportamento, a maioria dos métodos propostos na literatura têm a limitação de serem dependentes de imagens dinâmicas, inclusive com exigência de uma taxa mínima de quadros por segundo.

Ademais, as tecnologias utilizadas na prática para a estimação de pessoas em multidões são dispendiosas, por mobilizarem equipes ou exigirem equipamentos especiais, e imprecisas. Tais tecnologias poderiam ter uma significativa redução de custos e ganho de precisão com o desenvolvimento de novas técnicas. Nesse sentido, este trabalho busca trazer uma alternativa baseada em imagens estáticas, ou seja, livre da restrição temporal dos vídeos, e que seja aplicável a imagens que apresentam um grande número de pessoas (centenas e milhares).

1.3 Objetivos

O objetivo geral deste trabalho é realizar uma prova de conceito e atestar a validade das técnicas e métodos baseados em textura para estimar o número de pessoas em imagens de multidões densas.

Como objetivos específicos, pretende-se comparar técnicas de extração de características texturais, analisando os efeitos individuais da utilização de cada uma delas no resultado final,

bem como estudar as consequências de suas combinações para o desempenho do método proposto.

1.4 Organização do trabalho

O trabalho está organizado em seis capítulos, que estão divididos em seções e subseções. O Capítulo 1 apresenta uma introdução ao tema, expondo o que motivou sua escolha, bem como sua pertinência e, ao final, define os objetivos que deseja-se obter com este trabalho. O Capítulo 2 é destinado a uma revisão bibliográfica dos assuntos, técnicas e métodos relacionados ao tema e compara as suas tecnologias. O Capítulo 3, por sua vez, traz uma revisão teórica das principais ferramentas que subsidiaram a elaboração do método. Dando continuidade ao trabalho, o capítulo 4 apresenta o método proposto, tratando de duas abordagens: uma introdutória, baseada em detecção, e outra efetiva, que extende o alcance da estimação para multidões densas. Os resultados obtidos pelo método proposto e uma discussão a cerca deles são tratatos no Capítulo 5. Finalmente, no Capítulo 6, faz-se uma conclusão a respeito dos resultados obtidos e propõe-se sugestões para trabalhos futuros.

2 Revisão Bibliográfica

Este capítulo tem o intuito de mostrar os principais métodos de estimação do número de pessoas em multidões, tanto os que são de fato aplicados por institutos de pesquisa, empresas e organizações, quanto os últimos estudos na área e seu estado da arte.

2.1 Métodos não-automatizados

O método desenvolvido na década de 60 pelo professor da Universidade da Califórnia em Berkeley, Herbert Jacobs, é a fonte de diversos métodos utilizados por organizações, institutos e pesquisadores para contagem de pessoas. Sua fórmula, cunhada por ele próprio como *Jacobs Crowd Formula* em seu artigo original de 1967 (Jacobs, 1967), consiste em dividir a área total do espaço onde se encontra a aglomeração de pessoas em retângulos e multiplicar um fator de densidade pela soma da largura e do comprimento de cada retângulo. Então, a estimativa final $\hat{N}$ será dada pela soma das estimativas de cada retângulo, ou seja

$$\hat{N} = \sum_{i=1}^{n} (l_i + w_i) \cdot d_i \tag{2.1}$$

onde n é o número de retângulos, l_i, w_i e d_i são o comprimento, a largura e o fator de densidade do i-ésimo retângulo, respectivamente, sendo $d = 7$ para aglomerações relativamente bem espaçadas, nas quais as pessoas podem ser vistas tendo facilidade para se movimentarem e $d = 10$ para aglomerações mais compactas, nas quais as pessoas têm uma dificuldade maior para se movimentarem.

Recentemente, alguns refinamentos foram feitos, a partir do método proposto por Jacobs. A fórmula utilizada hoje se baseia no número de pessoas por quantidade de área e tem outros valores para quantificar o fator de densidade (Watson & Yip, 2011; Bialik, 2011; Choi-Fitzpatrick & Juskauskas, 2015). Os valores encontrados na literatura são os seguintes:

- 1 pessoa por metro quadrado, para densidades baixas;
- 2 pessoas por metro quadrado, para densidades médias; e
- 4 pessoas por metro quadrado, para densidade altas.

A grande vantagem que a fórmula de Jacobs traz é o tratamento da densidade heterogênea, possivelmente não utilizada em trabalhos anteriores, e que torna a estimativa mais acurada.

O Datafolha, instituto brasileiro de pesquisa pertencente ao Grupo Folha, se baseia nesta metodologia em suas estimativas, também dividindo a área do total em setores, mas acrescenta técnicas para contabilizar o efeito da rotatividade de pessoas, além de enviar profissionais a campo para atualizar periodicamente o valor das densidades em cada setor e com isso poder levantar até como se deu a evolução da quantidade de pessoas ao longo do tempo (Folha de São Paulo, 2013, 2016a,b). Deste modo, a estimativa do fator de densidade não se limita a valores pré-concebidos. Os pesquisadores levantam este número para cada setor levando em conta a estimativa, mais realista, dos profissionais que estão distribuídos na multidão.

O instituto Datafolha realiza a pesquisa, para a estimativa de pessoas presentes, de grandes eventos (em geral, manifestações) e para que possam efetuar seus cálculos, a área em que ocorrerá o evento deve ser um dado conhecido, tanto para eventos onde há um trajeto a ser percorrido, quanto para os eventos que se realizam no mesmo lugar. Inclusive, para o caso em que o instituto levanta também como se deu a variação do número de participantes ao longo do tempo, é necessário que haja um trajeto pré-definido (Folha de São Paulo, 2013). Como o instituto não utiliza imagens para chegar ao número final de pessoas em um determinado local (Folha de São Paulo, 2016a), seus valores para as áreas dos locais dependem de plantas baixas e inspeção local, o que traz alguma imprecisão para a estimativa.

Além de institutos de pesquisa, outras entidades constumam realizar esse tipo de levantamento como a Polícia Militar e os próprios organizadores dos eventos. A Polícia Militar não revela suas metodologias e se limita a dizer que "O cálculo foi elaborado pelo programa COPOM online, que realiza o georreferenciamento da área, definindo polígonos de concentração de pessoas por meios de inúmeras fotos aéreas e terrestres [...]" (SSP-SP, 2016).

Há também empresas privadas que prestam serviços para contagem do número de pessoas em multidões, como é o caso da *DigitalDesign & Imaging Service Inc.* (DDIS). Sua tecnologia se baseia também na ideia de Jacobs, mas utiliza, como é tendência atual, uma variedade maior de possíveis níveis de densidade. Além disso, a empresa realiza uma inspeção prévia a grandes eventos no intuito de prever onde as pessoas se concentrarão e assim soltar um balão, utilizado para a captura de imagens aéreas, em um local estratégico. O balão carrega uma camêra que tira fotos em 360° e a partir de uma variedade de altitudes. Com as imagens, a empresa faz um levantamento topográfico e com isso gera um mapa 3D.

A partir do mapa (exemplificado na Figura 2), que contém a divisão da imagem em setores, a empresa efetua a estimativa da densidade para cada setor e com isso contabiliza seu número final. A empresa alega que sua tecnologia fornece resultados com erros inferiores a 10% (Cariveau, 2015).

Uma nova abordagem que vem sendo adotada no contexto da contagem de pessoas é a utilização de *smartphones* para coleta e análise de dados. Há estudos em que se utiliza o sinal *bluetooth* para detectar dipositivos eletrônicos, que também disponham desta tecnologia, passíveis de serem descobertos e gerar uma relação entre o número de dipositivos encontrados

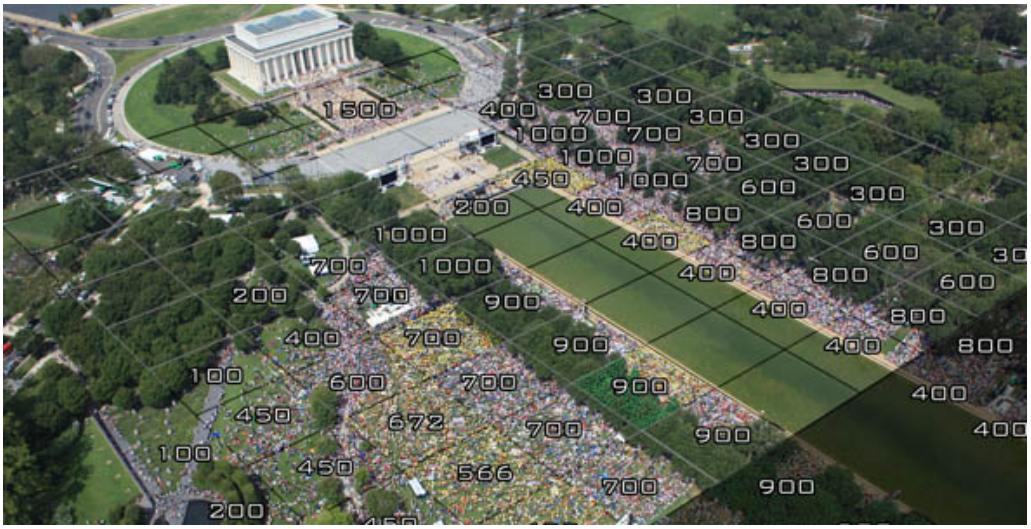


Figura 2 – Exemplo de um modelo gerado por meio de imagens aéreas da DDIS. (*Uso permitido.*)

e a quantidade de pessoas estimada em uma certa localidade (Nicolai & Kenn, 2007; Weppner & Lukowicz, 2011). Outros estudos mostram boas margens de erro na estimação (erro absoluto médio por volta de 13%) utilizando *Big Data*[1] para analisar os dados e projetar uma relação entre o número de chamadas e mensagens de texto, a quantidade de uso de internet e o volume de *tweets* em uma dada região com o número de pessoas presentes (Botta et al., 2015). Há ainda uma outra vertente que, por meio de um aplicativo, utiliza os *smartphones* como fontes geradoras e receptores de tons senoidais, controlando sua(s) caixa(s) de som e seu(s) microfone(s) (Kannan et al., 2012). O método se baseia no fato que, se cada aparelho transmitir um harmônico, uma análise espectral seria capaz de perceber a quantidade de tons, e consequentemente a quantidade de *smartphones*. Ademais, esses autores apontam uma acurácia de 90%.

Estas tecnologias baseadas na exploração de dados providos por *smartphones* ainda não vêm sendo utilizadas em larga escala por agentes que realizam pesquisas de contagem de pessoas em multidões na prática, mas o prognóstico é favorável devido, principalmente, à facilidade de implementação e à escalabilidade. Inclusive, a empresa israelense *StoreSmarts* (Store Smarts, 2016) já vem utilizando tecnologias similares para esta finalidade, com o uso do sinal de *Wi-Fi*.

2.2 Métodos Automatizados

A comunidade acadêmica produziu nos últimos anos uma variedade de métodos que visam automatizar a tarefa de detectar e contar objetos em imagens estáticas e também rastreá-los em imagens dinâmicas. Alguns desses métodos se destacam pelo alto nível de desempenho

[1] Tecnologia projetada para extrair valor de grandes volumes de dados (Gantz & Reinsel, 2011).

obtido e sua notoriedade os eleva ao patamar de *estado da arte* atual. Esses métodos possuem várias abordagens distintas, entre as quais a **contagem por detecção** — que utiliza classificadores treinados por *features*[2] locais da imagem para encontrar humanos ou partes de humanos, por exemplo —, a **contagem por agrupamento** — que localiza entidades individuais e as agrupa quando se comportam de modo semelhante, formando um objeto ou agrupamento contável —, e a **contagem por regressão** — que se baseia no aprendizado de uma função que relaciona *features* da imagem com a densidade do aglomerado de pessoas para estimar a contagem —, estão entre as principais. Aqui será apresentada uma breve descrição destas abordagens, consagradas pela literatura, com um destaque especial em contagem por regressão, pois é a estretégia que se mostrou mais efetiva em ambientes de maior densidade de pessoas e que é o foco deste trabalho.

2.2.1 Contagem por Detecção

A abordagem de contagem por detecção utiliza um classificador treinado que percorre a imagem por inteiro através de uma janela móvel. Os classificadores mais comuns para esta tarefa são *boosting*, Máquinas de Vetores de Suporte Lineares (*Support Vector Machines* – SVM) e Floresta Aleatória (*Random Forest Regression*) devido à agilidade oferecida (Loy et al., 2013).

Este paradigma de contagem tem grande aplicabilidade em casos de detecção de pedestres, por exemplo. Para estes casos, tipicamente utilizam-se classificadores treinados por *features* que representam características de formas gráficas com a aparência de um corpo humano inteiro ou de partes do corpo como cabeça ou o conjunto cabeça-ombros. Tais *features* podem ser descritores como *wavelets de Haar* (Viola & Jones, 2004) — que são *features* retangulares que exploram diferenças de sombreamento na imagem — ou histogramas de gradientes orientados (HOG - do inglês *Histogram of Oriented Gradient*) (Dalal & Triggs, 2005) — que descrevem bem formas e contornos. Os resultados obtidos com *features* de corpo inteiro podem fornecer resultados satisfatórios em cenários com distribuição esparsa de pessoas, entretanto, em cenários de maior concentração de pessoas, nos quais ocorre intensa oclusão (parcial ou total), a escolha de *features* de partes do corpo produz melhores resultados (Lin et al., 2001). Outra forma de amenizar o problema da oclusão é a utilização de técnicas de registro de imagem (Brown, 1992) a partir de imagens adquiridas por sensores dispostos em mais de um ponto de vista parcialmente interseccionados. Entretanto, essa configuração nem sempre está disponível, tornando essa solução restrita.

[2] O termo *feature* será utilizado neste trabalho por ser amplamente adotado na Ciência da Computação, em especial em Aprendizado de Máquina. O termo designa uma variável independente e possui diversos sinônimos, a depender do contexto, como por exemplo variável preditora, variável explicativa, variável regressora ou regressor são termos comuns em Econometria e Estatística; variável de entrada ou variável de controle são preferíveis em Engenharia (Dodge, 2006).

Um ponto importante a ser mencionado é a capacidade de transferência de aprendizado (ou capacidade de generalização). Um classificador com essa característica seria capaz de realizar a detecção em cenários genéricos de forma autônoma, o que é um grande desafio. A maioria das soluções apresentadas na literatura contam com detectores treinados para um cenário específico, porém estudos mais recentes vêm sendo publicados na intenção de propor soluções nesse sentido (Wang et al., 2012).

2.2.2 Contagem por Agrupamento (*Clustering*)

Clustering é o processo de organizar objetos em grupos cujos membros possuem similaridade, tal que a similaridade é definida por um critério arbitrário e seu desempenho depende da boa escolha desse critério. Em outras palavras, é um processo que lida com o problema de encontrar uma estrutura entre uma coleção de dados. Portanto, o paradigma de contagem por agrupamento se baseia na suposição de que os movimentos individuais dos *pixels*[3] observados são relativamente uniformes, então trajetórias de *pixels* coerentes podem ser agrupadas para representar grupos independentes.

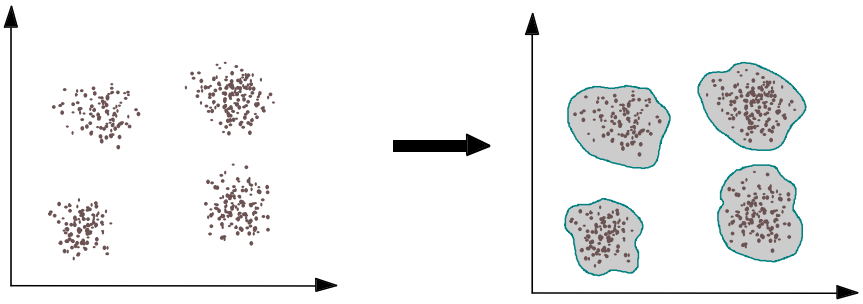


Figura 3 – Exemplo de classificação por agrupamento. No caso ilustrado na imagem, o critério utilizado para agrupamento é a posição relativa entre os dados.

É importante notar que a contagem por agrupamento depende da percepção de coerência entre trajetórias de objetos, ou seja, o método funcionará apenas com imagens dinâmicas e ainda exige uma taxa de quadros mínima para que a informação sobre a trajetória seja confiável. As contagens por detecção e por regressão não possuem essa restrição. (Chen et al., 2012)

2.2.3 Contagem por Regressão

As abordagens expostas acima produzem excelentes resultados para imagens de ambientes com distribuição espacial não concentrada de pessoas. Entretanto, a medida que a

[3] Contração de *picture element*, expressão em inglês que designa a menor área distinguível em uma imagem (Graf, 1999).

multidão fica mais densa, torna-se cada vez mais difícil utilizá-las, bem como seus resultados perdem qualidade (Chan & Vasconcelos, 2012). Desse modo, uma abordagem que ganhou destaque, por apresentar desempenho oposto aos outros paradigmas em cenários de maior densidade populacional, é a contagem por regressão (Loy et al., 2013). Tal método é capaz de estimar o número de pessoas em multidões por meio da aprendizagem de *features* em nível de *pixel* e de textura, além de poder ser aplicado a imagens estáticas ou dinâmicas. Outro fator interessante da abordagem por regressão reside no fato de que por dar um tratamento holístico ao processamento, é incapaz de identificar objetos individuais, o que pode ser útil em aplicações nas quais a privacidade das pessoas necessita ser preservada.

Vale observar que as primeiras soluções em detecção e/ou contagem de pessoas utilizavam a contagem por regressão. Contudo, a função que mapeava a contagem de *pixels* em número de pessoas num dado cenário apresentava falha quando este estava sujeito a severa oclusão (Davies et al., 1995). A oclusão é comum em configurações em que a vista é lateral, situação que era usual à época devido a aplicações concentradas em detecção de pessoas em ambientes como ruas e metrôs. Pensava-se que a principal causa para esta limitação era a utilização de um modelo linear para a regressão (Junior et al., 2010), todavia a maior deficiência desses métodos era a utilização de *features* baseados unicamente em *pixels*, obtidos através de segmentação de primeiro plano e de detecção de bordas, por exemplo. Essa deficiência foi corrigida posteriormente com o advento da introdução de novos *features* no modelo de regressão linear como os *features* de gradiente e textura (Loy et al., 2013).

Posto isto, para que o modelo de regressão forneça bons resultados, uma escolha criteriosa dos *features* é essencial. Uma estratégia comum é a de combinar *features* de natureza complementar. Nesse sentido, os mais adotados são *features* de segmentação, *features* de borda e *features* de textura e que serão discutidos em maior detalhe a seguir.

Features de segmentação visam estabelecer uma diferenciação entre o primeiro e o segundo plano (*foreground*$\,$*background*), em que no primeiro plano estão os objetos de interesse e no segundo plano estão elementos do ambiente, oferencendo destaque para aquele. Em suma, a operação estabelece uma subtração do segundo plano ou uma detecão do primeiro. Entre os *features* que podem ser extraídos desta segmentação estão:

- Área – número total de pixels no primeiro plano;
- Perímetro – número total de pixels que compões o perímetro do primeiro plano;
- Razão Área-Perímetro – visa medir a complexidade da forma segmentada;
- Orientação da borda do perímetro – histograma orientado do perímetro;
- Contagem de bolhas – número de componentes conectados com área maior que um limiar pré-estabelecido.

É importante notar que a detecção de primeiro plano depende de uma sequência de quadros (imagem dinâmica) para realizar a diferenciação dos objetos de interesse que, neste caso, precisam ser móveis.

Enquanto *features* de segmentação aprendem propriedades globais da imagem, outros *features*, como os de borda, trazem informações complementares a respeito de propriedades locais (Chan & Vasconcelos, 2012). Imagens de multidões pouco densas tendem a apresentar bordas mais homogêneas, ao passo que em imagens com multidões densas a tendência é o surgimento de bordas mais complexas e heterogêneas. Alguns dos *features* de borda mais comuns são os seguintes:

- Pixels de borda – número total de pixels na borda;
- Orientação da borda – histograma da orientação das bordas do primeiro plano;
- Dimensão de Minkowski – conta quantos elementos estruturais pré-estabelecidos são necessários para preencher o interior das bordas.

A textura da imagem traz informações muito relevantes em relação ao número de pessoas nela contida. Em destaque, há um padrão que sugere muito a esse respeito: regiões de intensa densidade tendem a exibir uma resposta mais significativa à *features* de textura. Isso se traduz no fato de que áreas de alta densidade são, normalmente, constituídas por padrões finos (que correspondem a altas frequências no domínio de Fourier), enquanto que áreas de baixa densidade são, usualmente, constituídas de padrões grossos (que correspondem a baixas frequências no domínio de Fourier), especialmente quando seu segundo plano também é constituído por este padrão (Marana et al., 1998).

2.3 Comparação das soluções

Os métodos utilizados para contagem de pessoas em multidões sem o devido aparato tecnológico exigem dispêndio de pessoal para realizar a contagem manual ou estimar densidades localmente. Por outro lado, métodos que utilizam tecnologia para aquisição de imagens aéreas ainda não dispõem de técnicas de processamento de imagens avançadas o suficiente para tornar a ação automatizada.

O uso de novas tecnologias que instituam a automatização na estimação são desejáveis para tornar o processo de contagem de multidões menos laborioso. Assim como, para aplicações específicas envolvendo monitoração, traria uma maior segurança para o sistema ao evitar a necessidade de observação humana ininterrupta, que é mais sucetível a falhas. Exemplos de aplicação como vigilância por circuito fechado de televisão, que já possuem tecnologias baseadas em elementos de alto nível, como objetos (Marana et al., 1998), são beneficiados por essas inovações.

No entanto, as abordagens de contagem por detecção ou por agrupamento dependem da segmentação explícita dos objetos ou da identificação de trajetórias por rastreamento de pontos coerentes. Elas não são adequadas para imagens com multidões densas com segundo plano complexo e oclusão. Em contraste, um modelo de contagem por regressão visa aprender um mapeamento direto entre as características a nível de *pixel* e contagem sem a segmentação explícita, ou identificação, ou rastreamento de indivíduos. Esta abordagem é, portanto, mais adequada para imagens com intensa densidade de pessoas (Chen et al., 2012).

3 Fundamentos Teóricos

Este capítulo foi escrito com a intenção de apresentar os fundamentos teóricos das ferramentas que são utilizadas no trabalho. Primeiramente, expõe-se o algoritmo de Viola-Jones, que é um algoritmo de uso bastante difundido e especialmente popular em trabalhos nos quais a tarefa de detecção de faces em imagens aparece. Em seguida, é apresentado o conceito de detecção de bordas em imagens, bem como as técnicas envolvidas e suas dificuldades. Também é apresentada a Matriz de Co-Ocorrência de níveis de cinza e as informações que ela carrega. Por fim, faz-se uma descrição do método de Regressão Linear.

3.1 Algoritmo de Viola-Jones

O problema de detecção de objetos pode ser trivial para um humano, contudo é uma tarefa não-trivial para um computador, que deve receber as instruções corretas para conseguir aprendê-la. Um algoritmo de detecção de objetos deve ser capaz de apontar se em uma dada imagem existe um certo objeto de interesse ou não e, caso exista, localizá-lo. Para que atinja critérios de desempenho satisfatórios, o algoritmo deve minimizar as taxas de falsos negativos e falsos positivos. Nesse sentido, surgiram diversas técnicas que podem ser empregadas para realizar essa tarefa, mas o detector de objetos conhecido por Viola-Jones, proposto por Paul Viola e Michael Jones, inicialmente em 2001 e aperfeiçoado em 2004 (Viola & Jones, 2004), ganhou destaque e a aceitação da comunidade acadêmica pela alta taxa de acertos e velocidade de execução e pelo baixo custo computacional. Embora o detector proposto por eles seja capaz de ser treinado para reconhecer outros objetos, a motivação principal do trabalho foi o reconhecimento facial.

A estrutura do algoritmo é fundada na adoção da seguinte estratégia: utiliza-se *features* baseados em *wavelets de Haar*, extraídos a partir da imagem integral, para treinar um conjunto de classificadores fracos, que por sua vez alimentam um conjunto de classificadores fortes dispostos em uma arquitetura em cascata. A seguir serão sucintamente descritas as etapas envolvidas nessa estrutura.

3.1.1 Extração dos *features*

A primeira etapa é a extração dos *features* baseados em *wavelets de Haar*, isto é, a operação de um produto escalar entre a imagem e os padrões *Haar-like*, que são retângulos formados por sub-retângulos brancos ou pretos e de mesmas dimensões. Isto é feito computando

o valor dos *features* em cada janela w através da seguinte expressão:

$$f_i(w) = \sum^{w} pixel_{preto} - \sum^{w} pixel_{branco} \tag{3.1}$$

na qual é computada a diferença dos somatórios dos pixels nas regiões preta e branca do i-ésimo padrão *Haar-like* utilizado na extração dos *features*, que contêm informação suficiente para caracterizar faces, uma vez que estas são regulares por natureza. Cada padrão tem sua forma de contribuir para a identificação de alguma característica. Por exemplo, os padrões dispostos na parte superior da Figura 4 permitem identificar áreas da imagem onde há diferença significativa de intensidade entre as metades inferior e superior, ou direita e esquerda de uma sub-região.

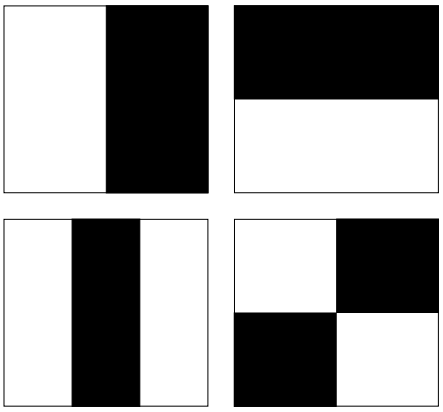


Figura 4 – Exemplos de alguns padrões *Haar-like* utilizados no Viola-Jones.

Para a computação dos somatórios, ao invés de simplesmente somar os valores de todos os *pixels* dentro de uma janela retangular, utiliza-se a *imagem integral*, que é um algoritmo criado em 1984 por Franklin Crow e conhecido também pelo nome tabela de soma das áreas (Crow, 1984). A vantagem da utilização da imagem integral é sua velocidade de computação: o algoritmo trabalha em tempo linear (Wang, 2014) e, portanto, com baixo custo computacional. Utilizando o sistema de coordenadas padrão para imagens, isto é, com a origem localizada no canto superior esquerdo de uma imagem $i(x, y)$, a imagem integral é dada por:

$$ii(x, y) = \sum_{\substack{x' \leq x \\ y' \leq y}} i(x', y') \tag{3.2}$$

Esta equação é válida dentro dos limites da imagem e diz que a imagem integral na coordenada (x,y) é composta pela soma dos valores dos *pixels* que estão acima e à esquerda de (x,y), x e y inclusos. A partir desta expressão, e notando que o valor de $ii(x, y)$ pode ser computado em uma única varredura, de modo eficiente, usando o fato de que:

$$ii(x_0, y_0) = i(x_0, y_0) + ii(x_0 - 1, y_0) + ii(x_0, y_0 - 1) - ii(x_0 - 1, y_0 - 1) \tag{3.3}$$

A Figura 4 exemplifica a computação da imagem integral de maneira recursiva, efetuada em tempo linear.

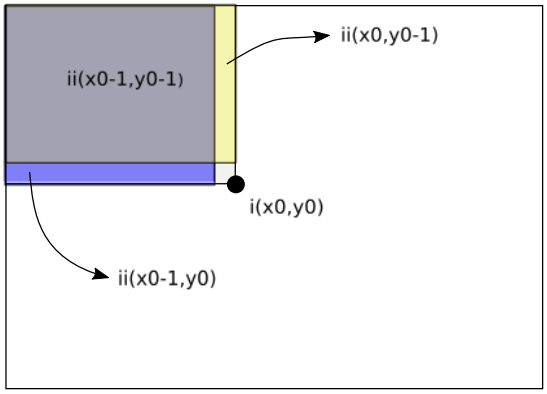


Figura 5 – Imagem integral: ilustração da equação 3.3. A distância entre os *pixels* adjacentes foram exageradas para melhor visualização.

Desse modo, com apenas uma operação, encontra-se a soma dos valores de todos os *pixels* contidos em qualquer região retangular na imagem. Assim, dada uma região retangular ABCD contida em uma imagem $i(x, y)$ (ilustrada na Figura 6), a soma das intensidades dos pixels nessa região é calculada da seguinte forma:

$$\sum_{(x,y \in ABCD)} = ii(A) + ii(D) - ii(B) - ii(C) \tag{3.4}$$

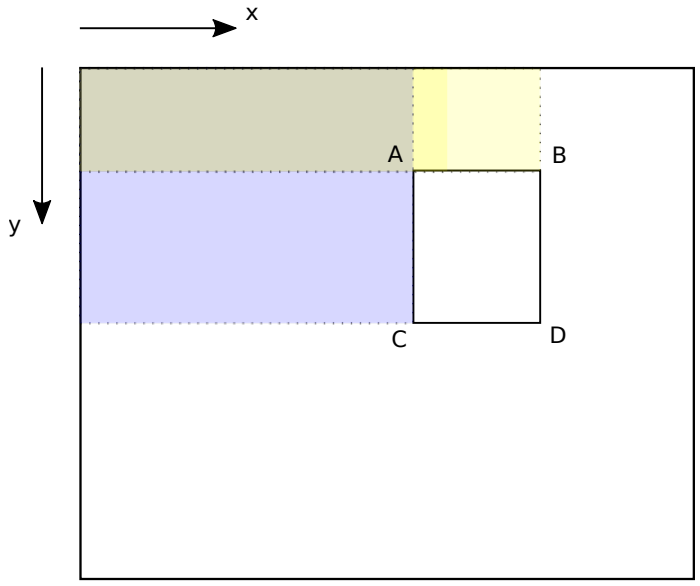


Figura 6 – Imagem integral: ilustração da equação 3.4.

Com isso, os somatórios da equação 3.1 são computados mais rapidamente e as características de sombreamento podem ser extraídas da imagem com a formação dos *features* $f_i(w)$.

3.1.2 Treinamento e classificação

O treinamento de um detector de faces baseado em um classificador binário é feito apresentando-no exemplos positivos, ou seja, faces, e exemplos negativos, que podem ser qualquer coisa que não seja uma face. O classificador então realizará várias iterações até aprender a diferenciar os exemplos.

O algoritimo de classificação escolhido por Viola e Jones é o *Adaboost*[1]. Este classificador é um aperfeiçoamento de um outro classificador (mais precisamente, família de algoritmos de classificação) chamado *boosting*, cuja ideia central é a de que é mais simples encontrar relações de acurácia modesta (regras fracas) do que relações de alta acurácia (regras fortes) e a partir de uma variedade de regras fracas, gerar uma regra forte. O algoritmo usado para encontrar regras fracas é chamado de classificador fraco (ou hipótese fraca), assim como o algoritmo usado para encontrar regras fortes é chamado classificador forte (ou hipótese forte). No caso de uma classificação binária, para ser considerado fraco, o classificador deve possuir acurácia maior que 50% para que seja útil (afinal, o erro de um palpite aleatório é de 50%).

O *Adaboost* é, basicamente, um algoritmo de *boosting* com uma estratégia melhor de atualização dos pesos de cada regra. Em cada iteração, um conjunto de hipóteses fracas h_t é ajustado para minimizar o erro de classificação. À cada uma dessas hipóteses é associada uma característica $f_t(x)$, um limiar θ_t e uma paridade p_t que estão relacionadas pela regra:

$$h_t(x) = \begin{cases} 1, & \text{se } p_t f_t(x) > p_t \theta_t \\ -1, & \text{caso contrário} \end{cases} \tag{3.5}$$

tal que a paridade p_t indica a direção da desigualdade.

Assumindo que a cada uma das T hipóteses também seja associado um peso α_t:

$$h = \{h_t : t = 1, ..., T\} \tag{3.6}$$

$$\alpha = \{\alpha_t : t = 1, ..., T\} \tag{3.7}$$

Sendo $F(x)$ a combinação linear das hipóteses fracas, ou seja:

$$F(x) = \sum_{t=0}^{T} \alpha_t h_t(x) = \langle \alpha, h(x) \rangle \tag{3.8}$$

[1] Expressão para *Adaptive boosting*, uma abordagem ao aprenizado de máquina baseada na ideia de criar uma regra de predição acurada combinando diversas regras relativamente fracas e inacuradas (Schapire, 2013).

Define-se uma hipótese forte como a função sinal deste produto interno:

$$H(x) = sign[F(x)] = \begin{cases} 1, & \text{se } F(x) > 0 \\ 0, & \text{se } F(x) = 0 \\ -1, & \text{se } F(x) < 0 \end{cases} \tag{3.9}$$

O objetivo é escolher h e α tal que o erro na hipótese forte seja minimizado. Erro este, que tende a ser menor utilizando o *boosting* adaptativo, mas que ainda precisa ser melhorado. Por esse motivo, adota-se uma estratégia em que os classificadores são dispostos em cascata, de tal maneira que cada estágio seja composto por um conjunto distinto de classificadores fortes, em que os primeiros estágios buscam por características mais gerais de faces (como formato e contorno), ao passo que os últimos estágios buscam por características mais específicas como posição relativa e sombramento de elementos internos — boca, nariz e olhos. Isso permite que os elementos que compõe o plano de fundo, por exemplo, sejam descartados logo nos primeiros estágios, gerando um ganho de eficiência ao despender maior processamento aos objetos que realmente sejam ou se assemelhem à faces. O fluxograma do sistema descrito é mostrado na Figura 7.

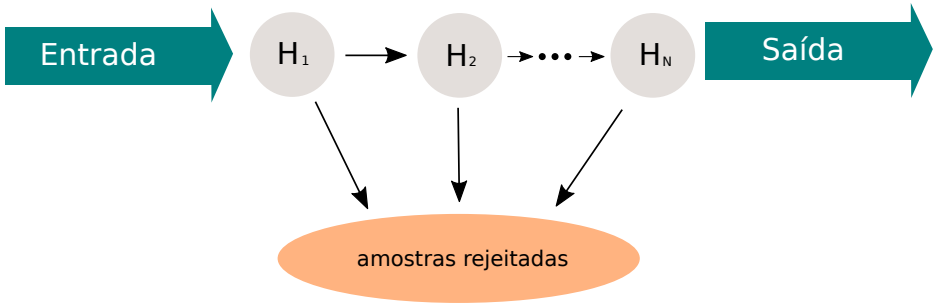


Figura 7 – Fuxograma da arquitetura em cascata dos classificadores do algoritmo de Viola-Jones.

Os objetivos principais do uso da arquitetura em cascata são acelerar a execução do algoritmo, além de diminuir a taxa de falsos positivos devido à multipla análise de cada exemplo, isto é, casos negativos são rejeitados nos primeiros níveis, porém um caso positivo necessariamente percorreu todos os estágios.

É importante ressaltar que na detecção, como não se sabe a posição nem o tamanho das faces nas imagens de teste, a janela que varre a imagem é escalonada de um tamanho mínimo até o tamanho da imagem para que a variabilidade de tamanhos permita, caso haja um positivo, que ele seja identificado. Além disso, vale registrar que, pelas próprias características que são exploradas pelos padrões *Haar-like*, faces de perfil, ou com muita inclinação, ou submetida à falta ou ao excesso de iluminação, ou faces com óculos escuros dificilmente são detectadas. O algoritmo que foi apresentado produz bons resultados em situações relativamente específicas,

nas quais as faces estão dispostas frontamelmente, sem oclusão e sem os fatores negativos citados acima.

3.2 Detecção de bordas

Bordas são regiões onde há uma transição de intensidade, abrupta ou suave, entre um objeto e outro, isto é, são os contornos dos objetos e assim os diferenciam. A detecção de bordas é composta por diversas técnicas que visam identificar essas regiões da imagem em que se percebe uma variação na intensidade. Quanto mais abrupta essa variação, maior a chance de haver um contorno. Este fator é conhecido por força do contorno ou contexto e é um dos fatores que mais causa dificuldades na detecção de bordas, pois é necessário que se defina quão intensa uma variação de intensidade deve ser para que seja caracterizado um contorno.

A maioria dos detectores de borda utilizados são baseados, fundamentalmente, no uso de operações de derivadas de primeira e segunda ordem. É importante registrar que esses filtros não encontram as bordas, propriamente ditas. Ao invés, eles fornecem uma indicação das regiões em que elas ocorrerão com maior probabilidade (Solomon & Breckon, 2011).

3.3 Matriz de co-ocorrência de níveis de cinza

Usualmente, uma imagem em níveis de cinza possui 256 níveis, cada um representando uma intensidade que varia desde o 0 (preto) ao 255 (branco). A chamada matriz de co-ocorrência (ou matriz de probabilidade conjunta) de níveis de cinza (GLCM, do inglês *Gray-Level Co-Ocurrence Matrix*) é obtida, primeiramente, por um processo de quantização dos níveis de cinza, que é feito tipicamente com oito níveis. Assim, será gerada uma nova imagem quantizada cujos valores variam de 1 a 8. A partir da imagem quantizada, a GLCM é construída pela contabilização da frequência com que pares de *pixels* com intensidades i e j, separados por um vetor deslocamento $\vec{d}$, ocorrem na imagem. Desse modo, cada elemento $p(i, j|\phi)$ da GLCM representa a frequência com que dois valores i e j de intensidade, separados por $\vec{d}$, ocorreram na imagem quantizada, sendo d o módulo e ϕ a direção de $\vec{d}$.

A variação do vetor $\vec{d}$ permite a captura de diferentes características da textura da imagem. Então é comum que algumas GLCM sejam geradas a partir de variações de d e de ϕ para obter-se mais informações dos *features*.

Como exemplo, considere a imagem 5×6 a seguir:

$$I = \begin{bmatrix} 42 & 221 & 67 & 85 & 22 & 100 \\ 203 & 16 & 92 & 131 & 62 & 138 \\ 50 & 58 & 157 & 182 & 130 & 193 \\ 250 & 30 & 127 & 97 & 171 & 95 \\ 73 & 41 & 41 & 159 & 133 & 0 \end{bmatrix}$$

Quantizando a imagem em oito níveis de intensidade, obtem-se:

$$I_q = \begin{bmatrix} 2 & 7 & 3 & 3 & 1 & 4 \\ 7 & 1 & 3 & 5 & 2 & 5 \\ 2 & 2 & 5 & 6 & 5 & 7 \\ 8 & 1 & 4 & 4 & 6 & 3 \\ 3 & 2 & 2 & 5 & 5 & 1 \end{bmatrix}$$

Assim, para a matriz que representa a imagem quantizada, mostrada acima, pode-se obter uma variedade de matrizes de co-ocorrência de níveis de cinza, especificando-se d e ϕ. Nas Tabelas 1 e 2 estão representadas as GLCMs para os casos em que $d = 2$ e $\phi = 0°$ e $d = 1$ e $\phi = 45°$, respectivamente.

	1	2	3	4	5	6	7	8
1	0	0	0	1	1	0	0	0
2	0	0	1	0	3	1	0	0
3	1	2	0	1	0	0	0	0
4	0	0	1	0	0	1	0	0
5	1	0	0	0	2	0	0	0
6	0	0	0	0	0	0	1	0
7	0	0	2	0	0	0	0	0
8	0	0	0	1	0	0	0	0

Tabela 1 – GLCM para $d = 2$ e $\phi = 0°$.

	1	2	3	4	5	6	7	8
1	0	0	1	0	1	0	0	0
2	1	0	1	3	0	0	0	0
3	1	0	1	0	0	0	0	0
4	0	0	0	0	1	1	0	0
5	1	0	1	0	2	1	0	0
6	0	1	0	0	0	0	1	0
7	0	0	0	0	0	0	1	0
8	0	1	0	0	0	0	0	0

Tabela 2 – GLCM para $d = 1$ e $\phi = 45°$.

A análise da Tabela 1 permite constatar que o elemento (7,3) tem valor igual a dois, porque há duas ocorrências em que pixels horizontalmente distantes por dois pixels têm valores 7 e 3, assim como na Tabela 2 nota-se que o valor do elemento (2,4) é três, pois há três ocorrências em que um quatro está acima e a direita de um *pixel* de valor dois.

A partir da GLCM, é possível extrair *features* estatísticos que carregam informações a respeito da distribuição espacial e recorrência de padrões da imagem. Haralick foi quem utilizou a técnica pioneiramente para obter essas informações de uma imagem a partir de uma análise da textura. Originalmente, e ainda hoje, a GLMC é chamada também por *Gray-Tone Spatial-Dependece Matrix* (Haralick et al., 1973; Haralick, 1979). Em seu trabalho, Haralick considerou

diversos features estatísticos. Os mais presentes nos trabalhos atuais são contraste, energia, entropia, homogeneidade e correlação e serão descritos a seguir.

O contraste mede a intensidade da diferença entre dois pixels, ou seja, o contraste terá maior presença quanto maiores os elementos mais distantes da diagonal principal de GLCM e será nulo para imagens constantes. Sua equação é dada por:

$$f_1 = \sum_{i,j} |i-j|^2 \, p(i,j) \tag{3.10}$$

A energia mede a uniformidade da textura, isto é, a chance de repetição dos pares de *pixels* ou a chance do vetor deslocamento cair repetidas vezes sobre o mesmo par. A energia será igual a um para imagens constantes.

$$f_2 = \sum_{i,j} p(i,j)^2 \tag{3.11}$$

Entropia é a propriedade que mede o grau de aleatoriedade da textura, ou seja, sua desordem. Se uma imagem tiver uma distribuição de *pixels* completamente aleatória, a GLCM não terá nenhum conjunto de pares preferido, então será uniforme[2].

$$f_3 = \sum_{i,j} -p(i,j) \; log \; p(i,j) \tag{3.12}$$

A proximidade da distribuição dos elementos da GLCM em relação à sua diagonal é medida pela homogeneidade. Visualmente, a homogeneidade mede a suavidade da textura da imagem.

$$f_4 = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \tag{3.13}$$

3.4 Modelo de Regressão

A regressão é um algoritmo de aprendizado supervisionado, ou seja, procura resolver o problema cujo objetivo é estimar as relações entre as variáveis independentes (*features*) e a variável dependente e com isso estabelecer um modelo capaz de prever o comportamento da variável dependente a partir de alterações nas variáveis independentes. Então, é dito que as amostras apresentadas ao modelo são treinadas para aprender uma certa função objetivo ou função alvo.

[2] Repare que a uniformidade da entropia é em relação à GLCM e da energia é em relação a própria imagem.

Uma vez que este trabalho visa a determinação de uma função que represente a estimativa do número de pessoas em uma imagem, tal que o aprendizado se dá pelo treinamento utilizando alguns *features* relacionados a borda e a textura, o modelo usado será o de regressão linear multivariada, que pode ser descrito matematicamente da seguinte forma: Dado um conjunto de dados composto por um vetor de n *features* $\left\{\mathbf{x}_j^{(i)}\right\}$, tal que $j = 0, 1, 2, ..., n$, $i = 1, 2, 3, ..., m$, e sendo m o número de exemplos (ou observações) — que estabelecem uma correspondência unívoca com os elementos do conjunto da variável dependente $\left\{y^{(i)}\right\}$ — o objetivo da regressão é prever o valor de y dado um novo valor de $\mathbf{x}$. A função de regressão (ou hipótese) é uma combinação linear das variáveis de entrada:

$$h_\theta = \theta_0 + \theta_1 \mathbf{x_1} + \theta_2 \mathbf{x_2} + \cdots + \theta_n \mathbf{x_n} \tag{3.14}$$

Também representada na forma compacta (ou matricial):

$$h_\theta\left(x\right) = \mathbf{X}\theta, \qquad \mathbf{x}_0 = 1 \tag{3.15}$$

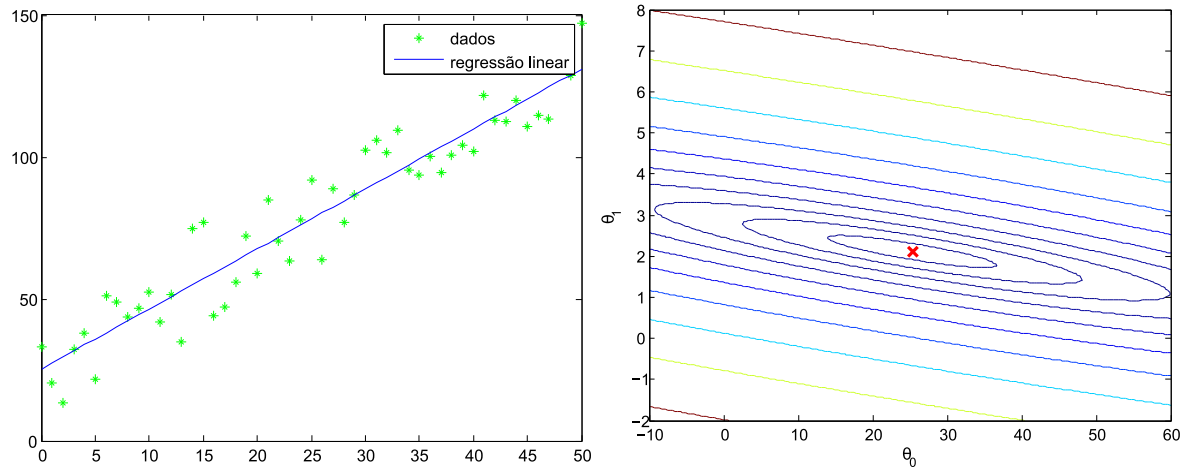
Os parâmetros θ_j são calculados para minimizar uma função de custo $J\left(\theta\right)$ que mede o erro quadrático médio entre o valor estimado e o valor real. A função de custo é dada por:

$$J\left(\theta_0, \theta_1, ..., \theta_n\right) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \tag{3.16}$$

Portanto, os parâmetros de regressão θ_j são a solução do problema. Para o caso univariado, por exemplo, o problema de otimização pode ser geometricamente interpretado como a busca pelos valores de θ_0 e θ_1 que produzem a reta que melhor se ajusta ao conjunto de dados. O resultado da aplicação da regressão a um conjunto de dados arbitrário é ilustrada da Figura 8a. A ideia de minimizar a função de custo, ou seja, o erro quadrático médio, pode ser vista geometricamente como a tentativa de minimizar a distância euclidiana de um ponto da amostra ao seu respectivo valor em uma reta de regressão, de modo que a reta produzida representa aquela que acumulou os menores valores para estas distâncias no agregado dos dados.

A Figura 8b representa as curvas de nível da função de custo $J\left(\theta_0, \theta_1\right)$, que é um parabolóide no caso univariado, ou seja, no caso em que são utilizados apenas dois parâmetros (θ_0 e θ_1) para construir a função de regressão. Repare que o par ordenado encontrado para a combinação dos parâmetros de regressão que melhor ajusta a curva aos dados é aquele que se encontra no vértice da função de custo.

Os métodos para a determinação dos parâmetros de regressão ótimos são diversos e podem ser numéricos ou analíticos. Em geral, os métodos numéricos são utilizados para casos em



(a) (b)

Figura 8 – A Figura da esquerda representa o gráfico de uma reta de regressão sobreposto ao conjunto de dados regredido; a Figura da direita mostra as curvas de nível da função de custo $J(\theta_0, \theta_1)$.

que o número de *features* utilizados no modelo é grande[3]. Como este não é o caso deste trabalho, optou-se pela utilização de um método analítico popular, chamado Mínimos Quadrados Linear e que será introduzido a seguir.

Considere o erro (ou resíduo) da estimativa do i-ésimo exemplo como sendo $e^{(i)} = h_\theta(x^{(i)}) - y^{(i)}$. Substituindo a expressão para o resíduo na Equação 3.16, teremos:

$$J(\theta_0, \theta_1, ..., \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left(e^{(i)}\right)^2 \tag{3.17}$$

Reescrita na forma matricial,

$$J(\theta) = \frac{1}{2m} \mathbf{e}^\intercal \mathbf{e} \tag{3.18}$$

Substituindo $\mathbf{e}$ por $h_\theta(\mathbf{x}) - \mathbf{y}$ e $h_\theta(\mathbf{x})$ por $\mathbf{X}\theta$, obtêm-se,

$$J(\theta) = \frac{1}{2m} (\mathbf{X}\theta - \mathbf{y})^\intercal (\mathbf{X}\theta - \mathbf{y}) = \frac{1}{2m} (\theta^\intercal X^\intercal \mathbf{X}\theta - \theta^\intercal X^\intercal \mathbf{y} - \mathbf{y}^\intercal X\theta + \mathbf{y}^\intercal y) \tag{3.19}$$

Uma vez que $\theta^\intercal X^\intercal \mathbf{y} = (\mathbf{y}^\intercal X\theta)^\intercal$ e que a dimensão de $\mathbf{X}$ é $m \times (n+1)$, de θ é $(n+1) \times 1$ e de $\mathbf{y}$ é $m \times 1$, a igualdade se dá entre escalares, e estes, por definição, são iguais a própria transposta. Então, $\theta^\intercal X^\intercal \mathbf{y} = \mathbf{y}^\intercal X\theta$, e assim a Equação 3.19, torna-se:

$$J(\theta) = \frac{1}{2m} (\theta^\intercal X^\intercal \mathbf{X}\theta - 2\theta^\intercal X^\intercal \mathbf{y} + \mathbf{y}^\intercal y) \tag{3.20}$$

[3] Em geral, métodos numéricos são preferíveis a partir de uma quantidade de *features* da ordem de 10^4, pois a computação da operação da inversa de uma matriz começa a ficar custosa (tempo polinomial).

A fim de minimizar os parâmetros de regressão θ, busca-se a combinação desses parâmetros para a qual o gradiente da função de custo é nulo:

$$\nabla J(\theta) = 0, \qquad \therefore \frac{\partial J}{\partial \theta_j} = 0, \qquad j = 0, 1, 2, ..., n \tag{3.21}$$

Desta maneira, utilizando o cálculo matricial,

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{2m} \left(\mathbf{X}^\intercal \mathbf{X} \theta - 2 \mathbf{X}^\intercal \mathbf{y} \right) = 0 \tag{3.22}$$

E, finalmente, isolando θ, obtêm-se a expressão para os parâmetros ótimos de regressão, que é conhecida por Equação Normal:

$$\theta = \left(\mathbf{X}^\intercal \mathbf{X} \right)^{-1} \mathbf{X}^\intercal \mathbf{y} \tag{3.23}$$

Este resultado encontra grande utilidade prática, uma vez que a matriz $\left(\mathbf{X}^\intercal \mathbf{X} \right)^{-1} \mathbf{X}^\intercal$, chamada matriz pseudo-inversa de Moore-Penrose, não precisa ser quadrada para que o sistema linear possua solução.

4 Método Proposto

Este capítulo trata da apresentação de uma prova de conceito a fim de evidenciar o resultado que as técnicas propostas podem alcançar. Para tanto, estuda-se algumas das técnicas utilizadas em processamento de imagens e visão computacional a fim de que seus resultados indiquem caminhos promissores para melhor compreensão do problema.

4.1 Estimação da contagem baseada em detecção de faces

A abordagem baseada em detecção de faces realizada neste trabalho utiliza o algoritmo de Viola-Jones para estimar a quantidade de pessoas em uma imagem. Sua aplicação é feita por meio da *toolbox*[1] de Sistemas de Computação Visual do ambiente de computação numérica *MATLAB*. O resultado da detecção se constitui nas imagens originais com janelas de detecção sobrepostas as faces econtradas. A estimação do número de pessoas contidas na cena é imediata: é igual ao número de janelas de detecção

Primeiramente, o algoritmo foi aplicado em imagens ideais, onde não há oclusão e as faces estão frontalmente dispostas para a câmera. Para este tipo de imagem em especial o algoritmo apresentou resultados robustos, com uma alta taxa de acerto e uma baixa taxa de falsos positivos. Aplicado a um banco de imagens públicas da internet (diponíveis em Padovani (2016)) que atendem aos critérios das imagens ditas ideais (ou triviais), o algoritmo apontou 2 falsos positivos e 5 falsos negativos para uma quantidade total de 191 faces contidas em 40 imagens.

Se utilizado de forma crua, ou seja, sem alterações, a aplicação do Detector em Cascata de Objetos (como é chamado no *MATLAB*) resulta em uma taxa maior de erro: 8 falsos positivos (com os mesmos 5 falsos negativos), o que significa que o algoritmo está classificando como faces objetos com características geométricas e de sombreamento similares a faces. Na busca pela diminuição dessas classificações equivocadas, propôs-se a implementação de duas técnicas de pós-processamento.

A primeira técnica se constitui na verificação detalhada dos objetos encontrados, visando encontrar falsos positivos. Para isso, as imagens foram reprocessadas, mas desta vez apenas nas vizinhanças das regiões onde objetos haviam sido detectados da primeira vez. A segunda técnica tem a finalidade de encontrar sobreposição entre caixas de detecção muito próximas e que na verdade representam a detecção de uma única face. Como o algoritmo de Viola-Jones é processado em multi-escala para detectar faces de diversos tamanhos e cada classificador procura por um conjunto de características na imagem, várias janelas de detecção

[1] Em tradução livre do inglês: caixa de ferramentas

são definidas para a mesma face e o classificador estima que quanto maior a quantidade de janelas maior probabilidade de ali haver uma face. Um critério chamado limiar de fusão (*merge threshold*) pode ser definido para estabelecer um nível de exigência maior ou menor para que uma certa quantidade de janelas represente uma detecção. Por vezes, mais de uma janela de detecção acaba sendo criada para a mesma face por um erro de estimação do classificador. A detecção de sobreposição procura identificar esses casos e manter apenas uma janela.

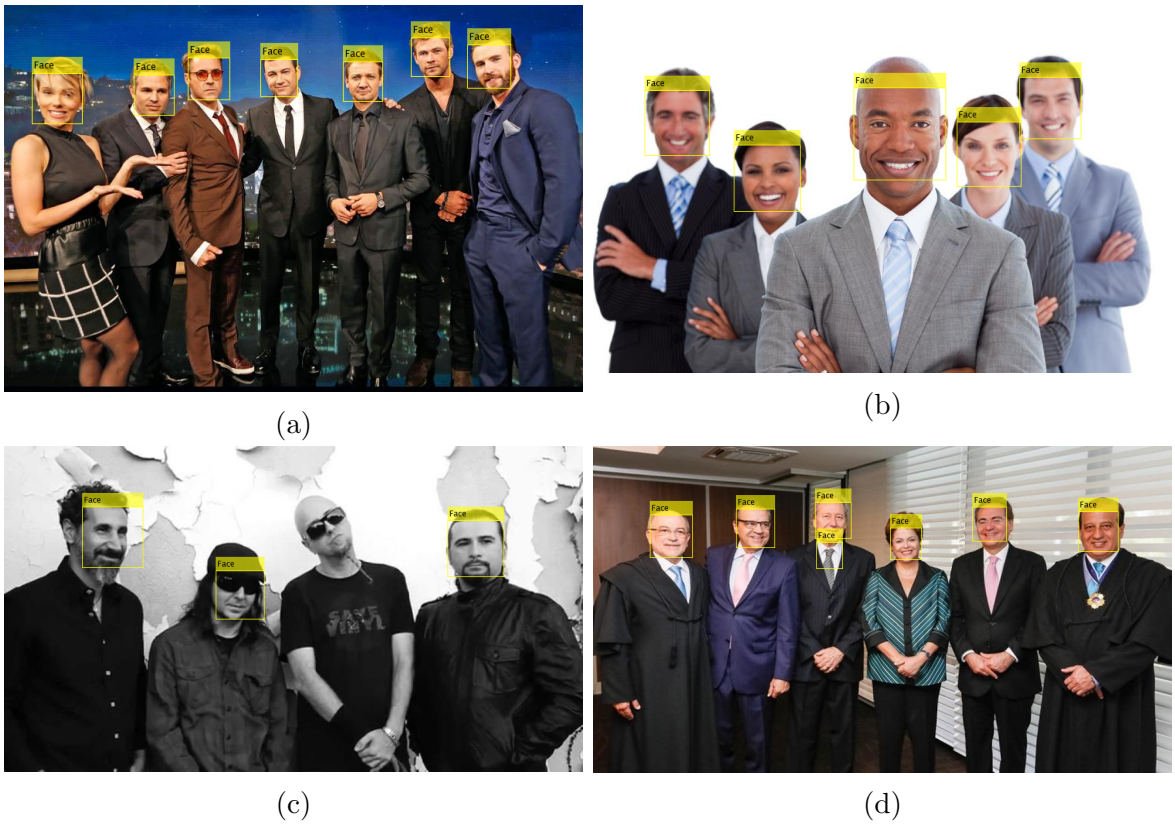
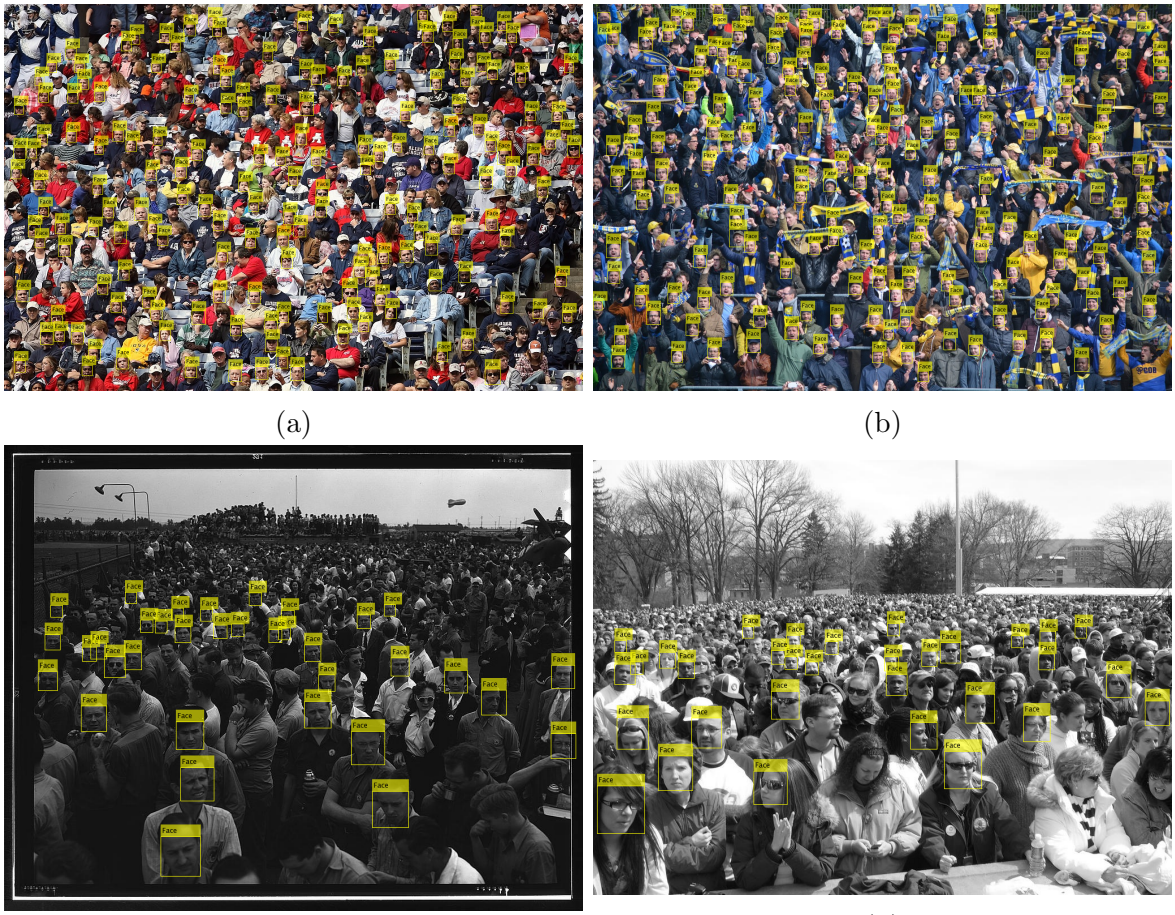


Figura 9 – Exemplos de resultados da detecção de faces baseada no algoritmo de Viola-Jones. Na parte superior encontram-se exemplos livres de erros; na inferior, há dois exemplos com erros. A imagem da esquerda possui um falso negativo e a da direita, um falso positivo.

A Figura 9 mostra alguns resultados das imagens testadas. As imagens localizadas na parte superior são exemplos de imagens sem erros. Para o conjunto de imagens ideais utilizado, esse tipo de resultado representa a quase totalidade dos exemplos. A ferramenta de detecção baseada no algoritmo de Viola-Jones disponível no *MATLAB* garante robustez e conseguiu detectar faces mesmo quando estas apresentavam leve rotação, ou continham óculos — como pode ser visto na Figura 9a — ou ainda se encontrava com algum grau de baixa nitidez — como mostra a Figura 9b. Mesmo com os aprimoramentos realizados, não foi obtida uma solução para casos com faces consideravelmente rotacionadas, como os da Figura 9c. Além disso, apesar da correção da quase totalidade dos erros desse tipo, a verificação de falsos positivos não filtrou o falso positivo encontrado na Figura 9d. Cenas como esta última, em que há uma gravata

contrastando com a cor da camisa, formam contornos e sombreamentos típicos que o algoritmo entende como face.

Outra base de dados utilizada[2] explora cenários com uma grande quantidade de pessoas, em situações não tão favoráveis a esta abordagem de detecção. Outrossim, um banco de imagens com cenários de multidões extremamente densas e sujeitas a um severo grau de oclusão (Center for Reaserch in Computer Vision - University of Central Florida, 2016) também é testado. Alguns exemplos são mostrados na Figura 10.



(a) (b) (c) (d)

Figura 10 – Aplicação do algoritmo de detecção facial Viola-Jones em imagens não triviais.

Podemos observar que a taxa de acerto cai significativamente para estas imagens. Principalmente nas duas imagens da parte inferior. As imagens da parte superior foram prejudicadas, principalmente, pelo fato de ocorrem muitas faces rotacionadas (em qualquer um dos três eixos de rotação), um grande número de pessoas com bonés ou chapéus e, em menor medida, devido a oclusão. Apesar de apresentar uma quantidade considerável de falsos negativos, isto é, ausência de detecções que deveriam ter sido feitas, o algoritmo detectou uma parcela expressiva

[2] Obtida em site fornecedor de imagens livres de direitos autorais (PixaBay, 2016) e disponibilizada para conferência (Padovani, 2016).

da população total das imagens 10a e 10b a ponto de poder ser suficiente para gerar resultados satisfatórios, a partir de ferramentas estatísticas, na estimação do número de pessoas. Entretanto, a posição privilegiada, com pouca oclusão — devido a elevação gradativa dos assentos — e vista majoritariamente frontal em relação à camêra, é rara de ser obtida em situações mais arbitrárias de turba. No caso destas imagens, especificamente, a vantagem da disposição espacial das cabeças é vencida pela perda do propósito de sua contagem, uma vez que este tipo de imagem é tipicamente proveniente de câmeras de estádios ou ginásios, lugares nos quais já há instrumentos para a contagem do número de presentes por mecanismo como catracas, por exemplo.

Os casos representados nas imagens 10c e 10d são os de maior aplicabilidade, uma vez que se dão em áreas abertas e consequentemente de menor possibilidade de controle de contagem, e os de maior dificuldade de tratamento, dado que apresentam, em algumas regiões, oclusão acentuada e baixa resolução (poucos *pixels* por cabeça) — o que avaria a detecção de objetos substancialmente. Estes efeitos são claramente percebidos nessas imagens: as faces em disposição ideal que se encontram na parte dianteira das imagens são detectadas, enquanto as faces na parte traseira, pouco distinguíveis, não o são.

A dificuldade que a abordagem baseada em detecção de faces encontra na estimação do número de pessoas nas imagens altamente povoadas a que este trabalho se põe a analisar, sugere a necessidade da utilização de outras abordagens nas regiões de extrema densidade, que possuem características texturais completamente diferentes das de baixa densidade. Estas características podem e serão exploradas na próxima seção para que se possa viabilizar uma solução alternativa para a estimação da contagem naquelas regiões.

4.2 Estimação da contagem não baseada em detecção de faces

O método desenvolvido neste trabalho tem por essência explorar técnicas — utilizadas em processamento de imagens e visão computacional e aplicadas à detecção e\ou estimação da contagem de pessoas — que sejam complementares, e com isso apontar as vantagens de cada uma delas para que sua utilização combinada produza resultados mais interessantes que sua utilização separada.

O objetivo, dada uma certa imagem, é estimar o número de pessoas que nela está contida. A densidade do números de pessoas em cenários de intensa aglomeração de pessoas raramente é uniforme, ou seja, varia de região para região dentro da imagem. Isto se deve ao fato da própria distribuição espacial das pessoas no ambiente ser heterogênea ou devido aos efeitos de perspectiva causados pelo ponto de vista da câmera. Este efeito sugere que a imagem deva ser analisada em *patches*[3] uniformemente distribuídos pela imagem, porque embora haja variação

[3] Conjuntos disjuntos cuja união é igual à própria imagem e que representam sub-regiões retangulares e de mesmo tamanho.

da densidade entre os *patches*, elas ocorrem suavemente entre *patches* adjacentes.

Desta forma, a estimação é realizada em cada *patch* individualmente, assumindo implicitamente a independência das densidades entre *patches*. Esta suposição é mantida neste trabalho, por efeitos de simplicidade, embora outros estudos tenham explorado as relações inter-*patches*(Idrees et al., 2013). A estrutura básica do modelo, então, se constitui na estimação da contagem nos *patches*, individualmente, por meio de detecção de objetos baseada no algoritmo de Viola-Jones, complementada pela determinação de uma função de regressão baseada em análise de textura e de características locais dos *pixels*, como bordas. É importante salientar que a detecção de objetos não tem seus resultados utilizados como *features* no modelo de regressão. A detecção de objetos tem o intuito de verificar em que regiões da imagem a característica textural, fortemente presente em imagens de multidões, não está presente. A ideia é a de que, se em determinado *patch* foi detectada uma face, esta região está a uma ditância da câmera que favorece a detecção de faces, mas desfavorece a análise de textura.

Como a natureza da textura de regiões da imagem com multidões densas é inerentemente repetitiva, *features* texturais são extraídos da imagem a fim de apontar em quais regiões há essa característica e ainda fornecer informações quantitativas a respeito do número de indivíduos. A seguir, apresenta-se os *features* texturais extraídos.

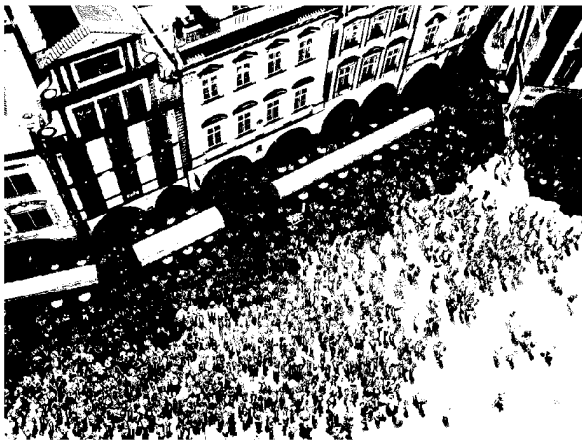
4.2.1 Contagem dos *pixels* das bordas

Primeiramente, a imagem analisada é dividida em n_p *patches*, tal que $n_p = 4^l, l = 0, 1, 2, 3....$ Então um vetor de *features* é criado para receber as características extraídas de cada *patch*. Esse vetor será utilizado como o vetor das variáveis de entrada no modelo de regressão. O primeiro *feature* extraído é o da quantidade de *pixels* nas bordas. Para tanto, aplica-se uma transformação na imagem para torná-la binária. Isto é feito escolhendo um limiar, a partir do qual os *pixels* cuja intensidade seja inferior são zerados (isto é, transformados em preto) e os *pixels* cuja intensidade seja superior são maximizados (isto é, transformados em branco). A binarização tem a tendência de realçar o contraste entre os objetos.

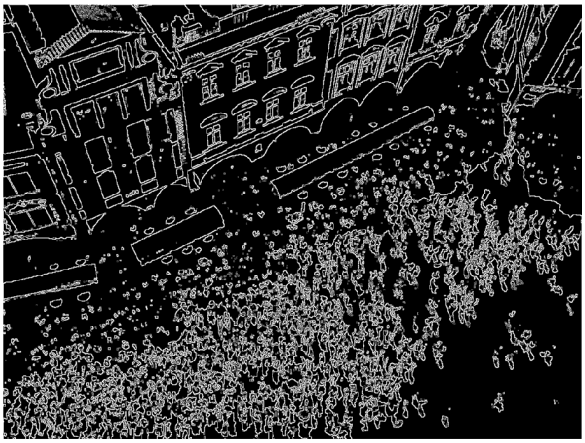
Em seguida, aplica-se o gradiente à imagem binarizada. O realce de contraste promovido pela binarização faz com que o gradiente da imagem binarizada valorize mais as bordas e menos os detalhes internos dos objetos. Deste modo, a Figura 11 representa bem as bordas dos objetos da imagem original, mas nota-se que este processamento é cego quanto a presença de pessoas: as bordas de todos os objetos são detectadas, inclusive as indesejadas. As regiões das imagens onde as multidões se concentram possuem uma característica textural inerentemente complexa e com uma tendência ao surgimento de padrões, se destacando de outras regiões. Essa peculiaridade textural permite que a binarização da imagem também seja explorada a fim de realizar um tipo básico de segmentação.

(a) Imagem original.

(b) Gradiente da imagem original.

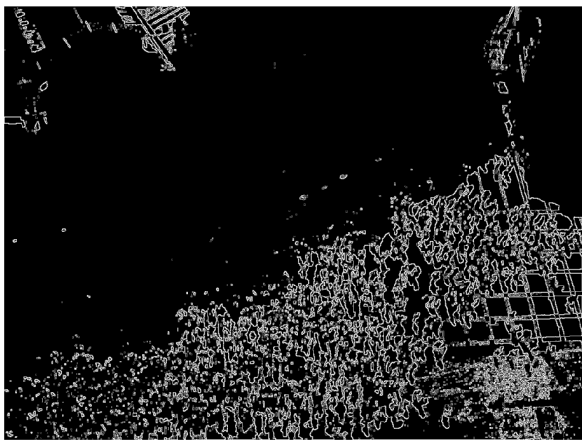


(c) Imagem original binarizada.

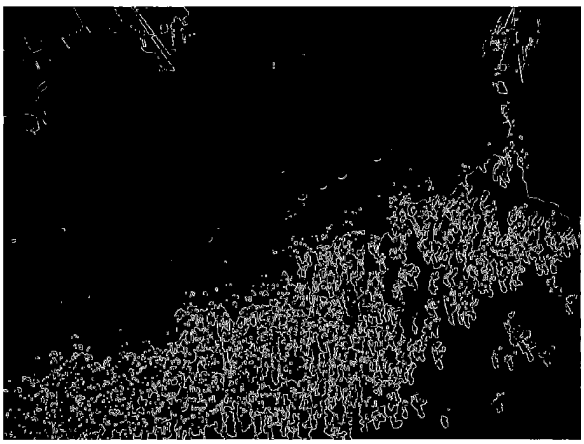


(d) Gradiente da imagem original binarizada.

Figura 11 – Imagem binarizada e não binarizada e seus gradientes. O gradiente da imagem binarizada elimina detalhes irrelevantes, fazendo com que as bordas remanescentes representem apenas o contorno dos objetos.



(a) Máscara de segmentação.



(b) Gradiente da imagem binarizada após a aplicação da máscara de segmentação.

Figura 12 – Efeito da aplicação da máscara de segmentação sobre o gradiente da imagem binarizada.

O gradiente da imagem binarizada com um limiar mais alto faz com que mais *pixels* sejam anulados e por uma análise por inspeção, notou-se que, de modo geral, uma parcela representativa dos *pixels* do plano de fundo foi anulada, enquanto que as regiões contendo multidões sofreu menos anulações. Dessa maneira, foi possível utilizar o gradiente da imagem binzarizada a partir de um limiar maior como máscara para anular *pixels* de plano de fundo do gradiente da imagem binarizada originalmente. Esta técnica de segmentação não é robusta, porém é de simples implementação e mostrou que, em algumas situações, repercute positivamente no resultado final.

Após a aplicação da máscara, a imagem recebe ainda uma operação morfológica para o afinamento das bordas. Tal operação anula um *pixel* se os quatro vizinhos adjacentes forem todos iguais a 1, e portanto remove um *pixel* interior.

A contagem dos *pixels* das bordas é uma técnica que explora características locais da imagem, bem como propriedades da textura da imagem, qualitativas e quantitativas, que informam as regiões mais propensas e menos propensas a conterem aglomerações de pessoas e a própria quantidade desses *pixels* tem uma relação com a quantidade de pessoas que ali estão.

4.2.2 Contagem de picos e vales

A contagem de picos é outra técnica que visa extrair características em nível de *pixels* a partir de elementos texturais da imagem. Como será mostrado no Capítulo 5, a contagem de picos tem correlação positiva com o número de pessoas na imagem. Existem diversas maneiras de obter-se uma contagem de picos de intensidade em uma imagem. A contagem direta, na imagem original convertida para níveis de cinza, simplesmente indicaria onde estão os *pixels* mais claros da imagem e isso por si só não traz informação revelante para a solução do problema, haja vista que *pixels* claros podem ocorrer numa imensa variedade de situações. Uma etapa de pré-processamento interessante no sentido de buscar as informações relevantes contidas na imagem é a aplicação do gradiente. A aplicação do gradiente elimina detalhes irrelevantes da imagem e destaca as bordas, portanto a chance de encontrar informação relevante no gradiente da imagem é maior.

Após a aplicação do gradiente, realiza-se a localização dos picos, que é feita pela varredura de uma janela que percorre a imagem, *pixel* a *pixel*, comparando o valor de todos os *pixels* contidos nela. Assim, um pico será detectado quando o *pixel* atual tiver valor igual ou superior ao *pixel* vizinho de maior intensidade.

A informação que o número de picos em uma região traz pode parecer redundante em relação à informação contida na quantidade de *pixels* de borda nessa região. Entretanto, uma análise mais minuciosa permite constatar que os picos detectados nem sempre se encontram na borda dos objetos, apesar da busca se dar no gradiente da imagem. Outra diferença é que em regiões onde a perspectiva da câmera ou a distância à camera promovem uma resolução excessivamente baixa, bordas não são distinguíveis, mas há contagem de picos.

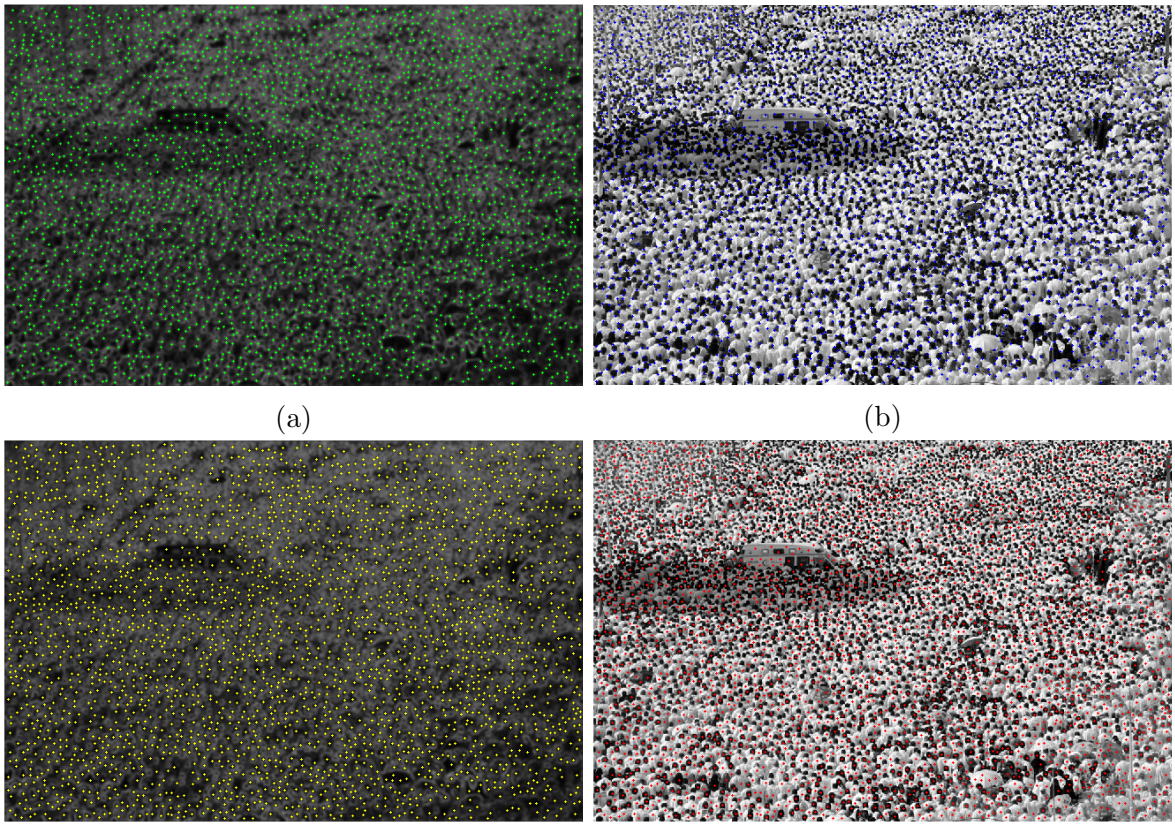


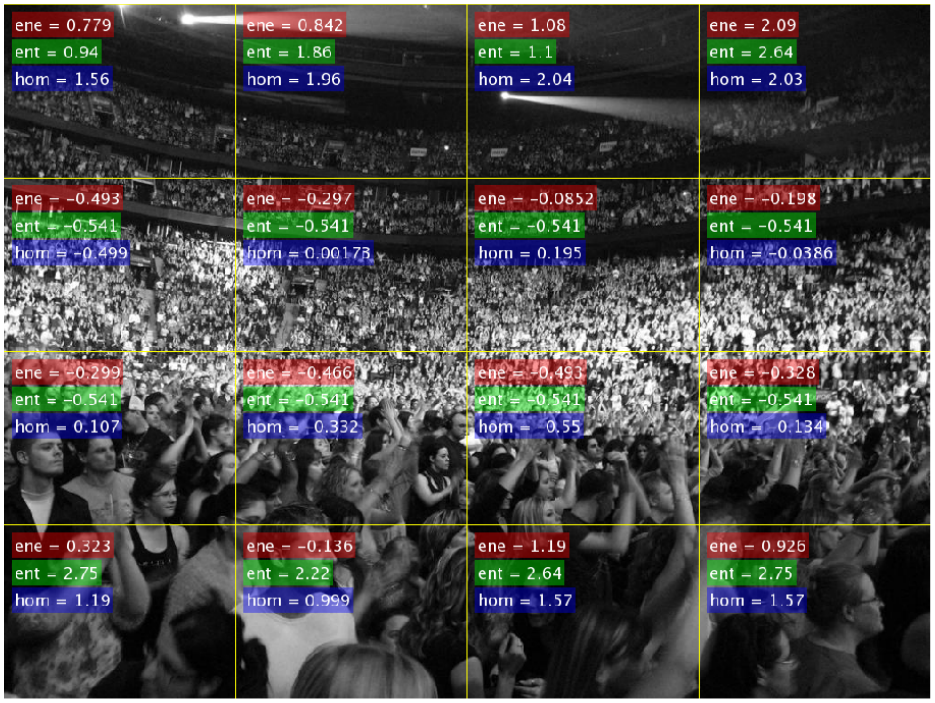
Figura 13 – Detecção de picos e vales na imagem. As imagens superiores apresentam a posição dos picos sobrepostos ao gradiente da imagem e a própria imagem, respectivamente, na esquerda e na direita; as imagens inferiores apresentam a posição dos vales sobrepostos ao gradiente da imagem e a própria imagem, respectivamente, na esquerda e na direita.

Vales são outros elementos de forte característica local que agregam informação espacial. Também são aplicados sobre o gradiente da imagem, que em regiões de densidade populacional elevada, apresentam formas fechadas ou quase fechadas nos lugares onda há cabeças. Em consequência, a aplicação de um detector de vales, que opera de modo análogo ao contador de picos, fará com que um número significativo de cabeças sejam detectadas, se calibrado para uma janela de varredura com dimensão de mesma ordem do tamanho das cabeças.

A Figura 13 mostra a correlação existente entre o número de picos e vales existentes no gradiente de uma imagem com presença de aglomeração intensa de pessoas e a real contagem de pessoas nesta imagem. Nesta Figura, os pontos verdes e azuis correspondem aos picos encontrados no gradiente da imagem e sobrepostos à imagem para comparação. O mesmo é feito com os pontos amarelos e vermelhos, que correspondem aos vales no gradiente da imagem e são, também, sobrepostos à imagem original para comparação com a localização das cabeças.

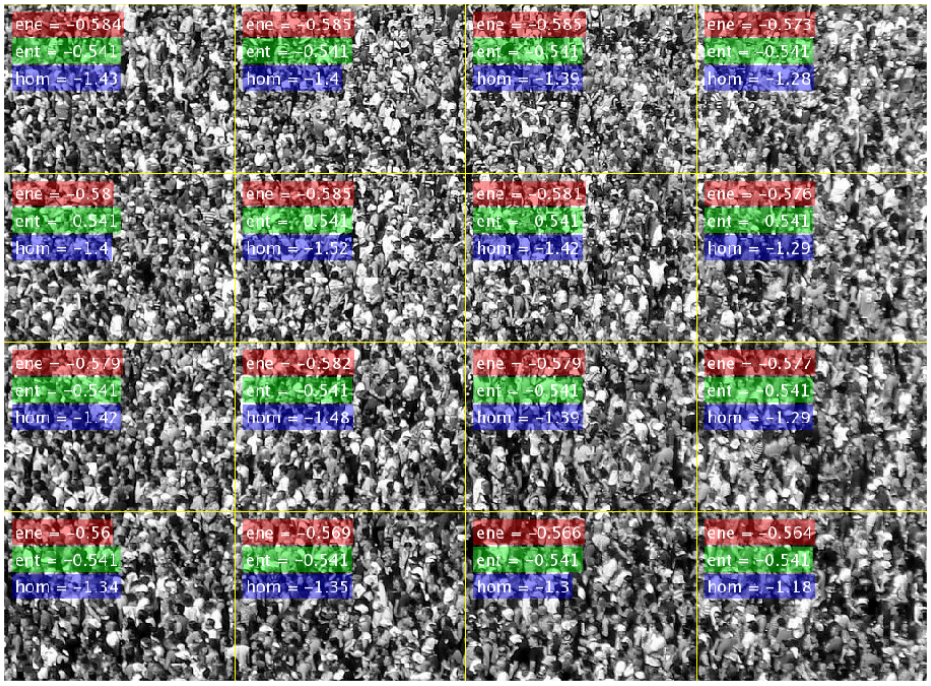
4.2.3 Propriedades da GLCM

A Matriz de Co-Ocorrência de Níveis de Cinza é extensivamente utilizada em análise de textura em imagens, porque é sensível à distribuições espaciais e é capaz de perceber periodicidade, o que é importante para que os *features* texturais possam evidenciar a repetição de certos padrões locais e arranjos regulares em regiões específicas da imagem. Pela forma como é computada, a GLCM captura a repetição de pixels que aparecem aos pares em uma certa direção e com uma certa distância e com isso, estimar métricas texturais como suavidade, rugosidade, regularidade.

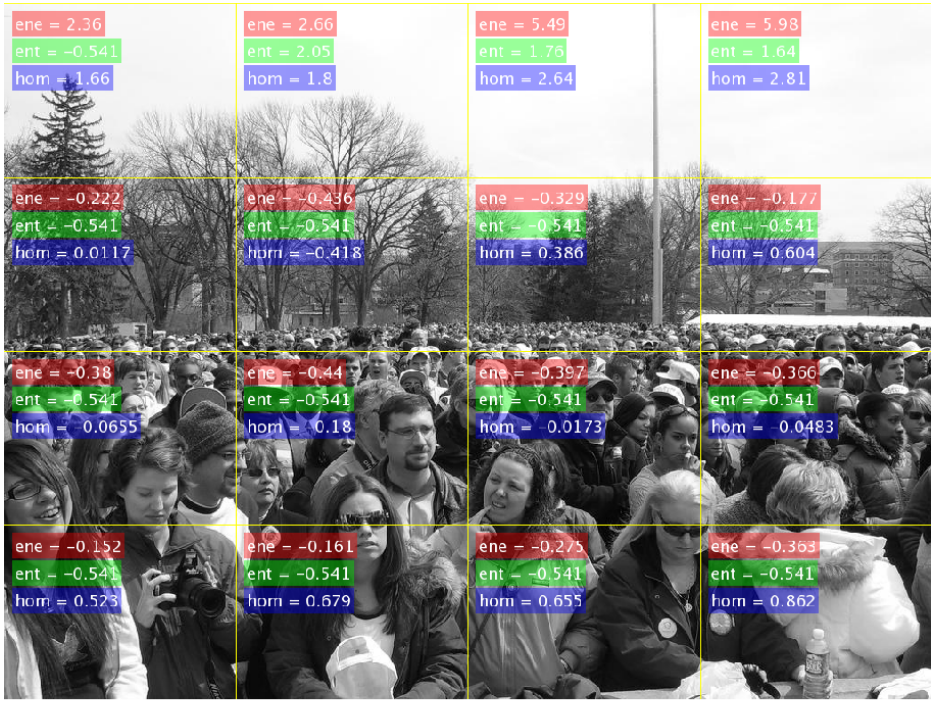


(a)

(b)



(c)



(d)

Figura 14 – Propriedades texturais das imagens a partir das métricas estatísticas da GLCM.

Neste trabalho foram utilizadas três propriedades estatísticas: energia, entropia e homogeneidade. A medida dessas propriedades foi extraída para cada *patch* de cada imagem e alguns exemplos são mostrados na Figura 14. É notável que nas regiões de textura suave e constante como o céu das imagens da direita, 14b e 14d, o valor da energia seja alto — o que mostra uniformidade. Regiões sem o aparecimento de um padrão repetitivo, isto é, com um

grau elevado de aleatoriedade, como os *patches* inferiores da imagem 14a possuem um valor de entropia maior. A homogeneidade é maior para regiões com menor contraste na direção do vetor de deslocamento usado. A expressiva similaridade dos *patches* da imagem 14c é evidenciada também por estas propriedades, que possuem valores que variam pouco entre os *patches* desta imagem.

A escolha do vetor deslocamento é determinante na definição das características texturais extraídas. Neste trabalho, foram realizadas simulações para estabelecer um valor a ser utilizado. A comparação dos resultados, feita com base em algumas métricas de avaliação que serão apresentadas no Capítulo 5, levou a escolha do valor de 11 *pixels* para o tamanho do vetor deslocamento. A forma como este vetor se relaciona com a GLCM está ilustrada na Figura 15.

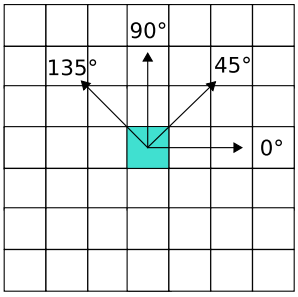


Figura 15 – Ângulos do vetor deslocamento utilizados na GLCM.

Todos os ângulos úteis da GLCM foram utilizados. Apesar de existirem 8 possíveis ângulos para utilização, ângulos explementares[4] produzem matrizes com os mesmos elementos, porém transpostas, então existem 4 ângulos redundantes dentre o total. A Figura 14 foi obtida utilizando o ângulo horizontal.

4.2.4 Treinamento e teste

Como visto, a contagem baseada em detecção é prejudicada pela oclusão e pela baixa resolução dos objetos de interesse (no contexto deste trabalho, faces) que se tornam inerentes às imagens a medida que estas apresentam multidões densas. Além disso, uma vez que a contagem por agrupaento é dependente da disponibilidade de uma certa taxa de quadros por segundo, a contagem baseada em regressão é escolhida, visto que o trabalho tem como foco imagens estáticas.

A contagem por regressão produz uma estimativa através do aprendizado de um mapeamento das características texturais da imagem com a contagem real de pessoas. Para tanto, a primeira etapa do método é a extração dos *features* baseados em textura. Os *features* escolhidos neste trabalho também foram utilizados em outros trabalhos, como por exemplo: *features*

[4] São ângulos cujas medidas subtraídas resultam em 180 graus.

relacionados aos *pixels* das bordas foram utilizados em Davies et al. (1995), Loy et al. (2013) e Chen et al. (2012); a contagem de picos no gradiente da imagem foi testada em Idrees et al. (2013); e as propriedades da GLCM também foram exploradas em Loy et al. (2013), Chen et al. (2012) e Marana et al. (1998). Um dos objetivos deste trabalho é trazer um estudo e apresentar os resultados a utilização das combinações desses *features*.

Após o processo de extração, as características extraídas das imagens são concatenadas em vetores, cada um respectivo a um *feature* e de tal maneira que essas características são referidas a uma região de extração que chamamos de *patch*. Desse modo, o tamanho de cada vetor é igual ao produto do número de imagens pela quantidade de *patches* em que a imagem foi dividida e representa o número de exemplos (ou observações). Como as imagens do banco utilizado estão anotadas, inclusive com a localização das cabeças disponível, é possível associar uma contagem real para cada *patch* e então estabelecer uma correspondência direta entre os vetores de características e o vetor objetivo.

Os vetores são então delimitados para que uma quantidade de exemplos seja utilizada no treinamento e o restante utilizado para teste. Assim, os exemplos correspondentes ao conjunto de treinamento são utilizados como variáveis de entrada no modelo de regressão linear, cujos parâmetros de saída são utilizados como estimadores no conjunto de teste, isto é, a função produzida na etapa de treinamento, pela regressão, é utilizada para estimar a contagem de pessoas nos *patches* das imagens do conjunto de teste. A soma da estimativa de todos os *patches* que compõe uma imagem fornece a estimativa da contagem no número de pessoas desta imagem. O algoritmo de aprendizado utilizado no modelo de regressão é o algoritmo conhecido por Mínimos Quadrados.

O método proposto foi aplicado a um banco de imagens que apresenta características particularmente pouco exploradas. Enquanto uma amplitude de métodos se concentraram em banco de dados com imagens contendo dezenas ou não mais que centenas de indivíduos, o banco de imagens utilizado neste trabalho possui imagens com uma contagem real de pessoas que varia de 94 a 4543. A amplitude de variação desses valores representa um fator dificultador para a análise textural, pois poderá apresentar ruído excessivo ao modelo de regressão nas imagens com menor quantidade de pessoas.

Outra característica deste banco de imagens é a de que as cenas pertencem a uma diversidade rica de eventos como concertos, protestos, estádios, maratonas e peregrinações (Idrees et al., 2013). Embora também seja um fator dificultador, essa diversidade garante robustez aos modelos que obtenham bons resultados.

O conjunto de dados possui um total de 50 imagens, contendo um total de 63705 pessoas, e que foi dividido em dois sub-conjuntos, um de treinamento e um de teste. Utilizou-se a proporção típica de 70% das imagens para treinamento e 30% para validação. Procurou-se dividir as imagens em cada sub-conjunto de modo equilibrado, para que nos dois sub-conjuntos houvesse uma boa representatividade das características gerais do banco de imagens.

5 Resultados e Discussões

Este trabalho promove uma investigação a respeito dos *features* de textura que favoreçam a estimação da contagem de pessoas, a partir de um modelo de regressão linear multivariada. A fim de analisar qual o conjunto de *features* fornece melhores resultados para o modelo de regressão, realizou-se um estudo das técnicas utilizadas para gerar os *features*, no contexto do problema tratado por este trabalho, e como a utilização das diferentes combinações destes influenciam no resultado final. Então, são estabelecidos quais técnicas proporcionaram os melhores resultados, a partir de algumas métricas de desempenho. Estas métricas ajudam a calibrar o método para que seja usada a combinação de *features* no modelo de regressão que efetivamente promove o melhor desempenho. Ademais, é apresentado o resultado da estimação do número de pessoas em cada uma das imagens do conjunto de testes avaliado. Ao fim, os resultados da estimação são mostrados para algumas imagens e é feita uma discussão sobre eles, inclusive comparando-os aos de outros trabalhos.

5.1 Métricas de avaliação

O indicador de desempenho adotado é composto pelo conjunto de métricas que se segue.

Erro quadrático médio (*Mean Square Error* – MSE):

$$\epsilon_1 = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{y}_n\right)^2. \tag{5.1}$$

Erro absoluto médio (*Mean Absolute Error* – MAE):

$$\epsilon_2 = \frac{1}{N}\sum_{n=1}^{N}\left|y_n - \hat{y}_n\right|. \tag{5.2}$$

Desvio médio[1] (*Mean Deviation Error* – MDE):

$$\epsilon_3 = \frac{1}{N}\sum_{n=1}^{N}\frac{|y_n - \hat{y}_n|}{y_n}. \tag{5.3}$$

[1] Proposto em (Conte et al., 2010).

Diferença absoluta normalizada (*Normalized Absolute Difference* – NAD):

$$\epsilon_4 = \frac{1}{N}\sum_{n=1}^{N}\frac{|y_n - \hat{y}_n|}{\sqrt{y_n^2 - \hat{y}_n^2}}. \tag{5.4}$$

O coeficiente de correlação entre o valor real do número de pessoas nas imagens anotadas e a estimativa também é utilizado como métrica de avaliação.

5.2 Resultados

O resultado do desempenho do modelo de regressão utilizando os *features* de textura — GLCM, *pixels* na borda, picos e vales — será mostrado nesta seção. As métricas são baseadas na relação entre a contagem estimada e a contagem real para cada imagem, sendo que o treinamento e o teste são feitos a partir dos *patches*, mas as métricas de desempenho são mostradas por imagem.

Os *features* extraídos das imagens foram utilizados um a um no modelo de regressão para que se pudesse ter uma intuição a respeito da repercussão de cada um no resultado final. A Tabela 3 resume o desempenho atingido. A primeira linha mostra as métricas de desempenho obtidas para as propriedades de energia, entropia e homogeneidade da Matriz de Co-Ocorrência de Níveis de Cinza. A segunda e a terceira linhas mostram o desempenho da contagem de *pixels* nas bordas para duas combinações de limiares de binarização: `Borda1` se refere a utilização de um limiar de 0.20 para a binarização da imagem e de 0.45 para a binarização da máscara de segmentação; `Borda2` utiliza 0.40 e 0.75, respectivamente.

As três linhas seguintes se referem à contagem de picos utilizando três técnicas diferentes. A primeira, `Picos_g`, é baseada no gradiente da imagem borrado por um filtro gaussiano; a segunda, `Picos_h`, é idêntica à primeira, exceto pelo fato de ter sido submetida à uma correção por histograma dos níveis de cinza para que fossem eliminadas regiões de baixa intensidade (o primeiro quintil foi anulado) causando um efeito de segmentação de planos de fundo nítidos[2]; a última também é idêntica à primeira, mas ao invés de ter sido borrada por um filtro gaussiano, foi utilizado um filtro passa-baixas ideal no espectro de frequências da imagem (obtido pela FFT[3]).

A última linha da Tabela 3 se refere à contagem de vales na imagem utilizando a mesma técnica da contagem de picos de `Picos_g`.

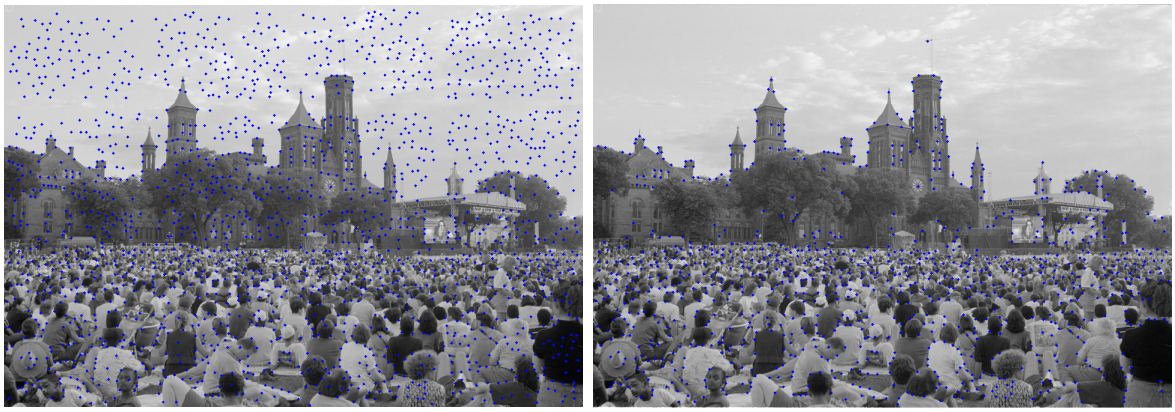
Dentre os *features* utilizados, dois se destacam: as propriedades da GLCM e a contagem de picos no gradiente da imagem. De maneira geral, a contagem de *pixels* de borda é a técnica que apresenta o pior desempenho. Dentre as técnicas utilizadas para a contagem de

[2] Planos de fundo que se diferem substancialmente do primeiro plano, sendo nítida a sua segregação.
[3] Transformada Rápida de Fourier, do inglês *Fast Fourier Transform*.

Tabela 3 – *Features* isolados

	MSE	MAE	MDE	NAD	CoefCorr
GLCM	7,3670E+05	634,9630	**0,4763**	**0,3479**	0,6654
Borda1	1,2641E+06	856,7925	0,6718	0,4748	-0,5644
Borda2	1,2120E+06	831,6240	0,6454	0,4623	0,5381
Picos_g	**6,8617E+05**	**591,3257**	0,5986	0,3489	**0,7144**
Picos_h	8,7009E+05	708,6655	0,5603	0,3990	0,5800
Picos_f	8,6239E+05	681,5803	0,5460	0,3976	0,6842
Vales	7,6002E+05	693,6585	0,8863	0,3910	0,6493

picos, a aplicação de um filtro gaussiano no gradiente da imagem foi a que se mostrou mais promissora. Como visam atingir o mesmo objetivo, utilizando um procedimento que provoca o mesmo efeito (de borrar a imagem), era esperado que as *features* do número de picos utilizando o filtro gaussiano e a FFT obtivessem desempenho semelhante, contudo, como mostra a Tabela 3 isso não ocorre. Outrossim, era esperado que a correção de níveis de cinza por histograma, que tem o efeito de eliminar picos no plano de fundo, aprimorasse o desempenho da contagem de picos, todavia isto não foi verificado. O efeito da utilização da correção por histograma está mostrado na Figura 16. Percebe-se, nesta figura, que os picos eliminados pertencem, majoritariamente, ao plano de fundo, porém alguns picos da região de interesse também são eliminados. Os resultados indicam que o *trade-off* entre a segmentação obtida e os picos perdidos não necessariamente compensa, ou seja, a utilização da correção por histograma não implica em uma melhora significativa da estimativa.



(a) (b)

Figura 16 – Imagens com os picos detectados sobrepostos. A imagem da direita recebeu correção de níveis de cinza por histograma e com isso retirou picos do plano de fundo.

Na Figura 13, do Capítulo 4, é possível observar que a sobreposição dos marcadores sobre as cabeças se dá de maneira mais acurada para os vales do que para os picos. Então, intuitivamente, esperava-se que os vales consistissem em melhor estimação para o número de pessoas em regiões da alta densidade, contudo isto também não se verificou. Uma possível

explicação para este fato se encontra na diversidade do banco de imagens. Se o banco fosse composto por imagens em que cabeças e corpos são facilmente distinguíveis, como a da Figura 13, os vales provavelmente consistiriam em melhores estimadores do que os picos. Na totalidade do conjunto de imagens analisado aqui, verificou-se o contrário.

As características dos *features* que as propriedades da GLCM oferece ao modelo de regressão se diferenciam das características oferecidas pelos outros modelos. A GLCM promove uma visão holística da textura, sendo capaz de detectar a repetição de padrões e inferir mais agudamente sobre as características texturais das diversas regiões da imagem, enquanto que os outros *features* detectam regiões de maior propensão à presença de multidões e infere diretamente sobre a densidade de pessoas. Por possuírem características distintas, mas complementares, a GLCM e os outros *features* foram combinados para aprimorar a estimação. Os resultados dessas combinação são apresentados na Tabela 4.

Tabela 4 – Demais *features* combinados com as propriedades da GLCM

	MSE	MAE	MDE	NAD	CoefCorr
GLCM + B1	7,4781E+05	659,2690	0,4826	0,3553	0,6869
GLCM + B2	8,0621E+05	704,5709	0,4866	0,3706	0,6490
GLCM + Pg	**5,8369E+05**	**540,5078**	0,4545	**0,3098**	**0,7864**
GLCM + Ph	6,5016E+05	587,7032	**0,4361**	0,3249	0,7760
GLCM + Pf	7,1480E+05	624,8568	0,4807	0,3465	0,6841
GLCM + V	5,9525E+05	592,6444	0,4748	0,3239	0,7575

Nota-se que a combinação de melhor desempenho é aquela em que o *feature* combinado apresentou o melhor desempenho individual. Nota-se, também, que muito embora a contagem de picos com correção por histograma não tenha apresentado um bom desempenho individual, sua combinação com a GLCM promove uma melhora substancial em suas métricas. Na realidade, essa melhora pode ser observada para todos os outros *features* (apesar da combinação supracitada ter sido aquela que obteve o segundo melhor desempenho), o que sugere que a combinação dos *features* gera um ganho para os resultados devido às características desses *features* estarem se complementando.

O desempenho da combinação com `Picos_h` é ligeiramente superior àquele da combinação com a contagem de vales. Entretanto, a contagem de vales será usada com GLCM para se combinarem com os demais *features* por oferecer à regressão informações diferentes e mais complementares que aquelas oferecidas por `Picos_h`. Os resultados para essas combinações são mostrados na Tabela 5.

Nota-se que a adição da contagem de vales melhorou o desempenho de algumas das métricas para todas as combinações com as técnicas de contagem de picos. A combinação que gerou o maior salto nas métricas foi aquela que envolve `Picos_f`, entretanto, as combinações envolvendo `Picos_g` e `Picos_h` ainda se mostraram com as melhores métricas. Exatamente

Tabela 5 – Contagem de picos combinados com a GLCM e a contagem de vales

	MSE	MAE	MDE	NAD	CoefCorr
GLCM + V + Pg	5,7721E+05	**540,3210**	0,4562	**0,3091**	0,7883
GLCM + V + Ph	**5,5479E+05**	570,9080	0,4357	0,3147	**0,8226**
GLCM + V + Pf	5,8759E+05	578,2244	0,4673	0,3182	0,7632
GLCM + V + Pg + Ph	5,5485E+05	570,1993	**0,4355**	0,3145	0,8224

por este resultado, foi realizada também uma combinação em que as duas técnicas `Picos_g` e `Picos_h` são adicionadas. No entanto, esta combinação não gerou diferença significativa no resultado.

Tabela 6 – Contagem de picos combinados com a GLCM e a contagem *pixels* na borda

	MSE	MAE	MDE	NAD	CoefCorr
GLCM + B1 + Pg	**6,0406E+05**	**573,6551**	0,4702	**0,3229**	0,7902
GLCM + B1 + Ph	6,6847E+05	623,7561	**0,4623**	0,3398	**0,7984**
GLCM + B1 + Pf	6,5213E+05	599,9228	0,4669	0,3324	0,7593
GLCM + B2 + Pg	6,4979E+05	619,2398	0,4666	0,3403	0,7546
GLCM + B2 + Ph	7,3031E+05	670,3292	0,4771	0,3563	0,7486
GLCM + B2 + Pf	7,4460E+05	663,3134	0,4751	0,3561	0,7172

A Tabela 6 mostra o resultado de um procedimento semelhante ao que foi apresentado na tabela anterior, porém com as técnicas de contagem de picos variando sobre a contagem de *pixels* na borda. O que se constata nesta tabela é que a utilização dos limiares para a binarização da imagem e para a binarização da máscara de 0.20 e 0.45, respectivamente produziu métricas melhores que aquelas produzidas pela utilização de limiares de 0.40 e 0.75, respectivamente. Além disso, nota-se que a adição de `B1` à combinação de GLCM e `Picos_f` melhorou seu desempenho, o que não aconteceu para as combinações de GLCM com `Picos_g` e `Picos_h` mostradas na Tabela 4. Isso indica a possibilidade de que as informações trazidas por `B1` podem ser redundantes com àquelas trazidas por `Picos_g` e `Picos_h`, causando uma dependência linear prejudicial ao modelo de regressão. O fato do desempenho da adição de `B1` à combinação com a técnica de contagem de picos que utiliza a FFT ter apresentado melhora sugere que as informações contidas nesses *features* tenham alguma complementariedade.

Tabela 7 – Combinação dos demais *features* adicionados à GLCM e à contagem *pixels* na borda

	MSE	MAE	MDE	NAD	CoefCorr
GLCM + B1 + Pg + Pf	6,2541E+05	597,3497	0,4626	0,3332	0,7840
GLCM + B1 + Ph + Pf	6,9122E+05	638,2690	0,4662	0,3458	0,7916
GLCM + B1 + Pg + Ph	6,7991E+05	640,4370	0,4655	0,3463	0,7645

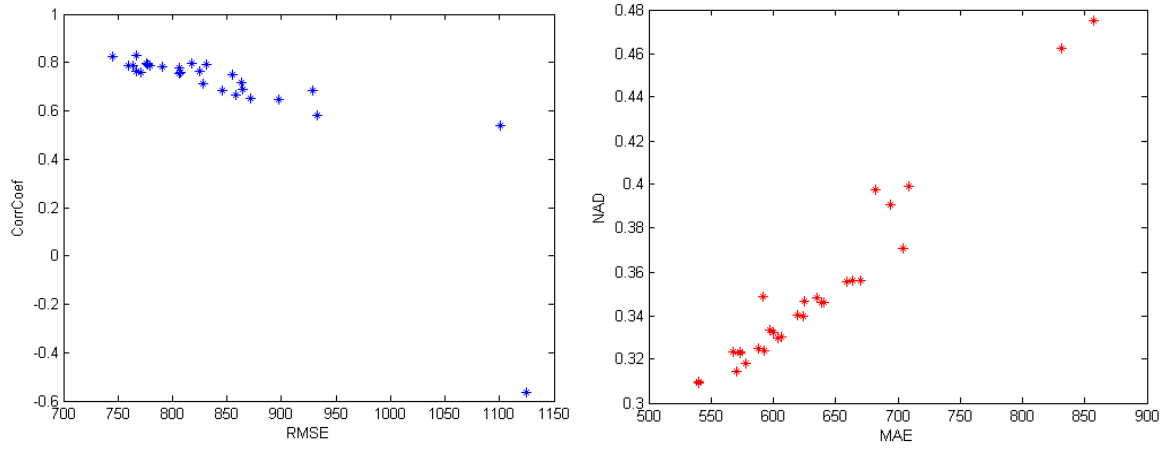
A Tabela 7 representa uma tentativa de combinar as técnicas de contagem de picos entre si somadas à GLCM e à `B1`. O que se constatou é que nenhuma das combinações apresenta ganho de desempenho quanto comparadas às técnicas combinadas isoladamente (Tabela 6).

Tabela 8 – Combinação de todos os *features*

	MSE	MAE	MDE	NAD	CoefCorr
GLCM + V + B1 + Pg	6,0692E+05	573,6903	0,4695	**0,3232**	0,7894
GLCM + V + B1 + Ph	**5,8844E+05**	606,7904	0,4595	0,3303	**0,8290**
GLCM + V + B1 + Pg + Ph	5,8849E+05	604,0847	**0,4589**	0,3295	0,8279
GLCM + V + B1 + Pg + Pf	6,0320E+05	**567,3350**	0,4725	0,3233	0,7980

A última combinação, envolvendo todos os *features*, tem suas métricas apresentadas na Tabela 8. Nela, os vales foram adicionados, a fim de se medir o efeito de sua presença, às combinações de melhor desempenho das Tabelas 6 e 7. O que se verifica é que, de maneira geral, a adição dos *features* de contagem de vale melhora o desempenho geral.

As melhores métricas de desempenho entre todas as avaliadas pertecem à Tabela 5 e foram destacadas. Dois dos casos mostrados nesta tabela são os melhores casos entre todos os estudados: A combinação de *features* da GLCM com a contagem de vales e contagem de picos no gradiente e o mesmo caso com a correção de histograma na contagem de picos. Cada um destes dois casos obteve o melhor desempenho em duas das cinco categorias de métricas de avaliação. As métricas nas quais cada um deles se destaca estão fortemente relacionadas, como pode ser visto na Figura 17. Por ser uma medida de erro representativa e utilizada em outros trabalhos (Idrees et al., 2013), o que permite a comparação direta dos resultados, o caso mostrado na primeira linha da Tabela 5 foi eleito como o que fornece o melhor desempenho dentre os avaliados neste trabalho.



(a) (b)

Figura 17 – Relação das métricas. A imagem da esquerda mostra a relação entre o coeficiente de correlação e a raiz do erro quadrático médio; a da direita mostra a relação entre a diferença absoluta normalizada e o erro médio absoluto.

O resultado da aplicação do método escolhido no conjunto de imagens de teste revela que a imagem com a melhor estimativa erra 94 pessoas na contagem, o que representa um taxa de 4.79% de erro. A imagem com a pior estimativa erra 981 pessoas, o que representa uma taxa de 598.17% de erro. O critério utilizado para mensurar a qualidade dessas estimativas é o erro absoluto normalizado. Essas imagens podem ser visualizadas nas Figuras 18 e 19, respectivamente. A síntese da estimação para todas as imagens do conjunto de teste está representada na Tabela 9.

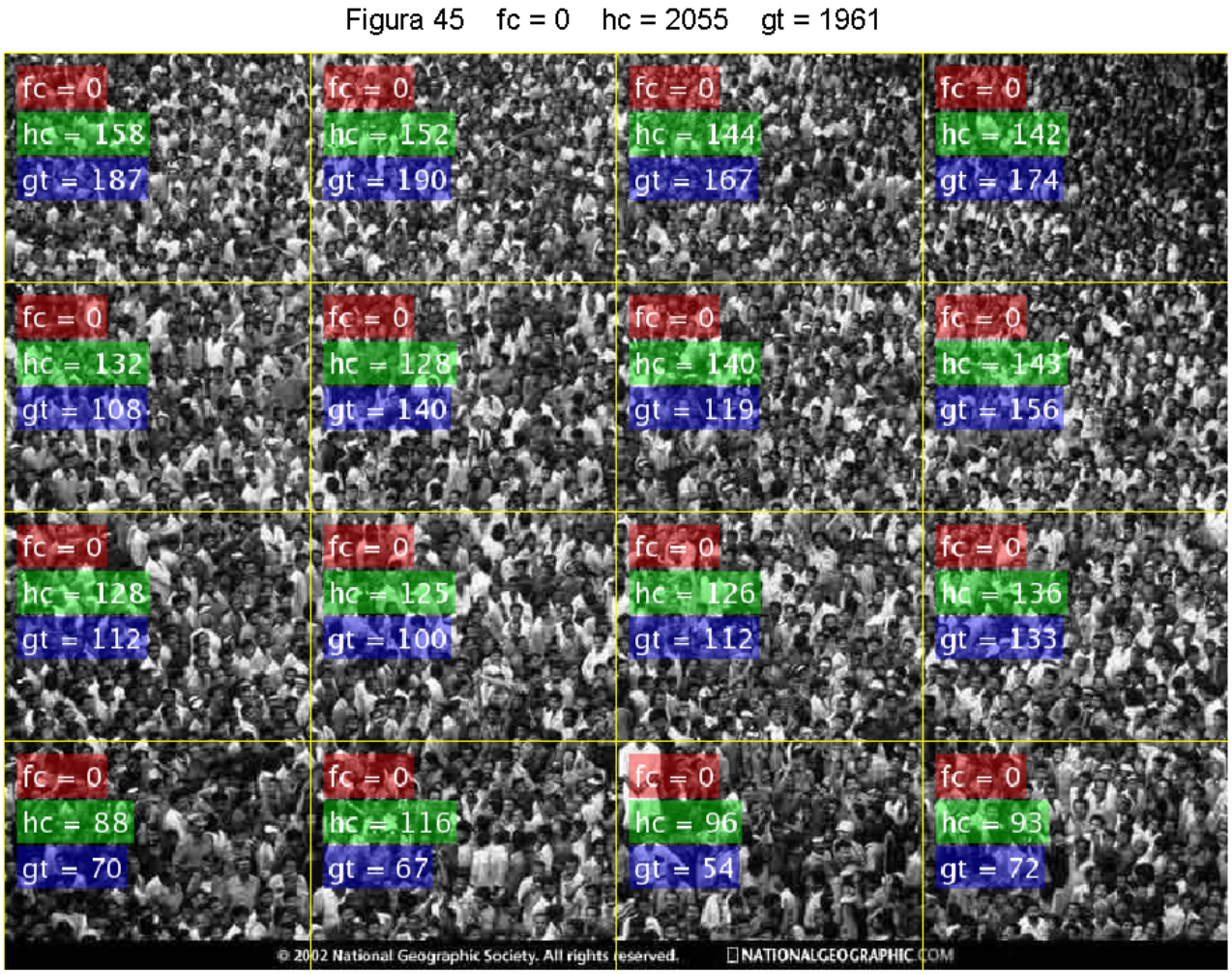


Figura 18 – Imagem com a melhor estimativa.

Nas Figuras 18 e 19 o índice *fc* indica a contagem de faces, o índice *hc* indica a contagem de cabeças e *gt* representa a contagem real.

A Tabela 9 apresenta a estimação para todas as imagens testadas. Tanto para esta tabela quanto para as métricas já apresentadas, foram utilizados os dados diretos da saída do modelo de regressão. Foi verificado que em alguns *patches* o modelo de regressão fornecia

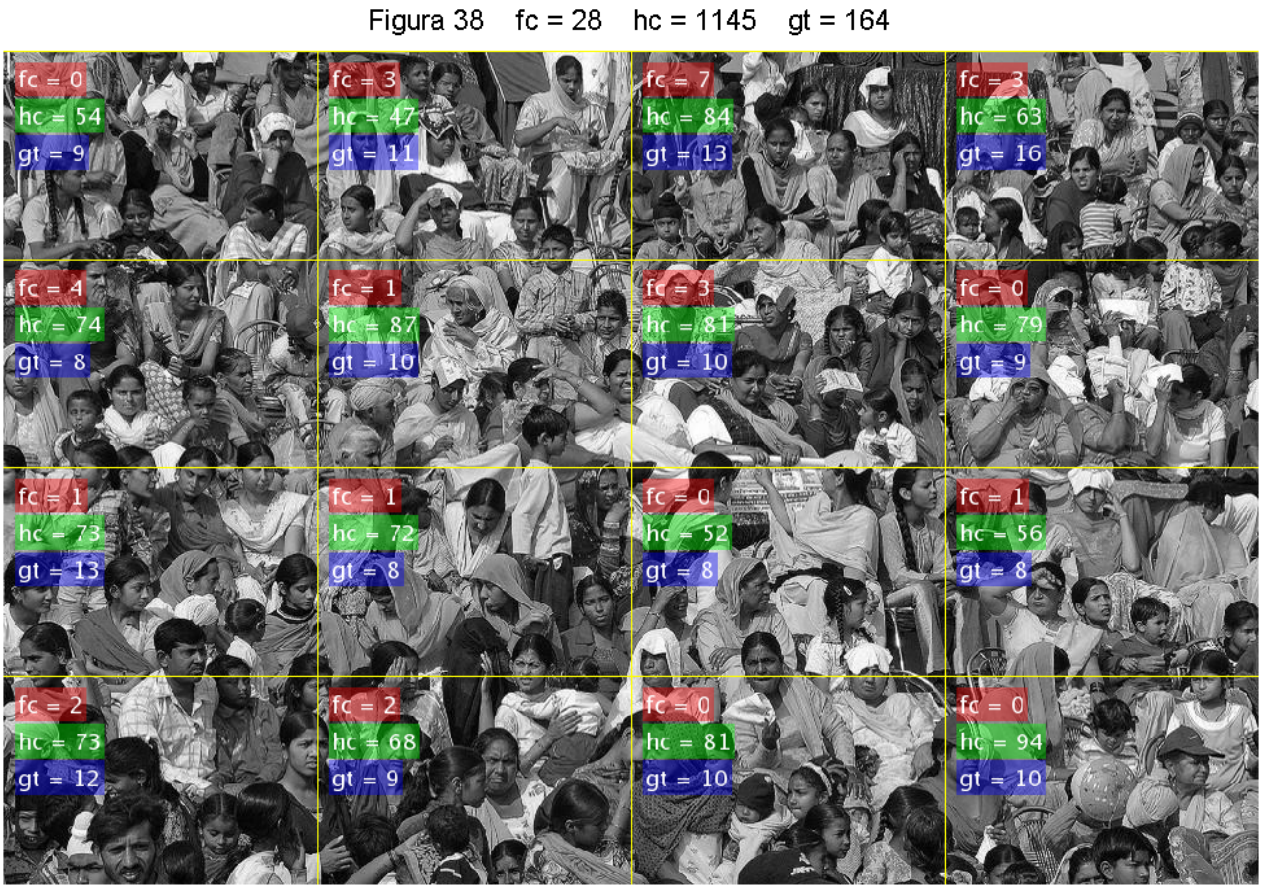


Figura 19 – Imagem com a pior estimativa.

Tabela 9 – Resultado da estimação para cada imagem

Imagem	Contagem real	Contagem estimada	Erro Absoluto	Erro Absoluto Normalizado
36	4631	2614	2017	43,55%
37	361	553	192	53,19%
38	164	1144	980	597,56%
39	754	1086	332	44,03%
40	1858	797	1061	57,10%
41	917	1300	383	41,77%
42	947	1628	681	71,91%
43	209	522	313	149,76%
44	440	881	441	100,23%
45	1961	2054	93	4,74%
46	967	1039	72	7,45%
47	890	835	55	6,18%
48	730	869	139	19,04%
49	2391	1212	1179	49,31%
50	1739	1575	164	9,43%
Média	1263,93	1207,27	540,13	83,68%

valores negativos para a estimação. Muito embora esses valores negativos tenham sido mantidos na apresentação do resultado geral do trabalho (a título de transparência), na apresentação das imagens com as contagens estimadas por *patch*, foram anulados aqueles com contagem negativa, por não possuir sentido físico. Os resultados do mesmo estimador, com os mesmos *features*, porém com a anulação de contagens negativas, é mostrado na Tabela 10.

Tabela 10 – Resultado da estimação para cada imagem, com anulação de *patches* negativos

Imagem	Contagem real	Contagem estimada	Erro Absoluto	Erro Absoluto Normalizado
36	4631	2614	2017	43,55%
37	361	767	406	112,47%
38	164	1144	980	597,56%
39	754	1248	494	65,52%
40	1858	1069	789	42,47%
41	917	1300	383	41,77%
42	947	1628	681	71,91%
43	209	543	334	159,81%
44	440	910	470	106,82%
45	1961	2054	93	4,74%
46	967	1039	72	7,45%
47	890	945	55	6,18%
48	730	895	165	22,60%
49	2391	1255	1136	47,51%
50	1739	1590	149	8,57%
Média	1263,93	1266,73	548,27	89,26%

Também foi feita uma análise de desempenho das imagens separadas e classificadas em grupos, segundo sua contagem real total, a fim de observar em qual dos grupos, separados em tercis, o método atinge o melhor resultado. O resultado desta comparação é mostrado na Figura 20 e revela que o método aqui proposto apresenta melhor desempenho nas imagens do conjunto de teste que não são nem as mais povoadas, nem as menos, variando em uma faixa de 500 a 1000 pessoas por imagem.

Por fim, o método proposto é comparado com os resultados obtidos por outros métodos para o mesmo banco de imagens. O método proposto por Rodriguez et al(Rodriguez et al., 2011) é baseado em detecção de cabeças, enquanto que o método apresentado por Lempitsky e Zisserman(Lempitsky & Zisserman, 2010) utiliza SIFT[4] *features*. O método proposto por Idrees et al (Idrees et al., 2013) foi apresentado no Capítulo 2. A comparação entre os trabalhos está representada na Tabela 11.

[4] Sigla em inglês para Características Invariantes a Transformações Geométricas (Scale-Invariant Feature Transform)

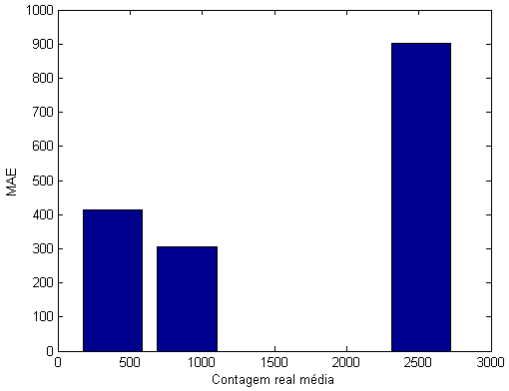


Figura 20 – Desempenho do método dividido em relação à contagem total em tercis.

Tabela 11 – Erro Absoluto Médio comparado com outros trabalhos

Método	MAE
Idrees et al	419.5
Rodriguez et al	655.7
Lempitsky et al	493.4
Proposto	540.3

A comparação permite concluir que o método proposto aqui encontra validade no mundo científico. Entretanto, é mister mencionar que o método proposto em (Idrees et al., 2013) passou por uma validação cruzada 5-fold e por isso sua métrica de avaliação possui maior robustez. Todos as métricas e estimativas apresentadas neste trabalho se referem ao conjunto de teste, composto por 30% do total das imagens analisadas. Nos trabalhos citados, a validação cruzada 5-fold permitiu que todas as imagens fossem testadas.

6 Conclusões e trabalhos futuros

6.1 Conclusões preliminares

O presente trabalho abordou o problema da estimação da contagem de pessoas em imagens de multidões. Primeiro, tratou-se das vantagens trazidas pela automatização do processo de estimação da contagem de pessoas, como agilizade, segurança e disponibilização de dados para análises posteriores, envolvendo análise do comportamento de multidões, por exemplo. Devido às dificuldades que algumas abordagens apresentam ao tratar de imagens que contêm uma alta densidade populacional — principalmente devido à baixa resolução dos objetos de interesse e ao elevado grau de oclusão encontrados em imagens dessa natureza; e por se basearem na detecção dos objetos ou na detecção de trajetórias coerentes — foi proposta uma abordagem baseada em características locais de *pixels* e em análise de textura para efetuar a estimação do número de pessoas nessas imagens.

A partir de um estudo das técnicas de extração de características não baseadas em detecção de faces, com o qual obteve-se o respaldo da literatura especializada, um conjunto de *features* locais e globais, baseados em *pixels* e em textura, foi estabelecido para alimentar um modelo de regressão linear multivaridada que cumpre estimar a contagem de pessoas na imagem.

O método apresentado tem o propósito de complementar a atuação de outros métodos baseados em detecção, ainda que a combinação destes métodos não tenha sido objeto deste trabalho. O método proposto foi idealizado para trabalhar com técnicas que foram escolhidas visando a extração de informações texturais das imagens, o que reflete um ganho de representação das características inerentes à multidões densas.

Os resultados gerados pelo modelo revelam que, de fato, os *features* utilizados parecem acompanhar a contagem real de pessoas. Nota-se, também, que o método obtém o melhor desempenho no tercil intermediário do conjunto de imagens de teste e apresenta o pior desempenho do tercil das imagens mais densas (com mais de 2 mil pessoas), indicando que em algum momento o algoritmo satura e passa a não perceber os detalhes em menor escala da imagem. Esse resultado reflete a incapacidade das transformações morfológicas serem generalizadas. Isso se deve pelo fato de que foi adotado um conjunto único de parâmetros, avaliado como o que apresentou o melhor desempenho para uma avaliação da totalidade do conjunto de imagens. Como consequência, a escolha dos parâmetros fixos não permitem que as peculiaridades de cada imagem sejam exploradas com mais profundidade. Por exemplo, escolhas personalizadas para

o tamanho do vetor deslocamento utilizado na GLCM e dos filtros passa-baixas utilizados na contagem de picos e vales permitiriam que resultados mais acurados fossem encontrados para cada imagem, porém a automação do método seria perdida.

Por fim, estabeleceu-se uma comparação com os resultados obtidos por outros trabalhos no intuito de mensurar se a acurácia obtida pelo método proposto se enquadra dentro das espectativas de desempenho almejadas. A conclusão é pela validade do método apresentado.

6.2 Trabalhos futuros

A seguir, são apresentadas sugestões para trabalhos futuros.

- Utilização da contagem de faces em cada *patch*, baseada no algotitmo de Viola-Jones, como *feature* balizador para os *features* texturais e de pixels no modelo de regressão, a fim de que os *features* se complementem, fornecendo informação para que o modelo pondere em quais regiões da imagem devam pesar mais um ou outro grupo de *features*, extraindo assim as vantagens de cada abordagem.
- Implementação de uma correção de perspectiva a fim de padronizar as densidades encontradas para os *patches*.
- Implementação de parâmetros adaptativos por meio de um algoritmo que seja capaz de ingerir características globais da imagem e assim adaptar parâmetros de calibração como o tamanho dos filtros utilizados, por exemplo.
- Realização de uma comparação de desempenho por meio da utilização de outros estimadores, incluindo estimadores não lineares, mais sofisticados, para a identificação de tendências não lineares de resposta ao *features* utilizados.
- Utilização de Belief Propagation como redes Bayesianas ou MRF[1] para estabelecer relações entre *patches* adjacentes e assim inferir probabilidades da existência de pessoas em cada *patch*.
- Utilização de estrutura multi-escala, principalmente para conseguir extrair características de *pixels* e de textura em regiões onde as faces são reconhecíveis.
- Explorar outros *features* como HOG e SIFT.
- Utilização de validação cruzada k-fold para garantir maior robustez e confiabilidade dos resultados.

[1] *Markov Random Fields*, em tradução do inglês: Campos Estocásticos de Markov.

A Apêndice

Códigos, bancos de imagens e imagens processadas disponíveis em Padovani (2016).

Referências

Bialik, C. (2011). Sizing up crowds pushes limits of technology. *The Wall Street Journal*.

Botta, F., Moat, H. S., & Preis, T. (2015). Quantifying crowd size with mobile phone and twitter data. *Royal Society Open Science*, 2(5), 150–162.

Brown, L. G. (1992). A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4), 325–376.

BU Center for Remote Sensing (1997). Million man march. "`http://www.bu.edu/remotesensing/research/completed/million-man-march/`". [Acessado em: 16/11/2016].

Cariveau, D. (2015). Crowd size estimation.

Center for Reaserch in Computer Vision - University of Central Florida (2016). "`http://crcv.ucf.edu/data/crowd_counting.php`". [Acessado em: 16/11/2016].

Chan, A. B. & Vasconcelos, N. (2012). Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing*, 21(4), 2160–2177.

Chen, K., Loy, C. C., Gong, S., & Xiang, T. (2012). Feature mining for localised crowd counting. In *BMVC*, volume 1 (pp.$\tilde{3}$).

Choi-Fitzpatrick, A. & Juskauskas, T. (2015). Up in the air: Applying the jacobs crowd formula to drone imagery. *Procedia Engineering*, (pp. 273–281).

Conte, D., Foggia, P., Percannella, G., & Vento, M. (2010). A method based on the indirect approach for counting people in crowded scenes. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*.

Crow, F. C. (1984). Summed-area tables for texture mapping. *ACM SIGGRAPH computer graphics*, 18(3), 207–212.

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1 (pp. 886–893).: IEEE.

Davies, A. C., Yin, J. H., & Velastin, S. A. (1995). Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, 7(1), 37–47.

Dodge, Y. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Folha de São Paulo (2013). Entenda como o datafolha calcula multidões. "http://www1.folha.uol.com.br/cotidiano/2013/06/1297948-entenda-como-o-datafolha-calcula-multidoes.shtml". [Acessado em: 16/11/2016].

Folha de São Paulo (2016a). Datafolha vai calcular tamanho de protesto na paulista. "http://www1.folha.uol.com.br/poder/2016/03/1749205-datafolha-vai-calcular-tamanho-de-protesto-na-paulista.shtml". [Acessado em: 16/11/2016].

Folha de São Paulo (2016b). Protesto na av. paulista é o maior ato político já registrado em são paulo. "http://www1.folha.uol.com.br/poder/2016/03/1749528-protesto-na-av-paulista-e-o-maior-ato-politico-ja-registrado-em-sao-paulo.shtml". [Acessado em: 16/11/2016].

Gantz, J. & Reinsel, D. (2011). Extracting value from chaos. *IDC iview*, 1142, 1–12.

Graf, R. F. (1999). *Modern dictionary of electronics*. Newnes.

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5), 786–804.

Haralick, R. M., Shanmugam, K., et al. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610–621.

Idrees, H., Saleemi, I., Seibert, C., & Shah, M. (2013). Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2547–2554).

Jacobs, H. (1967). To count a crowd. *Columbia Journalism Review*, 6(1), 37–40.

Junior, J. S. J., Musse, S., & Jung, C. (2010). Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, 5(27), 66–77.

Kannan, P. G., Venkatagiri, S. P., Chan, M. C., Ananda, A. L., & Peh, L.-S. (2012). Low cost crowd counting using audio tones. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems* (pp. 155–168).: ACM.

Lempitsky, V. & Zisserman, A. (2010). Learning to count objects in images. In *Advances in Neural Information Processing Systems* (pp. 1324–1332).

Lin, S.-F., Chen, J.-Y., & Chao, H.-X. (2001). Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6), 645–654.

Loy, C. C., Chen, K., Gong, S., & Xiang, T. (2013). Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds* (pp. 347–382). Springer.

Marana, A., Velastin, S. A., Costa, L. d. F., & Lotufo, R. (1998). Automatic estimation of crowd density using texture. *Safety Science*, 28(3), 165–175.

Nicolai, T. & Kenn, H. (2007). About the relationship between people and discoverable bluetooth devices in urban environments. In *Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology*, Mobility '07 (pp. 72–78).: ACM.

Padovani, L. H. (2016). Toolbox de estimação do número de pessoas em imagens de multidões. "`https://github.com/luizpadovani/crowd-count`". [Acessado em: 16/11/2016].

PixaBay (2016). "`https://pixabay.com/`". [Acessado em: 16/11/2016].

Portal de notícias G1 (2015). Manifestação na av. paulista reuniu 210 mil neste domingo, diz datafolha. "`http://g1.globo.com/sao-paulo/noticia/2015/03/manifestacao-na-av-paulista-reuniu-210-mil-neste-domingo-diz-datafolha.html`". [Acessado em: 16/11/2016].

Portal de notícias G1 (2016). Manifestantes fazem maior protesto nacional contra o governo dilma. "`http://g1.globo.com/politica/noticia/2016/03/manifestacoes-contra-governo-dilma-ocorrem-pelo-pais.html`". [Acessado em: 16/11/2016].

Rabaud, V. & Belongie, S. (2006). Counting crowded moving objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1 (pp. 705–711).: IEEE.

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

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37–52). Springer.

Solomon, C. & Breckon, T. (2011). *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons.

SSP-SP (2016). Manifestações no estado de são paulo foram pacíficas e reuniram 1 milhão e oitocentas mil pessoas. "`http://www.ssp.sp.gov.br/noticia/lenoticia.aspx?id=37054`". [Acessado em: 16/11/2016].

Store Smarts (2016). "`http://www.storesmarts.com/`". [Acessado em: 16/11/2016].

The Washinton Post (2015). The million man march: Its effect may be debatable. its significance is not. "https://www.washingtonpost.com/business/economy/the-million-man-march-its-effect-may-be-debatable-its-significance-is-not/2015/10/09/5fde11cc-67a4-11e5-8325-a42b5a459b1e_story.html?utm_term=.1a554b8ef936". [Acessado em: 16/11/2016].

Tian, Y.-l., Brown, L., Hampapur, A., Lu, M., Senior, A., & Shu, C.-f. (2008). Ibm smart surveillance system (s3): event based video surveillance system with an open and extensible framework. *Machine Vision and Applications*, 19(5-6), 315–327.

Viola, P. & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2), 137–154.

Wang, M., Li, W., & Wang, X. (2012). Transferring a generic pedestrian detector towards specific scenes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3274–3281).: IEEE.

Wang, Y.-Q. (2014). An analysis of the viola-jones face detection algorithm. *Image Processing On Line*, 4, 128–148.

Watson, R. & Yip, P. (2011). How many were there when it mattered? *Significance*, 8(3), 104–107.

Weppner, J. & Lukowicz, P. (2011). Collaborative crowd density estimation with mobile phones. In *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones*: Citeseer.