Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Estimação de Número de Pessoas em Vídeos

Marina Martins de Miranda

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Engenharia de Computação

Orientador
Prof. Dr. Eduardo Peixoto Fernandes da Silva

Brasília
2017

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Engenharia de Computação

Coordenador: Prof. Dr. Ricardo Pezzuol Jacobi

Banca examinadora composta por:

Prof. Dr. Eduardo Peixoto Fernandes da Silva (Orientador) — ENE-FT/UnB
Prof. Dr. Alexandre Ricardo Soares Romariz — ENE-FT/UnB
Prof. Dr. Camilo Chang Dorea — CIC/UnB

CIP — Catalogação Internacional na Publicação

Miranda, Marina Martins de.

Estimação de Número de Pessoas em Vídeos / Marina Martins de Miranda. Brasília : UnB, 2017.
121 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2017.

1. Estimação do Número de Pessoas, 2. Fluxo Óptico, 3. Agrupamento Hierárquico

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Estimação de Número de Pessoas em Vídeos

Marina Martins de Miranda

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Engenharia de Computação

Prof. Dr. Eduardo Peixoto Fernandes da Silva (Orientador)
ENE-FT/UnB

Prof. Dr. Alexandre Ricardo Soares Romariz
ENE-FT/UnB

Prof. Dr. Camilo Chang Dorea
CIC/UnB

Prof. Dr. Ricardo Pezzuol Jacobi
Coordenador do Bacharelado em Engenharia de Computação

Brasília, 7 de dezembro de 2017

Dedicatória

Dedico à minha encantadora mãe, Kátia, e à minha doce e inesquecível avó, Lygia.

Agradecimentos

Agradeço primeiramente a Deus pela minha vida, pela minha família, pela força para enfrentar as dificuldades, por tudo que me proporcionou e por todas as oportunidades que eu tive.

Agradeço à minha família por todo o amor e por todo o apoio que eu tive durante a minha vida inteira e principalmente durante o período da graduação e da monografia. Agradeço à minha mãe, Kátia, pelo carinho inefável que me deu e por todos os abraços, afagos e acalentos que me ajudaram a seguir em frente; à minha avó, Lygia, por todos os anos de amor e felicidade que eu tive ao seu lado; ao meu pai, Carlos, por todo o amor e cuidado a mim concedidos; ao meu padrasto, Carlos, por toda a alegria que trouxe; ao meu irmão Gabriel, meu eterno companheiro, por todo o apoio e pelas tantas vezes me ajudou; e aos meus dois gatinhos, Katiolino e Carlotina, que tantas vezes me fizeram sorrir quando eu já estava cansada.

Agradeço ao meu namorado, Guilherme, por todo carinho, todo incentivo e todo cuidado. Também agradeço aos seus pais, Lydia e Beto, por todo apoio durante todos esses anos.

Agradeço à Universidade de Brasília, ao Departamento de Ciência da Computação-CIC e ao Departamento de Engenharia Elétrica-ENE/FT pela minha formação profissional e por todo o aprendizado.

Agradeço ao Professor Eduardo, meu Orientador, por todo o apoio e ajuda durante a minha pesquisa e escrita da monografia.

Agradeço a todos os Professores da UnB e a todos os Professores que tive durante a vida acadêmica, por todo o ensinamento passado e por terem incentivado o meu amor ao estudo.

Resumo

Devido ao aumento populacional e à grande disponibilidade de câmeras, existe uma demanda para saber quantas pessoas transitam em certas áreas, seja por questões comerciais ou mesmo por questões de segurança, aumentando o interesse na área da Estimação do Número de Pessoas. A contagem de pessoas, vinda da análise de imagens de vídeo, têm várias aplicações, principalmente para sistemas de vigilância e segurança. Neste trabalho, um método de Estimação de Número de Pessoas é discutido e implementado, utilizando duas técnicas combinadas. A primeira é o Fluxo Óptico de Lukas-Kanade, que é utilizada para estimar o movimento entre os frames do vídeo. Após uma filtragem espacial (ou blocagem) e uma filtragem temporal, há uma inferência de onde os objetos estão localizados. Posteriormente, a técnica de Agrupamento Hierárquico é empregada para agrupar tais objetos em clusters. Por fim, o número de pessoas é mapeado pelo número de clusters distintos. Além disso, uma filtragem de vetores similares foi proposta antes do Agrupamento Hierárquico, gerando dois resultados para cada vídeo testado: um Resultado sem a filtragem e um Resultado com a Filtragem. Foram testados 5 vídeos, com número de pessoas variando de 0 a 5. A acurácia variou de 69,0% a 98,2%.

Palavras-chave: Estimação do Número de Pessoas, Fluxo Óptico, Agrupamento Hierárquico

Abstract

Estimating the number of people based on video imagery has attracted attention due to population growth and high availability of cameras. People counting has several applications for commercial and security reasons. In this work, a method for estimating the number of people in a video sequence is discussed and implemented using two combined techniques. The first technique is the Lukas-Kanade Optical Flow method, which is used to estimate the motion between the frames of the video. After this, a spatial filtering (or blocking) and a temporal filtering are employed to infer where the objects are located. Subsequently, the Hierarchical Clustering technique is used to group such objects into clusters, and the number of people is mapped by the number of different clusters. In addition, a similar vector filtering is proposed before the Hierarchical Clustering, generating two results for each video: a Result without filtering and a Result with Filtering. Five videos were tested, with the number of people ranging from 0 to 5. Results show an accuracy ranging from 69,0% to 98,2%.

Keywords: Estimation of Number of People, Optical Flow, Hierarchical Clustering

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A análise da dinâmica dos grupos de pessoas transeuntes e de seu comportamento é um tópico de grande interesse da sociologia, psicologia, segurança e também da Visão Computacional [1]. Com o aumento populacional e a urbanização, aliados à grande disponibilidade de câmeras, cresce uma demanda para saber quantas pessoas transitam pelos espaços públicos, aumentando assim também a relevância e a importância da Estimação do Número de Pessoas. Dessa forma, a Estimação do Número de Pessoas baseada em imagens de vídeo vindas de sistemas de vigilância e de segurança tem grande utilidade para espaços públicos como aeroportos, estádios, shoppings, rodoviárias e metrôs [2], possuindo aplicações comerciais e aplicações de segurança.

Primeiramente, a Estimação do Número de Pessoas pode ser uma ferramenta inteligente de coleta de informações muito proveitosa na área do comércio [3]. Esses registros são interessantes para os comerciantes, que querem saber se suas lojas estão bem movimentadas, se houve um crescimento médio do número de pessoas que visitam o seu estabelecimento, ou ainda os horários e as épocas em que possuem mais clientes, a porcentagem de clientes em cada horário, quais produtos são mais procurados e o tamanho de filas, auxiliando na decisão de quando contratar mais funcionários, de como planejar a planta da loja, ou de como dispor os produtos na prateleira, por exemplo [1]. Da mesma maneira, são úteis para saber onde passam mais pessoas a fim de decidir onde colocar propagandas. Também são úteis para os clientes, que podem decidir o melhor horário para realizar suas atividades, buscando horários menos cheios para ir à academia ou ao banco, por exemplo. Da mesma forma, podem ajudar no design de espaços públicos [4], provendo diretrizes para a arquitetura de tais espaços, e também contribuem para a operação de ambientes inteligentes [4], podendo ser utilizados para tomar decisões em como separar grupos de pessoas em museus baseado no seu comportamento, por exemplo. Ademais, embora a principal motivação seja contar pessoas em vídeos, há também aplicações para grupos mais gerais de objetos tais como rebanhos (contagem de animais) ou migração de células [5].

Além disso, esses registros de Estimação do Número de Pessoas podem ser consultados como fonte auxiliar em caso de crimes, como roubos ou furtos, apoiando policiais e seguranças e colaborando com a segurança pública. Em aplicações de sistemas de vigilância, os vídeos capturados por um número cada vez maior de câmeras monitorando espaços públicos podem ser assistidos por um número limitado de observadores humanos, e algoritmos de visão computacional podem ser usados para a identificação ou detecção automática de

eventos anômalos como acidentes ou até fuga por pânico (quando há um grande aumento ou grande diminuição no número de pessoas circulando), alertando os observadores [4] e permitindo, por exemplo, a chamada automática de policiais e ambulâncias e auxiliando no design de rotas de evacuação [1].

Com a grande facilidade de acesso a câmeras, a Visão Computacional pode ser empregada para auxiliar nessa estimação. As atividades podem ser monitoradas por meio da detecção e rastreamento de pessoas. Tais sistemas de detecção e rastreamento já existem, mas para os casos acima não é necessária a contagem de indivíduo a indivíduo, de modo que o intuito principal é ter uma estimação do número de pessoas que transitam ao invés de ter um número exato. Entretanto, a estimação do número de pessoas baseada em imagens de vídeo permanece como um problema não trivial em cenários lotados [1]. As principais dificuldades enfrentadas se dão devido às condições de ambiente tais como iluminação não uniforme, projeção de sombras [2], interferência do background e devido ao ângulo de visão da câmera, distância da câmera da região de interesse, ou ainda devido às oclusões entre pedestres, além de ambiguidades visuais como movimento dos membros do corpo.

Dessa forma, um método baseado no trabalho de A. S. Rao *et al* [2] para Estimação de Número de Pessoas é discutido e implementado, utilizando Fluxo Óptico (do termo em inglês *Optical Flow*) e Agrupamento Hierárquico (do termo em inglês *Hierarchical Clustering*). O Fluxo Óptico é interessante pois se baseia na velocidade aparente dos pixels, sendo menos susceptível às variações de iluminação do que outras técnicas comumente utilizadas. A metodologia consiste em cinco fases principais. A primeira fase é o Pré-processamento, no qual os frames do vídeo são preparados para passarem pela Estimação do Movimento; na Estimação do movimento há o cálculo do Fluxo Óptico de Lukas-Kanade [6], gerando os vetores velocidade dos objetos; na Filtragem, ocorrem a Filtragem Espacial para a blocagem dos frames e a Filtragem Temporal para computar o máximo valor de velocidade dentro de certo tempo para cada bloco, ajudando a contornar o problema de oclusão; no Agrupamento Hierárquico há a clusterização das velocidades a fim de agrupar as velocidades que correspondem a um mesmo objeto em um mesmo grupo, em que se assume que picos de velocidade distintos correspondem a objetos distintos; e por fim, na Estimação do número de pessoas, há a contagem de clusters distintos, e se mapeia o número de pessoas pelo número de clusters distintos. Além disso, uma filtragem de vetores similares foi proposta antes do Agrupamento Hierárquico, gerando dois resultados para cada vídeo testado: um Resultado sem a filtragem e um Resultado com a Filtragem. Embora o método possa ser aplicado a outros tipos de vídeo, ele foi testado em cinco vídeos da página CAVIAR Test Scenarios [7], nos quais o número de pessoas varia entre 0 e 5.

Desse modo, esta monografia é dividida da seguinte maneira: o Capítulo 2 apresenta os Fundamentos, explicando os métodos de Fluxo Óptico e Agrupamento Hierárquico, entre outros assuntos relevantes para o entendimento do método proposto; o Capítulo 3 apresenta uma breve Revisão Bibliográfica, discorrendo sobre alguns trabalhos relacionados que contribuíram para a pesquisa na área de estimativa de número de pessoas; o Capítulo 4 se refere à Metodologia utilizada neste trabalho, detalhando os métodos utilizados para a Estimação do Número de Pessoas; o Capítulo 5 apresenta os Resultados para os cinco vídeos testados, assim como algumas métricas de avaliação; e o Capítulo 6 apresenta a Conclusão e algumas propostas para trabalhos futuros.

Capítulo 2

Fundamentos

Este capítulo tem por objetivo explicar os principais assuntos utilizados para resolver o problema de Estimação de Número de Pessoas. Entre eles serão abordados os conceitos de Imagem, Vídeo, Registro para Estimação de Movimento, Fluxo Óptico (do termo em inglês *Optical Flow*), Agrupamento e Agrupamento Hierárquico (do termo em inglês *Hierarchical Clustering*).

2.1 Imagem

A maioria das imagens é gerada por uma combinação de uma fonte de iluminação e da reflexão, absorção, refração ou difração da energia dessa fonte por elementos da cena sendo imageada [8].

Uma cena pode ser definida como uma entidade tridimensional formada a partir da reflexão de energia radiante. Um fluxo de radiação de uma fonte, ao se propagar pelo espaço, pode interagir com a superfície dos objetos de uma cena. A energia absorvida por tais objetos geralmente causa aquecimento. Já a energia refletida carrega informações a respeito desses objetos. Podem-se registrar tais informações com sensores que sejam sensíveis à faixa de radiação refletida [8]. Embora uma cena real seja uma entidade tridimensional, uma imagem pode ser representada por uma função bidimensional, $f(x, y)$, em que x e y são coordenadas espaciais.

Uma imagem pode ser contínua com respeito às coordenadas espaciais x e y, e com respeito à intensidade $f(x, y)$. Assim, para criar uma imagem digital, é necessário converter esses dados contínuos para a forma digital. O processo de digitalização é resultado da amostragem, que é a discretização das coordenadas espaciais, e da quantização, que é a discretização da intensidade. A amostragem consiste em discretizar o domínio de definição da função $f(x, y)$, transformando-o em uma grade de pontos regularmente espaçados entre si. A quantização, por sua vez, define um conjunto finito de níveis de cinza com os quais uma imagem pode ser representada [8].

Na representação de imagens digitais, convenciona-se situar a origem dos eixos de uma imagem no canto superior esquerdo. O processo de amostragem gera uma grade M x N de pontos igualmente espaçados entre si [8]. Em forma de equação, escreve-se a

representação de uma imagem digital como um *array* numérico:

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,N-1) \\ f(1,0) & f(1,1) & \dots & f(1,N-1) \\ \vdots & \vdots & \ddots & \vdots \\ f(M-1,0) & f(M-1,1) & \dots & f(M-1,N-1) \end{bmatrix} \tag{2.1}$$

ou ainda

$$A = \begin{bmatrix} a_{0,0} & a_{0,1} & \dots & a_{0,N-1} \\ a_{1,0} & a_{1,1} & \dots & a_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M-1,0} & a_{M-1,1} & \dots & a_{M-1,N-1} \end{bmatrix} \tag{2.2}$$

Essas equações são maneiras equivalentes de expressar quantitativamente uma imagem digital. Cada elemento dessa matriz pode ser chamado de elemento de uma imagem ou de pixel (do termo em inglês *picture element*).

O nível de cinza ou intensidade de uma imagem monocromática em uma coordenada (x_0, y_0) é dada por

$$l = (x_0, y_0) \tag{2.3}$$

A escala formada por todos os possíveis valores de l é chamada de escala de cinza, que vai do intervalo de 0 a $L-1$, em que $l = 0$ é considerado preto e $l = L-1$ é considerado branco. Os valor intermediários são tons de cinza que variam de preto a branco. Geralmente, utilizam-se 8 bits para representar cada pixel, o que resulta em $2^8 = 256$ níveis de cinza.

Uma imagem colorida é formada no espaço RGB com componentes vermelho, verde e azul, respectivamente. Cada pixel de uma imagem RGB tem três componentes, que podem ser organizados em forma de um vetor coluna:

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \tag{2.4}$$

em que z_1 é a intensidade do pixel na imagem vermelha, z_2 na imagem verde e z_3 na imagem azul. Assim, uma imagem colorida RGB de tamanho $M \times N$ pode ser representada por 3 imagens componentes deste mesmo tamanho, ou por um total de vetores $M \times N$ 3D.

2.2 Vídeo

Um sinal de vídeo é qualquer sequência de imagens que variam com o tempo. Uma imagem estática é uma distribuição espacial de intensidades que permanecem constantes com o tempo, enquanto uma imagem que varia com o tempo possui uma distribuição espacial de intensidades que varia com o tempo. Um sinal de vídeo é tratado como uma série de imagens chamados frames ou quadros [9]. A taxa de frames (do termo em inglês *frame rate*) é a frequência com que os frames são exibidos e criam a ilusão de um vídeo contínuo.

Em um vídeo digital, a informação passa por uma amostragem temporal, uma amostragem espacial e as intensidades resultantes dos pixels são quantizadas. Assim, a amostragem espacial e a quantização são as mesmas descritas para as imagens na seção anterior. Porém, a amostragem do vídeo envolve tirar amostras de uma nova dimensão: o tempo. A dimensão do tempo tem uma direção associada, ao contrário das dimensões espaciais, cujo sistema de coordenadas é artificialmente imposto. O tempo procede do passado em direção ao futuro, com uma origem que existe apenas no momento corrente [10]. O resultado das amostras no tempo são os frames citados anteriormente, ou uma série de imagens completas, cada qual composto de amostras espaciais.

Dessa forma, cada frame de um vídeo pode ser representado por uma função $f(x, y, t)$ ou $I(x, y, t)$, como representado na Figura 2.1. Por exemplo, para o instante $t = t_0$, tem-se:

$$f(x, y, t_0) = \begin{bmatrix} f(0,0,t_0) & f(0,1,t_0) & \dots & f(0,N-1,t_0) \\ f(1,0,t_0) & f(1,1,t_0) & \dots & f(1,N-1,t_0) \\ \vdots & \vdots & \ddots & \vdots \\ f(M-1,0,t_0) & f(M-1,1,t_0) & \dots & f(M-1,N-1,t_0) \end{bmatrix} \tag{2.5}$$

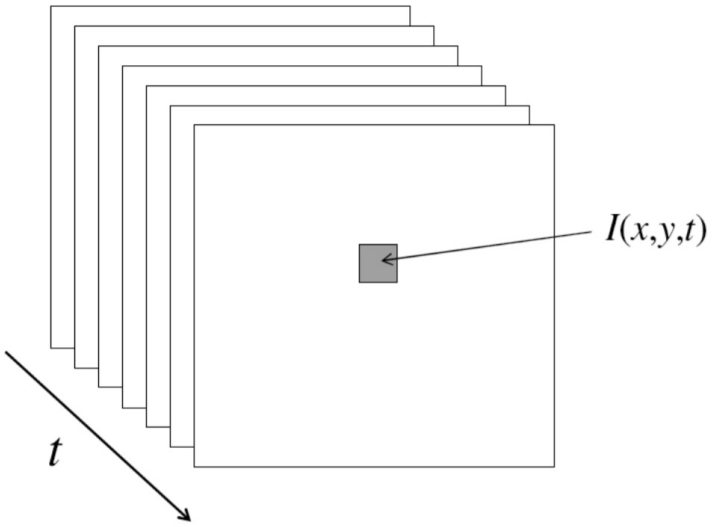


Figura 2.1: Representação de um Vídeo [11]

2.3 Registro para Estimação de Movimento

A Estimação de Movimento em sequências de imagens é muito importante para a Visão Computacional. A extração e uso dos indícios de movimento têm várias aplicações, como a detecção e rastreamento de objetos. Essas aplicações levam a diferentes tipos de representação do movimento, de vetores a silhuetas, que podem ser vistos como um problema de registro (do termo em inglês *registration*) [12].

Pode-se entender a estimação de movimento na análise de vídeos como um registro temporal entre os frames compondo o vídeo. O registro é um processo de encontrar a

transformação espacial que melhor mapeia uma imagem em outra imagem, portanto o movimento durante a sequência pode ser detectado comparando as duas imagens dentro do mesmo espaço. Dessa forma, o registro é um *framework* geral que relaciona as melhores técnicas de estimacão de movimento. A transformação geométrica que relaciona dois frames de vídeo I_1 e I_2 pode ser expressa na Equação 2.6

$$I_2(\hat{x}, \hat{y}) = I_2(f(x, y), g(x, y)) = I_1(x, y) \tag{2.6}$$

Obter as funções de transformação f e g permite descobrir a relação entre os dois frames de modo que o dado inicialmente localizado na posição (x, y) no frame inicial está localizado na posição $(\hat{x}, \hat{y})$ no novo frame. A abordagem mais comum para essa transformação é um campo vetorial, expresso nas Equações 2.7 e 2.8:

$$f(x, y) = x + \Delta_x(x, y) \tag{2.7}$$

$$g(x, y) = y + \Delta_y(x, y) \tag{2.8}$$

em que Δ_x define o deslocamento horizontal e Δ_y representa o deslocamento vertical de um determinado ponto (x, y).

Como um vídeo possui 3 dimensões, a equação 2.6 pode ser estendida para 3 dimensões para a modelagem de estimação de movimento em registro de vídeo. Considerando que I_1 representa o frame no instante t e I_2 representa o frame no instante $t + \Delta_t$, há as novas notações:

$$I_1(x, y) = I(x, y, t) \tag{2.9}$$

$$I_2(f(x, y), g(x, y)) = I(x + \Delta_x(x, y), y + \Delta_y(x, y), t + \Delta_t) \tag{2.10}$$

Dessa forma, obtém-se a equação conhecida por Restrição de Brilho 2.11, representada na Figura 2.2.

$$I(x, y, t) = I(x + \Delta_x(x, y), y + \Delta_y(x, y), t + \Delta_t) \tag{2.11}$$

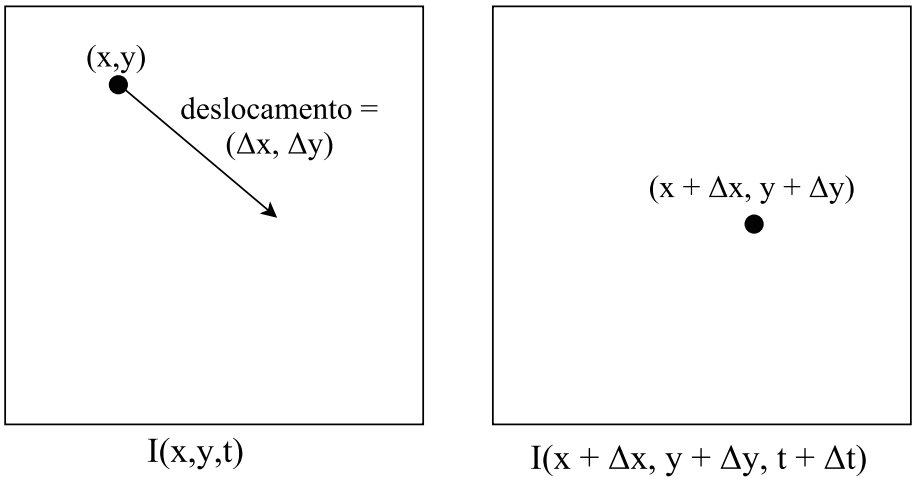


Figura 2.2: Representação da Restrição de Brilho

2.3.1 Fluxo Óptico ou *Optical Flow*

O Fluxo Óptico (do termo em inglês *Optical Flow*) é uma das técnicas mais utilizadas para calcular o campo vetorial nas aplicações de estimação de movimento. Tal método representa a distribuição das velocidades aparentes dos padrões de brilho em uma imagem [13] e provém do movimento relativo entre os objetos e o observador [14].

Utilizam-se as mudanças de brilho entre os pixels nas imagens I_1 e I_2 para tentar calcular o campo vetorial local. Para tanto, assume-se que as mudanças de brilho são originadas por movimentos absolutos dos objetos dentro do campo de visão [12]. Há muitas variações de Fluxo Óptico na literatura, porém todos têm em comum o fato de aproximar o campo vetorial por meio da comparação de pixels com o mesmo brilho ou intensidade.

Dada uma posição (x, y) no instante t, o Fluxo Óptico tenta encontrar a nova localização $(\hat{x}, \hat{y}) = (x + \Delta_x(x, y), y + \Delta_y(x, y))$ em um determinado instante $t + \Delta_t$. Dessa forma, o Fluxo Óptico pode ser entendido como descrito na Equação 2.11. Garantindo a preservação de tal equação 2.11, o Fluxo Óptico é capaz de aproximar um movimento 3D real como uma projeção 2D, resultando em um campo vetorial [12].

Assumindo pequenos deslocamentos espaciais e temporais, a equação 2.11 é aproximada por uma função como uma série de Taylor de primeiro grau, como expresso na Equação 2.12 .

$$\begin{aligned} I(x, y, t) &= I(x + \Delta_x(x, y), y + \Delta_y(x, y), t + \Delta_t) \\ &\simeq I(x, y, t) + \Delta_x \frac{\partial I(x, y, t)}{\partial x} + \Delta_y \frac{\partial I(x, y, t)}{\partial y} + \Delta_t \frac{\partial I(x, y, t)}{\partial t} \end{aligned} \tag{2.12}$$

Dividem-se os dois lados da equação por Δ_t, obtendo a Equação 2.13:

$$0 = \frac{\Delta_x}{\Delta_t} \frac{\partial I(x, y, t)}{\partial x} + \frac{\Delta_y}{\Delta_t} \frac{\partial I(x, y, t)}{\partial y} + \frac{\partial I(x, y, t)}{\partial t} \tag{2.13}$$

Utilizando uma notação mais simples (2.14) para as derivadas parciais:

$$\frac{\partial I(x, y, t)}{\partial x} = I_x, \frac{\partial I(x, y, t)}{\partial y} = I_y, \frac{\partial I(x, y, t)}{\partial t} = I_t \tag{2.14}$$

e adotando a notação 2.15

$$u = \frac{\Delta_x}{\Delta_t}, v = \frac{\Delta_y}{\Delta_t} \tag{2.15}$$

obtém-se a Equação 2.16, conhecida por Equação do Fluxo Óptico [15]:

$$I_x u + I_y v + I_t = 0 \tag{2.16}$$

Assim, o Fluxo Óptico é definido como o vetor velocidade dado pela Equação 2.17 [15],

$$O := \begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} \tag{2.17}$$

em que u e v são as componentes horizontal e vertical do vetor velocidade para cada pixel (x, y) da imagem e são definidos pela Equação 2.15.

Porém, a Equação 2.16 é indeterminada, pois possui infinitas soluções, já que há duas incógnitas e apenas uma equação. De fato, todos os pontos (u, v) dentro da reta definida pela Equação 2.16 são soluções possíveis. Esse problema é conhecido como "Problema de Abertura" [12] e vem da tentativa de estimar o movimento utilizando apenas informação local. Algumas restrições podem ser aplicadas para estimar uma solução única, geralmente relacionadas à suavidade do campo vetorial, seja local ou global. Há dois principais métodos de Fluxo Óptico: o algoritmo de Horn-Schunck [13] e o algoritmo de Lukas-Kanade [6].

Método Horn-Schunck

O método Horn-Schunck [13] introduz uma Restrição Global de Suavidade para resolver o "Problema da Abertura", a fim de permitir a estimação de uma solução única, e apresenta uma implementação iterativa para calcular o Fluxo Óptico para uma sequência de imagens. Assume-se que a velocidade do padrão de brilho varia suavemente em quase todos os pontos da imagem, tentando minimizar as distorções no fluxo.

Nas equações a seguir, as componentes $u(x, y)$ e $v(x, y)$ serão denotadas apenas por u e v para melhor visualização. Assim, seja o Erro associado à Restrição de Suavidade denotado pela Equação 2.18, que penaliza os desvios da suavização na velocidade do fluxo,

$$E_s(u, v) = \iint (u_x^2 + u_y^2 + v_x^2 + v_y^2)\, dx\, dy \tag{2.18}$$

e seja o Erro associado à Restrição de Brilho denotado por 2.19, derivado da Equação 2.11 [12], que representa a taxa de mudança da iluminação da imagem.

$$E_b(u, v) = \iint (I_x u + I_y v + I_t)^2\, dx\, dy \tag{2.19}$$

O método de Horn-Schunk tenta encontrar os valores (u, v) que minimizam a Equação 2.20, que é uma soma ponderada das Equações 2.18 e 2.19 [13].

$$E_{HS}(u, v) = E_b(u, v) + \alpha E_x(u, v) \tag{2.20}$$

em que α é um parâmetro de regularização, de modo que α maiores levam a fluxos mais suaves.

Minimizando a Equação 2.20, foi encontrada uma solução iterativa, como apresentado nas Equações 2.21 e 2.22:

$$u^{k+1} = \overline{u}^k - \frac{I_x(I_x\overline{u}^k + I_y\overline{v}^k + I_t)}{\alpha^2 + I_x^2 + I_y^2} \tag{2.21}$$

$$v^{k+1} = \overline{v}^k - \frac{I_y(I_x\overline{u}^k + I_y\overline{v}^k + I_t)}{\alpha^2 + I_x^2 + I_y^2} \tag{2.22}$$

em que o índice $k+1$ indica a próxima iteração a ser calculada e o índice k mostra o último resultado calculado e em que $\overline{u}$ e $\overline{v}$ são as médias ponderadas de u e v, respectivamente, calculadas na vizinhança do pixel localizado em (x, y).

Esse método tem a vantagem de produzir um Fluxo Óptico mais suave, porém é mais sensível a ruídos e mais oneroso computacionalmente do que outros métodos.

Método Lukas-Kanade

O método de Lukas-Kanade [6] introduz uma restrição de suavidade dentro de uma vizinhança, ou seja, assume-se que o fluxo de vetores é quase constante na vizinhança de um determinado pixel. Essa limitação permite resolver a ambiguidade da Equação do Fluxo Óptico 2.16 aplicando o método dos mínimos quadrados dentro da vizinhança [12].

Dessa forma, uma vizinhança pode ser determinada como uma janela J centrada em um pixel (x, y). Por exemplo, se tal janela J possui $2C + 1$ colunas e $2L + 1$ linhas, J pode ser definida pela matriz 2.23.

$$J = \begin{bmatrix} (x-C, y-L) & \dots & (x-C, y+L) \\ \vdots & \vdots & \vdots \\ \dots & (x, y) & \dots \\ \vdots & \vdots & \vdots \\ (x+C, y-L) & \dots & (x+C, y+L) \end{bmatrix} \tag{2.23}$$

Assim, a Equação do Fluxo Óptico 2.16 deve valer para todos os pixels p_i dentro da janela J, como descrito na Equação 2.24.

$$I_x(p_i)u + I_y(p_i)v = -I_t(p_i), \forall i \in J \tag{2.24}$$

Supondo que há n pixels dentro de J, essas equações podem ser escritas em forma de matriz $Ad = b$, em que A, d e b estão representados em 2.25.

$$A = \begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \vdots & \vdots \\ I_x(p_n) & I_y(p_n) \end{bmatrix}, d = \begin{bmatrix} u \\ v \end{bmatrix}, b = \begin{bmatrix} -I_t(p_1) \\ -I_t(p_2) \\ \vdots \\ -I_t(p_n) \end{bmatrix} \tag{2.25}$$

Esse sistema é sobredeterminado, já que possui mais equações do que incógnitas. Assim, utiliza-se o método dos mínimos quadrados para resolver este sistema, como apresentado nas Equações 2.26 e 2.27.

$$\begin{aligned} Ad &= b \\ A^T Ad &= A^T b \end{aligned} \tag{2.26}$$

ou

$$d = (A^T A)^{-1} A^T b \tag{2.27}$$

em que A^T é a matriz transposta de A.

Isso leva à solução local final, apresentada na Equação 2.28.

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i I_x^2(p_i) & \sum_i I_x(p_i)I_y(p_i) \\ \sum_i I_y(p_i)I_x(p_i) & \sum_i I_y^2(p_i) \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i I_x(p_i)I_t(p_i) \\ \sum_i I_y(p_i)I_t(p_i) \end{bmatrix} \tag{2.28}$$

Ao contrário do método Horn-Schunck, o método Lukas-Kanade é menos afetado por ruídos, porém não pode prover estimação de movimento dentro de áreas muito uniformes devido à sua natureza local [12].

A Figura 2.3 apresenta 2 frames consecutivos de um vídeo e o Fluxo Óptico correspondente nas versões dos algoritmos de Horn-Schunck [13] e de Lukas-Kanade [6]. Além disso, apresenta um zoom nos Fluxos Ópticos para melhor visualização dos vetores velocidade.

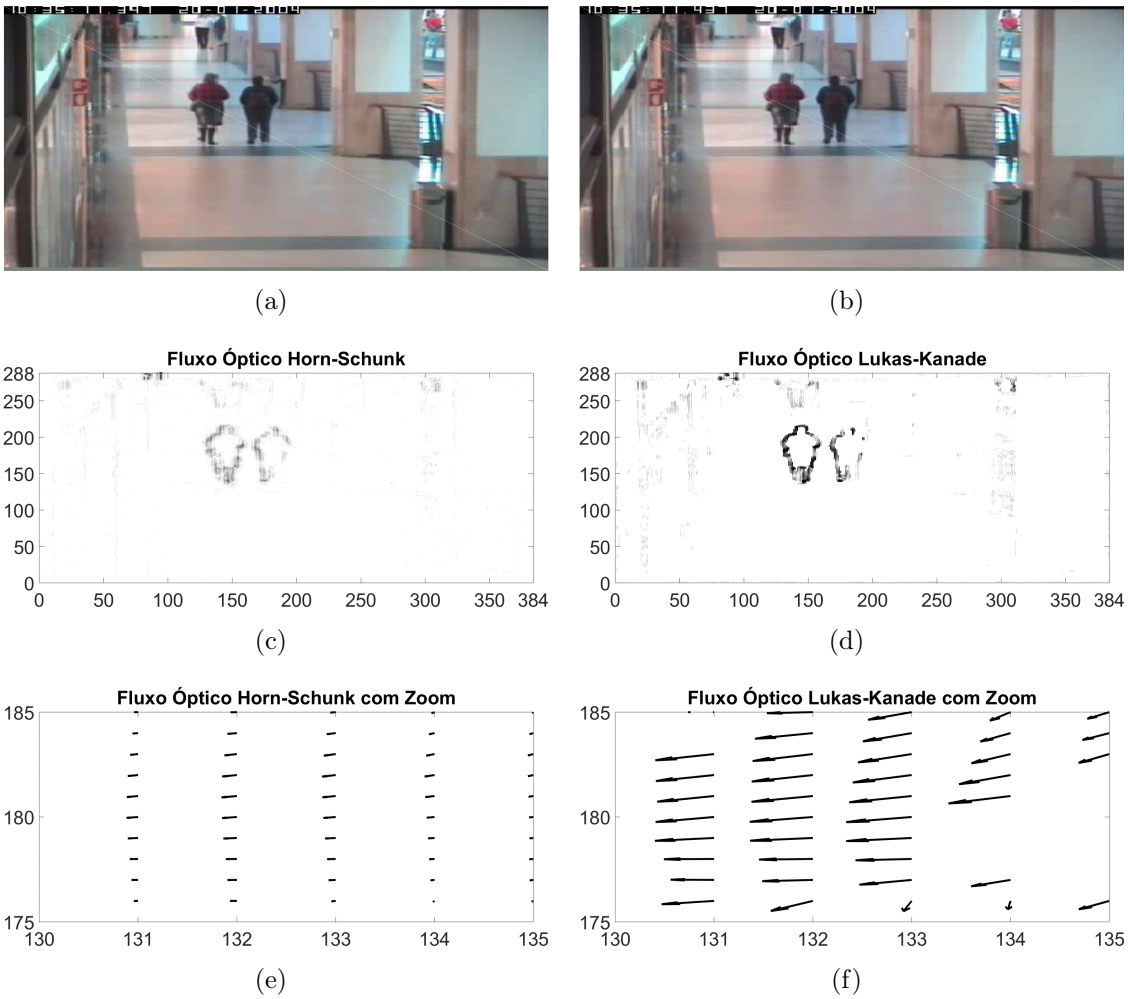


Figura 2.3: Exemplo de Fluxo Óptico: (a) Frame 1 (b) Frame 2 (c) Velocidades do Fluxo Óptico de Horn-Schunck (d) Velocidades do Fluxo Óptico de Lukas-Kanade (e) Velocidades de uma região da imagem (Horn-Schunk) (f) Velocidades de uma região da imagem (Lukas-Kanade)

2.4 Agrupamento ou *Clustering*

Agrupamento (do termo em inglês *Clustering*) é uma técnica computacional para fazer agrupamentos, ou separação de elementos de um conjunto em grupos, de acordo com suas características. Fundamenta-se em colocar elementos similares segundo algum critério pré-determinado em um mesmo grupo [16].

Tal critério é baseado em uma Função ou Métrica de Dissimilaridade, que recebe dois elementos e retorna a distância entre eles. As principais funções ou métricas são:

- Distância Euclidiana [17]

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_i (x_i - y_i)^2} \tag{2.29}$$

- Geometria do Táxi ou Distância de Manhattan [18]

$$d(\vec{x}, \vec{y}) = \sum_i |x_i - y_i| \tag{2.30}$$

- Distância Máxima [19]

$$d(\vec{x}, \vec{y}) = \max_i |x_i - y_i| \tag{2.31}$$

- Distância de Mahalanobis [20]

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})} \tag{2.32}$$

  em que S é a matriz de covariância.

2.4.1 Agrupamento Hierárquico ou *Hierarchical Clustering*

O Agrupamento Hierárquico (do termo em inglês *Hierarchical Clustering*) cria uma hierarquia de relacionamentos entre os elementos e pode ser divido em duas classes [21]:

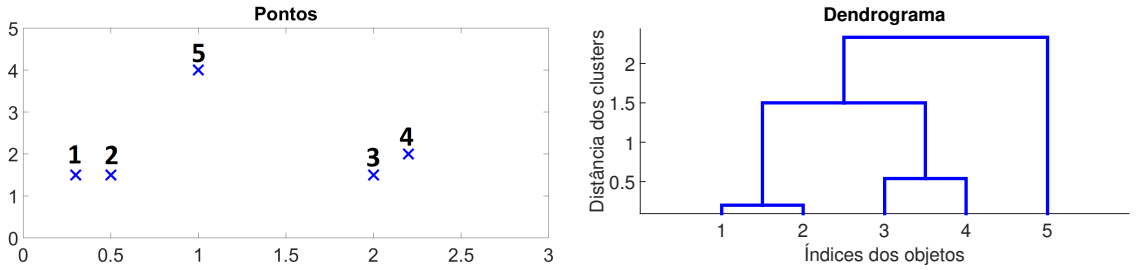
- Agrupamento aglomerativo - No início, cada elemento é considerado como um grupo ou cluster individual, e os grupos são fundidos recursivamente até haver apenas um agrupamento final.
- Agrupamento divisivo - No início, o conjunto de todos os elementos é considerado como um agrupamento único, que é dividido recursivamente até todos os elementos estarem em grupos unitários separados.

Assim, seja S um conjunto de n elementos, $\{E_1, ..., E_n\}$, em que há uma métrica de dissimilaridade entre cada par de elementos, $p_{i,j}$. Segundo Johnson [22], cada Agrupamento Hierárquico produz um esquema de agrupamento hierárquico, que é uma sequência ou hierarquia de partições de S, denotadas por $P_0, P_1, ..., P_{n-1}$ vindas das informações das medidas de proximidade. Para o Agrupamento aglomerativo, por exemplo, a partição P_0 contém todos os elementos em clusters separados; P_{n-1} contém apenas um cluster com todos os elementos; e P_{k+1} é derivado de P_k após unir um par de clusters em P_k.

Dessa forma, o Agrupamento aglomerativo pode ser descrito de acordo com o algoritmo a seguir:

1. Atribua um cluster para cada elemento - correspondente à partição P_0. Então se há N itens, agora há N clusters, cada um contendo um único item.
2. Encontre o par de clusters mais próximos ou mais similares de acordo com uma das Medidas de Distância ou Critérios de Ligação (do termo em inglês *Linkage Criterion*) escolhida, que serão detalhadas posteriormente nesta Seção.
3. Funda-os em um único cluster maior, então agora há um cluster a menos, gerando a partição P_{k+1}.
4. Calcule a distância desse novo cluster para todos os outros clusters, de acordo com uma das Métricas de Dissimilaridade descritas anteriormente.
5. Repita os passos 2, 3 e 4 até sobrar apenas um único cluster de tamanho N - correspondente à partição P_{n-1}.

Com isso, pode-se formar um dendrograma, que é uma árvore hierárquica binária representando graficamente o histórico de fusões dos clusters. Na Figura 2.4(a) é apresentado um conjunto de pontos e na Figura 2.4(b) é apresentado o Dendrograma associado ao Agrupamento Hierárquico correspondente.



(a) (b)

Figura 2.4: Exemplo de Agrupamento Hierárquico: (a) Pontos (b) Dendrograma.

Realizando o algoritmo para o conjunto de pontos da Figura 2.4(a):

Primeiramente, o algoritmo começa com um cluster para cada elemento, ou seja, como há 5 elementos, P_0 será um conjunto com 5 clusters. A notação para cluster será $\{\}$. Assim, $P_0 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$.

Agora deve-se encontrar o par de clusters mais próximos, que são $\{1\}$ e $\{2\}$, e juntá-los em um só cluster, formando $\{1, 2\}$ e criando a partição $P_1 = \{\{1, 2\}, \{3\}, \{4\}, \{5\}\}$, que possui 4 clusters.

Após isso, os clusters mais próximos são $\{3\}$ e $\{4\}$. Eles são unidos em um novo cluster $\{3, 4\}$, formando a partição $P_2 = \{\{1, 2\}, \{3, 4\}, \{5\}\}$ com 3 clusters.

Após a união, há apenas 3 clusters. Entre eles, os mais próximos são $\{1, 2\}$ e $\{3, 4\}$. Eles são unidos em um novo cluster $\{1, 2, 3, 4\}$, formando a partição $P_3 = \{\{1, 2, 3, 4\}, \{5\}\}$ com 2 clusters.

Por fim, há apenas dois clusters, que são obviamente os mais próximos: $\{1, 2, 3, 4\}$ e $\{5\}$. Eles são unidos em um novo cluster $\{1, 2, 3, 4, 5\}$, , formando a partição $P_4 = \{\{1, 2, 3, 4, 5\}\}$, composta por um único cluster de tamanho 5 e o algoritmo termina.

Deve-se notar que a estrutura hierárquica formada pela união entre os elementos foi representada como um dendrograma, que é a forma mais usual de representação dos resultados de algoritmos hierárquicos e mostra a ordem do agrupamento de forma intuitiva. Quanto mais alta a linha ligando dois clusters, mais tarde foi feito seu agrupamento. Logo, a altura da linha ligando dois clusters é proporcional à sua distância [16].

Para esse exemplo das Figuras 2.4(a) e 2.4(b), foi utilizada a Medida de Distância Single-Linkage Clustering 2.33 (a ser explicada a seguir) com Métrica de Dissimilaridade do tipo Distância Euclidiana 2.29.

Entre as Medidas de Distância ou Critério de Ligação (do termo em inglês *Linkage Criterion*), podem-se citar as mais importantes [16] [21] [23] [24]:

- Single-linkage clustering ou Método Mínimo: A distância entre dois clusters é a distância entre os seus pontos mais próximos, como representado na Figura 2.5.

$$D_{min} = \min_{x \in X, y \in Y} d(x, y) \tag{2.33}$$

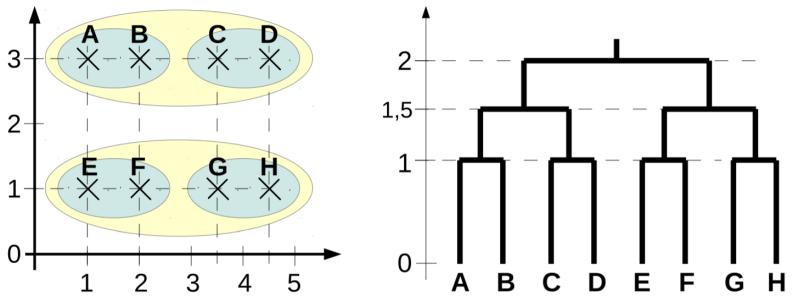


Figura 2.5: Representação de Single-linkage clustering [21]

- Complete-linkage clustering ou Método Máximo: A distância entre dois clusters é a distância entre os seus pontos mais distantes, como representado na Figura 2.6.

$$D_{max} = \max_{x \in X, y \in Y} d(x, y) \tag{2.34}$$

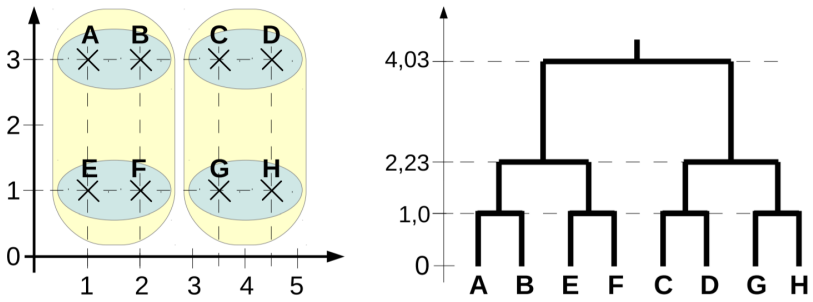


Figura 2.6: Representação de Complete-linkage clustering [21]

- Average-linkage clustering ou Método da Média: A distância entre dois clusters é dada pela distância entre os seus centróides.

$$D_{avg} = \frac{1}{|X|.|Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y) \tag{2.35}$$

em que $d(x, y)$ pode ser qualquer distância dada pelas Métricas de Dissimilaridade descritas nas Equações 2.29 a 2.32.

Capítulo 3

Revisão Bibliográfica

Este capítulo tem por por intuito apresentar trabalhos relacionados que contribuíram para a pesquisa na área de estimativa de número de pessoas.

O trabalho de C. C. Loy *et al* [1] descreve e compara os métodos do estado-da-arte para a contagem de pessoas em imagens de vídeo e provê uma avaliação para eles. Os principais paradigmas podem ser classificados em: contagem por detecção, contagem por agrupamento e contagem por regressão, que vem ganhando interesse devido à maior viabilidade ao lidar com ambientes com um número maior de pessoas. A seguir será dada uma breve explicação sobre cada uma destas estratégias.

Os métodos de **contagem por detecção** (do termo em inglês *counting by detection*) [1] detectam pedestres através de um escaneamento da imagem utilizando um detector treinado com características de imagens locais. Podem-se dividir nos seguintes casos:

- **Detecção Monolítica:** Consiste em realizar o treinamento de um classificador para reconhecer a aparência de um corpo inteiro em um conjunto de imagens ou vídeos de pedestres. As principais características utilizadas para representar a aparência de um corpo inteiro são as Transformadas *Wavelet* de Haar [25], as características de Histograma de Gradiente Orientado (HOG) [26], as *edgelets* [27] e as *shapelets* [28]. Os classificadores mais utilizados são do tipo linear, tais como *boosting* [29], SVM (*Supported Vector Machine*) lineares e *Random/Hough Forests* [30]. Embora a Detecção Monolítica tenha bons resultados em ambientes esparsos, há a dificuldade na identificação de um corpo inteiro no caso de oclusões.

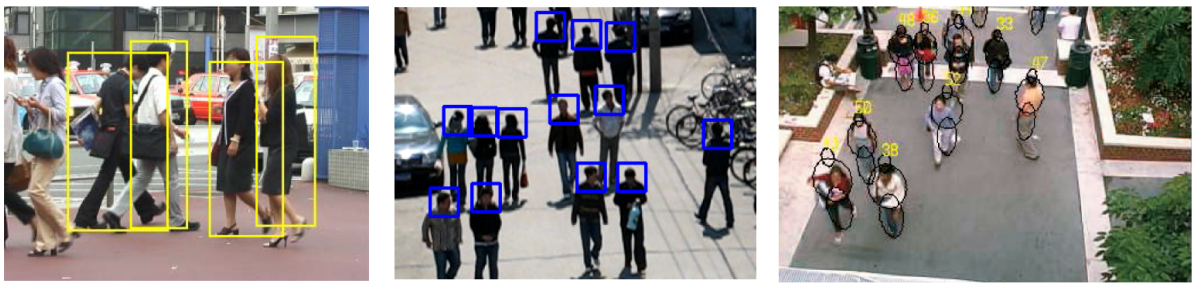
- **Detecção baseada em Partes:** Tenta contornar o problema das oclusões parciais, tentando identificar apenas partes do corpo [31]. Dessa forma, os classificadores são treinados para reconhecer essas partes do corpo humano, principalmente cabeça e ombros, a fim de estimar a quantidade de pessoas [32]. Este tipo de detecção é mais robusto que a detecção monolítica em cenários mais cheios porque não precisa ter a imagem do corpo inteiro.

- **Correspondência de Formatos (do termo em inglês *Shape Matching*):** Tenta aproximar o corpo humano por diferentes formatos e estimar a quantidade destes formatos nas imagens de vídeos. Alguns trabalhos utilizam formatos de corpos compostos por elipses e processos estocásticos para a estimação do número de pessoas [32], enquanto outros trabalhos mais recentes utilizam formatos de corpos

mais realistas e flexíveis, treinando classificadores para reconhecer formatos de modelos combinados de Bernoulli (do termo em inglês *mixture model of Bernoulli*) nas imagens [33].

- **Detecção com Multi-Sensores:** Tenta resolver problemas de oclusão inter-objeto utilizando informação de diferentes campos de visão vindas de várias câmeras. Pode-se empregar uma rede de câmeras para extrair a silhueta de seres humanos e estabelecer limites na quantidade e possíveis localizações de pessoas, como no trabalho de Yang *et al* [34]. Ou ainda estimar o número de pessoas e suas localizações impulsionando restrições geométricas de diferentes campos de visão, como no trabalho de Ge *et al* [35].
- **Tranferência de aprendizado:** Estuda a transferência de detectores genéricos de pedestres para novos cenários sem supervisão humana, enfrentando as dificuldades de resolução, iluminação e diferentes backgrounds. Tenta contornar essas situações utilizando as estruturas do cenário, ocorrências espaço-temporais e tamanho dos objetos [36].

A Figura 3.1 apresenta exemplos de Contagem por Detecção por Detecção Monolítica, Detecção baseada em Partes e Correspondência de Formatos.



(a) (b) (c)

Figura 3.1: Contagem por Detecção [1]: (a) Detecção Monolítica [37], (b) Detecção baseada em Partes [32] e (c) Correspondência de Formatos [38].

Os métodos de **contagem por agrupamento** [1] assumem que uma multidão é composta de indivíduos, cada qual possuindo um padrão de movimento único porém coerente que pode ser agrupado para aproximar o número de pessoas.

Rabaud *et al* [5] fazem uso da contagem por agrupamento, utilizando uma versão altamente paralelizada do rastreador Kanade-Lucas-Tomasi (KLT), que permite popular eficientemente o volume espaço-temporal com um grande conjunto de trajetória de características, para processar o vídeo, obter um conjunto de características rastreadas de baixo-nível e agrupar a trajetória para estimar o número de pessoas no cenário. Já por sua vez, Brostow e Cipolla [39] rastreiam aspectos locais e os agrupam em clusters utilizando agrupamento Bayesiano.

O trabalho de A. S. Rao *et al* [2] foi o artigo principal no qual a implementação deste Trabalho de Conclusão de Curso foi baseada e pode ser classificado como pertencente à classe de contagem por agrupamento, porque utiliza Agrupamento Hierárquico após uma filtragem espaço-temporal para agrupar as velocidades dos objetos e mapear o número

de pessoas. Mais especificamente, utiliza Fluxo Óptico de Horn-Schunck [13] baseado em blocos com filtragem espacial e temporal para obter as velocidades, a fim de inferir a localização de objetos em cenários nos quais há pessoas andando, e posteriormente utiliza Agrupamento Hierárquico para agrupar os objetos, com métrica de distância do tipo Euclidiana, e mapear o número de pessoas pelo número de clusters distintos.

Uma característica interessante sobre os métodos de contagem por agrupamento é que eles não utilizam aprendizado supervisionado como no paradigma de contagem por detecção acima, porém assumem coerência de movimento. Assim, pode haver falsas estimações quando as pessoas ficam paradas articulando os membros ou mesmo caminhando, mas fazendo muitos movimentos com os membros ou ainda quando dois ou mais objetos compartilham trajetórias comuns. Além disso, a contagem por agrupamento só funciona com sequências de frames, e não com imagens estáticas, ao contrário da contagem por detecção e por regressão, que funcionam em ambos os casos.

A Figura 3.2 apresenta exemplos para Contagem por Agrupamento, mostrando os resultados do agrupamento de movimentos coerentes dos métodos de Rabaud *et al* [5] e Brostow e Cipolla [39].

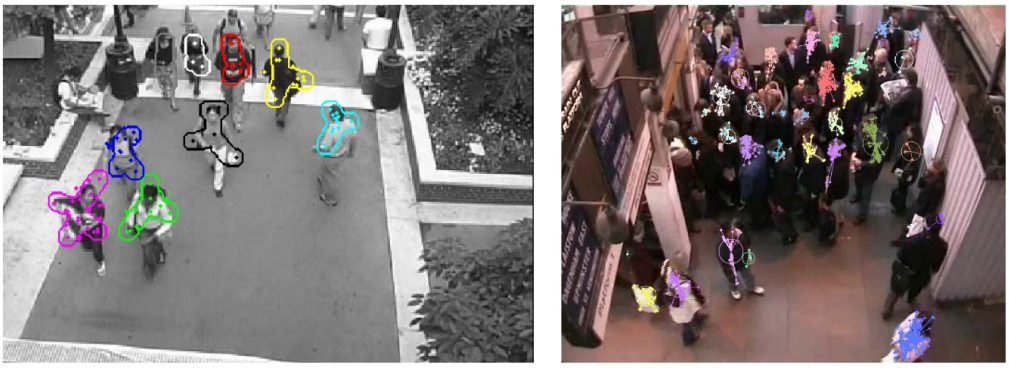


Figura 3.2: Contagem por Detecção [1]: Resultados do agrupamento de movimentos coerentes dos métodos de (a) Rabaud *et al* [5] e (b) Brostow e Cipolla [39]

Os métodos de **contagem por regressão** [1] contam pessoas em multidões aprendendo um mapeamento direto das caracaterísticas de baixo-nível da imagem para a densidade de multidões. Estima-se a densidade de multidão baseada na descrição coletiva de padrões de multidões e evita-se a segregação de indivíduos ou trajetórias como nos métodos anteriores. Assim, a contagem por regressão pode ser empregada em ambientes muito cheios em que a detecção e o rastreamento são impraticáveis.

Como exemplo, Davies *et al* [40] foram um dos primeiros a utilizar a contagem por regressão. Extraem-se as características de baixo-nível como os pixels referentes ao foreground (característica de segmento) e as características de borda de cada frame (característica de estrutura) e dessas informações se derivam características gerais como área de foreground e área total de borda. Após isso, utiliza-se um modelo de regressão linear para estabelecer um mapeamento entre esses padrões globais e o real número de pessoas.

Dessa forma, pode-se observar um pipeline de regressão na contagem por regressão, composto principalmente de representação de características, correção geométrica e modelo de regressão [1].

A representação de características [1] diz respeito à extração, seleção e transformação de propriedades visuais de baixo nível em uma imagem a fim de construir uma entrada intermediária para um modelo de regressão. Podem ser do tipo características de segmento foreground (como área ou perímetro), características de bordas (como área da borda e orientação) ou características de textura e gradientes (como Matriz de Ocorrência de Níveis de Cinza [41] ou características HOG [42]).

A correção geométrica ou normalização de perspectiva [1] tenta transformar o tamanho percebido dos objetos em diferentes profundidades para a mesma escala, tentando resolver o problema de distorção de perspectiva, como quando objetos afastados da câmera parecem menores do que os objetos próximos.

Após a extração de características e da correção geométrica, um modelo de regressão [1] é treinado para estimar o número de pessoas dadas as características normalizadas. Há várias funções diferentes utilizadas para o modelo de regressão. Entre elas estão a Regressão Linear, a Regressão dos Mínimos Quadrados Parciais (do termo em inglês *Partial least squares regression*), a *Kernel ridge regression*, a Regressão de Vetores de Suporte (do termo em inglês *Support vector regression*), a Regressão de Processos Gaussianos (do termo em inglês *Gaussian processes regression*) e a Regressão da Floresta Aleatória (do termo em inglês *Random forest regression*) [1].

Outros vários métodos foram propostos seguindo esse modelo de pipeline, mas com diferentes conjuntos de características e modelos de regressão mais sofisticados. Por exemplo, alguns trabalhos mais recentes como Lin *et al* [43] utilizam as características de segmentos, bordas e gradientes com modelos de regressão do tipo de processos Gaussianos a um nível de segmento. Por sua vez, Ke *et al* [44] utilizam características de segmentos, bordas e texturas com modelos de regressão do tipo *kernel ridge regression* a nível de segmento também.

A Figura 3.3 apresenta um pipeline típico do modelo de regressão: primeiramente uma região de interesse é definida e o mapeamento de normalização de perspectiva da cena é encontrada. Posteriormente as características são extraídas e usadas por um modelo de regressão, que é treinado para estimar o número de pessoas.

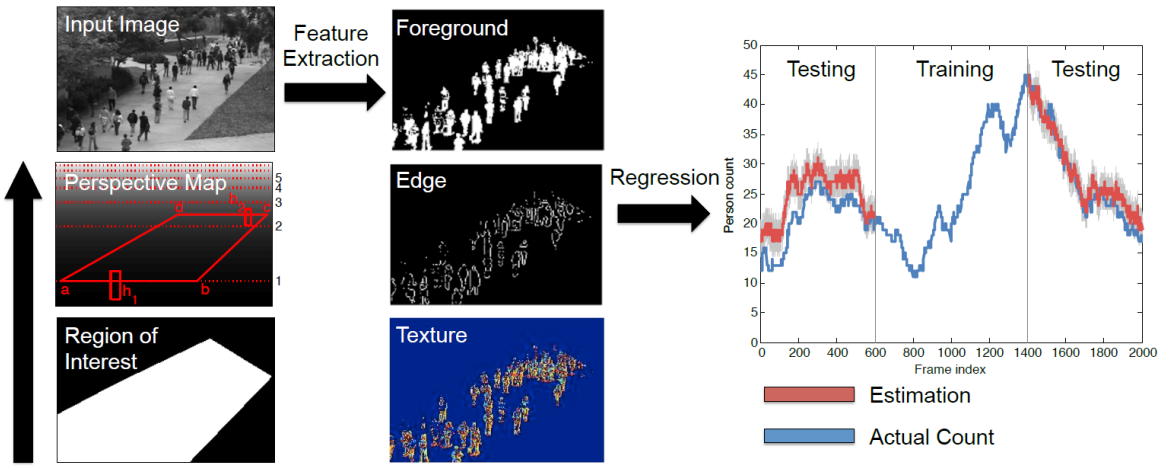


Figura 3.3: Contagem por Regressão [1]: Pipeline com representação de características, mapeamento de normalização de perspectiva e modelo de regressão

Capítulo 4

Metodologia

Esse capítulo apresenta o detalhamento dos métodos utilizados para a estimação do número de pessoas. A metodologia foi fundamentada principalmente no processo descrito no trabalho de A. S. Rao *et al* [2].

Embora o método possa ser aplicado a outros tipos de vídeo, ele foi testado em cinco vídeos de padrão half-resolution PAL, com 384 x 288 pixels, 25 frames por segundo e método de compressão MPEG2 [45]. Tais vídeos fazem parte de um banco de dados que pode ser encontrado na página CAVIAR Test Scenarios [7]. Este banco de dados é composto majoritariamente de cenas de interior, ruidosas e com câmera estática.

As implementações foram feitas no MATLAB R2015b [46], utilizando funções de Motion Estimation da Computer Vision System Toolbox [47] para o cálculo do Fluxo Óptico e também funções de Hierarchical Clustering (Cluster Analysis) da Statistics and Machine Learning Toolbox [48] para o cálculo do Agrupamento Hierárquico.

Um fluxograma resumindo os passos é apresentado na Figura 4.1. Há cinco passos principais. Primeiramente, há o Pré-processamento, no qual os frames do vídeo são preparados para passarem pela Estimação do Movimento; na Estimação do movimento ocorre o cálculo do Fluxo Óptico de Lukas-Kanade [6], gerando os vetores velocidade dos objetos; na Filtragem, ocorrem a Filtragem Espacial para a blocagem dos frames e a Filtragem Temporal para computar o máximo valor de fluxo óptico dentro de certo tempo para cada bloco; no Agrupamento Hierárquico há a clusterização das velocidades a fim de agrupar as velocidades que correspondem a um mesmo objeto em um mesmo grupo; e por fim, na Estimação do Número de Pessoas, há a contagem de clusters distintos, e se mapeia o número de pessoas pelo número de clusters distintos.

Figura 4.1: Diagrama com os passos para estimação do número de pessoas

4.1 Pré-processamento

Esta seção detalha os passos do Pré-processamento. O objetivo é preparar os frames do vídeo para a Estimação do Movimento e cálculo do fluxo óptico. Assim, os frames são selecionados, convertidos para tons de cinza, filtrados e uma Região de Interesse é definida. Um fluxograma com as etapas do pré-processamento é apresentado na Figura 4.2.

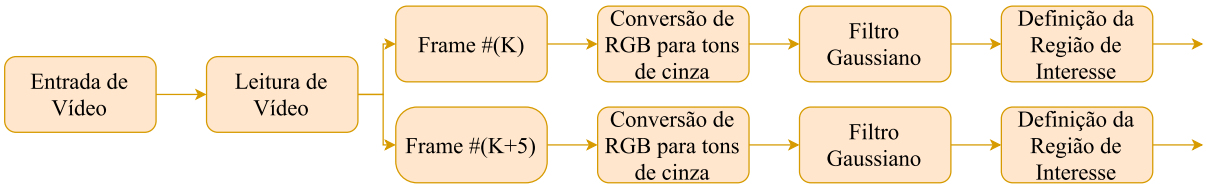


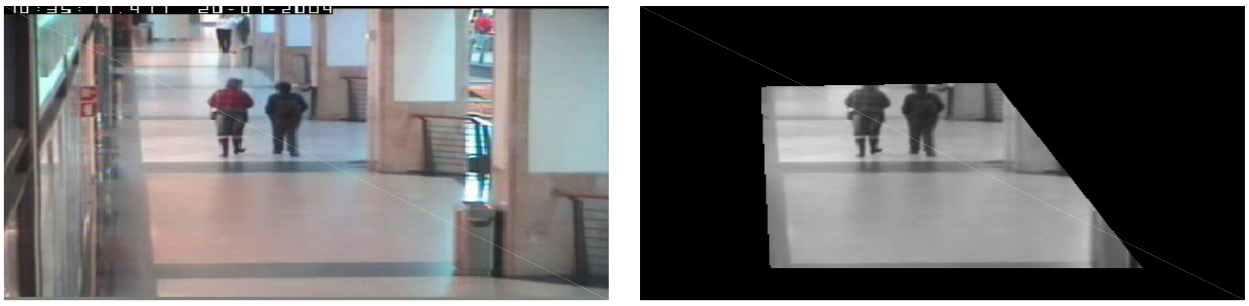
Figura 4.2: Diagrama com os passos para o pré-processamento

Inicialmente, recebe-se a entrada do vídeo, que serão os frames RGB a serem lidos. Em seguida, é selecionado apenas um frame a cada 5 frames do vídeo, desprezando os outros frames. Isso facilita o procedimento, evitando ruídos de movimento, e não gera perdas significativas, pois o frame se mantém praticamente o mesmo, já que a taxa dos vídeos utilizados é de 25 frames por segundo. Posteriormente, cada frame é convertido de RGB para escalas de cinza a fim de facilitar o cálculo do fluxo óptico.

Depois disso, filtra-se a informação de alta frequência do vídeo para que o algoritmo subsequente não a interprete como ruído e também para não prejudicar a detecção do movimento de baixa frequência. Assim, cada frame é filtrado por um filtro gaussiano 2D de 5x5 pixels e desvio padrão de $\sigma = 0,5$.

Por fim, define-se uma Região de Interesse, ou máscara, que vai delimitar a seção dos frames a ser analisada. Esse processo é feito manualmente apenas uma vez no começo de cada vídeo. Os pixels exteriores à tal região não serão utilizados no cálculo do fluxo óptico.

Na Figura 4.3(a) é apresentado um exemplo de um frame RGB antes do pré-processamento e na Figura 4.3(b) é apresentado um exemplo do frame após o pré-processamento, com a conversão para tons de cinza, o filtro gaussiano e com a Região de Interesse ou máscara já definida.



(a) (b)

Figura 4.3: Exemplo de Pré-processamento: (a) Frame RGB; (b) Frame após pré-processamento.

4.2 Estimação de Movimento

Esta seção apresenta os detalhes da Estimação de Movimento. O principal intuito é calcular o fluxo óptico para os frames, gerando os vetores velocidade dos objetos, e realizar um filtro de mediana para filtrar os ruídos. Um fluxograma com as etapas da estimação de movimento é apresentado na Figura 4.4.

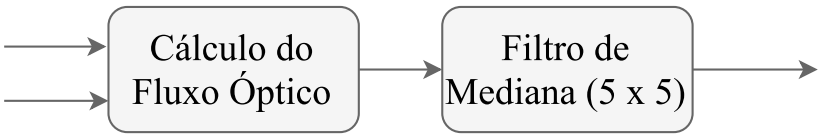


Figura 4.4: Diagrama com os passos para estimação de movimento

Depois de as imagens passarem pelo Pré-processamento, calcula-se o fluxo óptico entre dois frames utilizando método de Lucas-Kanade [6]. Notar que aqui os frames não são imediatamente consecutivos, mas na verdade têm um distância de 5 frames. Para esse método, assume-se que a velocidade de um objeto é constante ao longo dos frames.

O Fluxo Óptico é definido como o vetor velocidade dado pela Equação 4.1 [15]

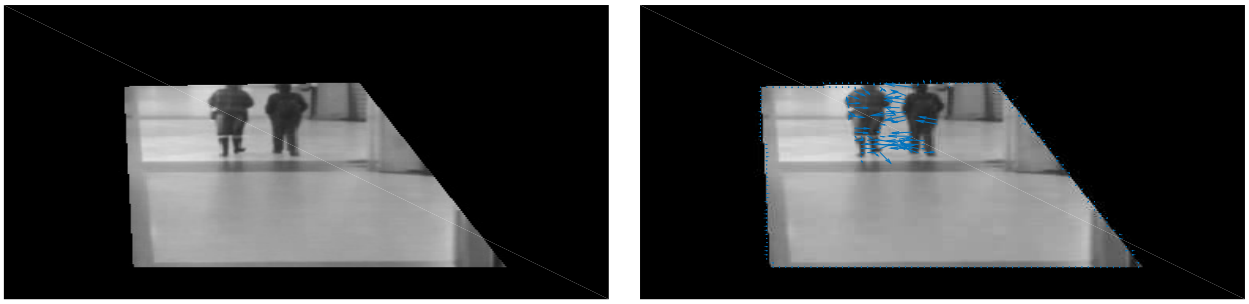
$$O := \begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} \tag{4.1}$$

em que $u, v \in \mathbb{R}$ são as componentes horizontal e vertical respectivamente do vetor velocidade para cada pixel (x, y) do frame, ou seja, é computado um fluxo óptico denso (para cada pixel).

O uso do Fluxo Óptico para a estimação do número de pessoas pode ter potenciais problemas, como quando há pessoas paradas ou quando há outros objetos se movendo, porém aqui se assume que o grupo de objetos é composto apenas de pessoas e que elas estão de fato se movimentando.

Após o cálculo do Fluxo Óptico, utiliza-se um filtro de mediana de 5x5 pixels para filtrar o ruído dos valores obtidos acima.

Na Figura 4.5 se apresenta um exemplo da Estimação do Movimento. A Figura 4.5(a) mostra o frame recebido do Pré-processamento e a Figura 4.5(b) mostra a saída da Estimação de Movimento, ou seja, o frame com os vetores velocidade calculados pelo Fluxo Óptico e filtrados pelo filtro de mediana.



(a) (b)

Figura 4.5: Exemplo de Estimação de movimento: (a) Frame vindo do Pré-processamento; (b) Frame com as velocidades obtidas pelo Fluxo Óptico e com o Filtro de Mediana.

4.3 Filtragem

Esta seção apresenta o detalhamento da etapa de Filtragem. A Filtragem é dividida em duas fases: a Filtragem Espacial, para realizar a blocagem dos frames, e a Filtragem Temporal, para computar o máximo valor de fluxo óptico dentro de certo intervalo de tempo para cada bloco. A Figura 4.6 apresenta os principais passos:

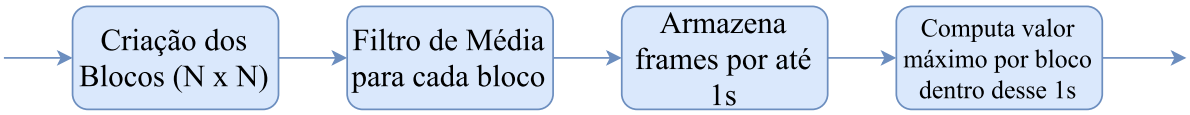


Figura 4.6: Diagrama com os passos da filtragem

4.3.1 Filtragem Espacial ou Blocagem

Primeiramente, realiza-se uma blocagem: divide-se o fluxo óptico de cada frame em blocos de $N \times N$ pixels, cujo valor será a média dos vetores velocidade pertencentes ao bloco, atribuindo-se assim um único vetor de fluxo óptico para cada bloco, com magnitude, direção e sentido. A magnitude é dada por

$$mag := \{\sqrt{u^2 + v^2}\} \tag{4.2}$$

em que $u, v \in \mathbb{R}$ são as componentes horizontal e vertical respectivamente do vetor velocidade para cada bloco (x, y) do frame.

O melhor valor de N para os vídeos testados foi de 32 pixels, que foi o valor utilizado na implementação, porém também foram testados outros valores, como 16 ou 8 pixels.

A Figura 4.7(a) mostra os vetores velocidade por blocos de 32×32 pixels, de modo que os eixos representam os índices dos blocos no frame. Já a Figura 4.7(b) mostra as magnitudes destas velocidades: quanto mais clara, maior a magnitude. Para fins de ilustração, também estão representadas as blocagens de 16×16 pixels e 8×8 pixels. Assim, pode-se comparar o mesmo frame pela ótica de blocos tamanho 32×32, 16×16 e 8×8 pixels.

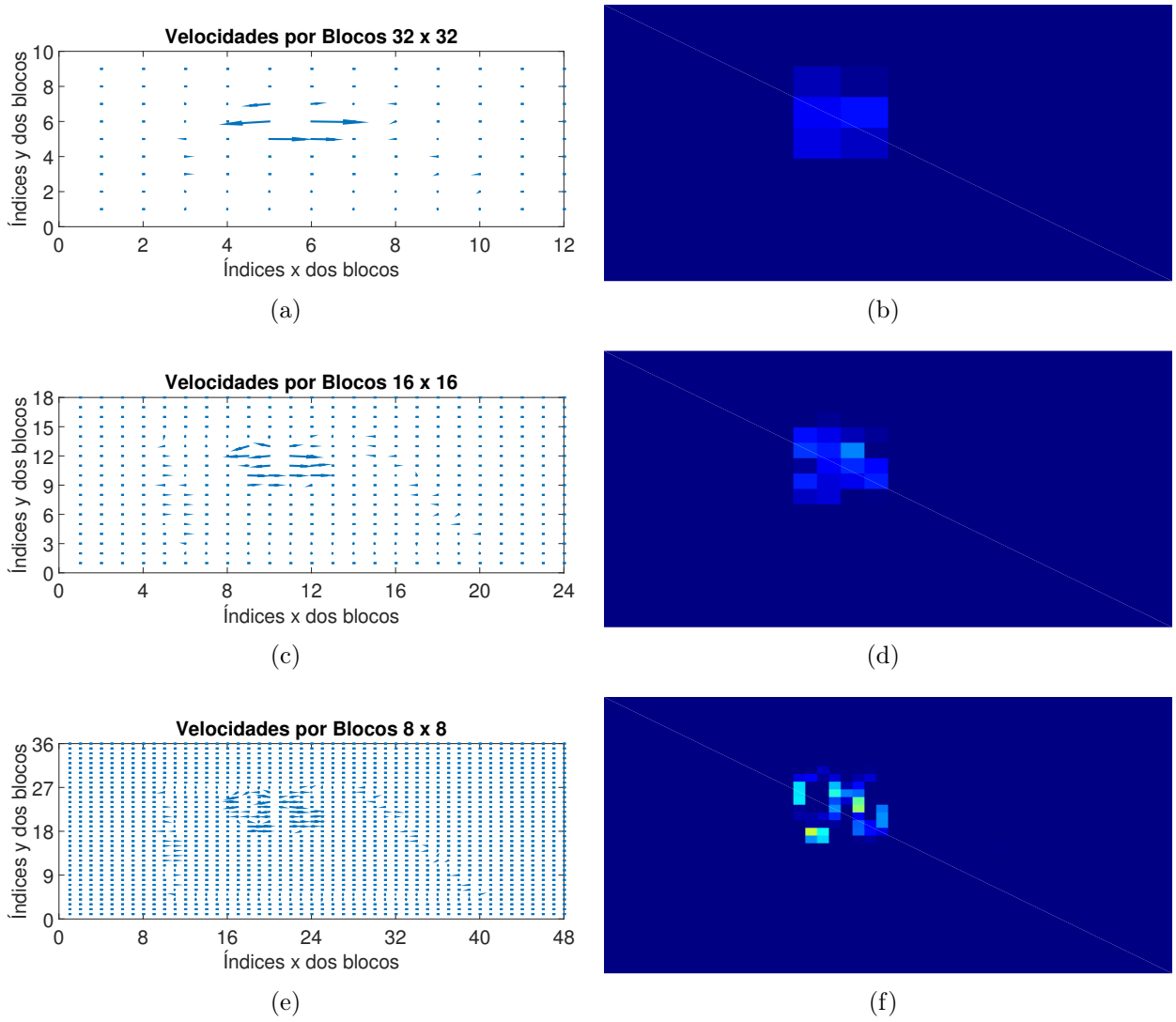


Figura 4.7: Exemplo de Filtragem Espacial: (a) Velocidades por Blocos 32 x 32; (b) Magnitude das Velocidades por Blocos 32 x 32; (c) Velocidades por Blocos 16 x 16; (d) Magnitude das Velocidades por Blocos 16 x 16; (e) Velocidades por Blocos 8 x 8; (f) Magnitude das Velocidades por Blocos 8 x 8.

4.3.2 Filtragem Temporal

A cada 5 frames que já sofreram blocagem, computa-se a velocidade de maior magnitude para para cada bloco entre esses 5 frames, ou seja, o máximo valor de velocidade ao longo do tempo (1 segundo) é atribuído ao bloco, que permanece estático. O tempo é 1 segundo porque a taxa dos vídeos é de 25 frames por segundo, mas para o cálculo de fluxo óptico só se considera 1 frame a cada 5, como explicado na Seção 4.1, resultando em 5 frames por segundo, e a filtragem temporal considera grupos de 5 frames para gerar 1 frame com o maior valor de velocidade para cada bloco. Dessa forma, a cada 1 segundo é computado 1 frame com blocos $N \times N$ de modo que cada bloco contém o maior valor de fluxo óptico encontrado dentro desse tempo nesse bloco.

A Figura 4.8 ilustra um exemplo de filtragem temporal. Os gráficos 4.8(a) a 4.8(e) mostram os blocos com os vetores velocidade para 5 frames e o gráfico 4.8(f) mostra o frame resultante com cada bloco contendo a velocidade de maior magnitude dentre os 5 frames, ou seja, após a filtragem temporal. Notar que, ao se referir ao termo **velocidades** nas próximas seções, considera-se que tais velocidades já passaram pelo processo de filtragem espacial e temporal.

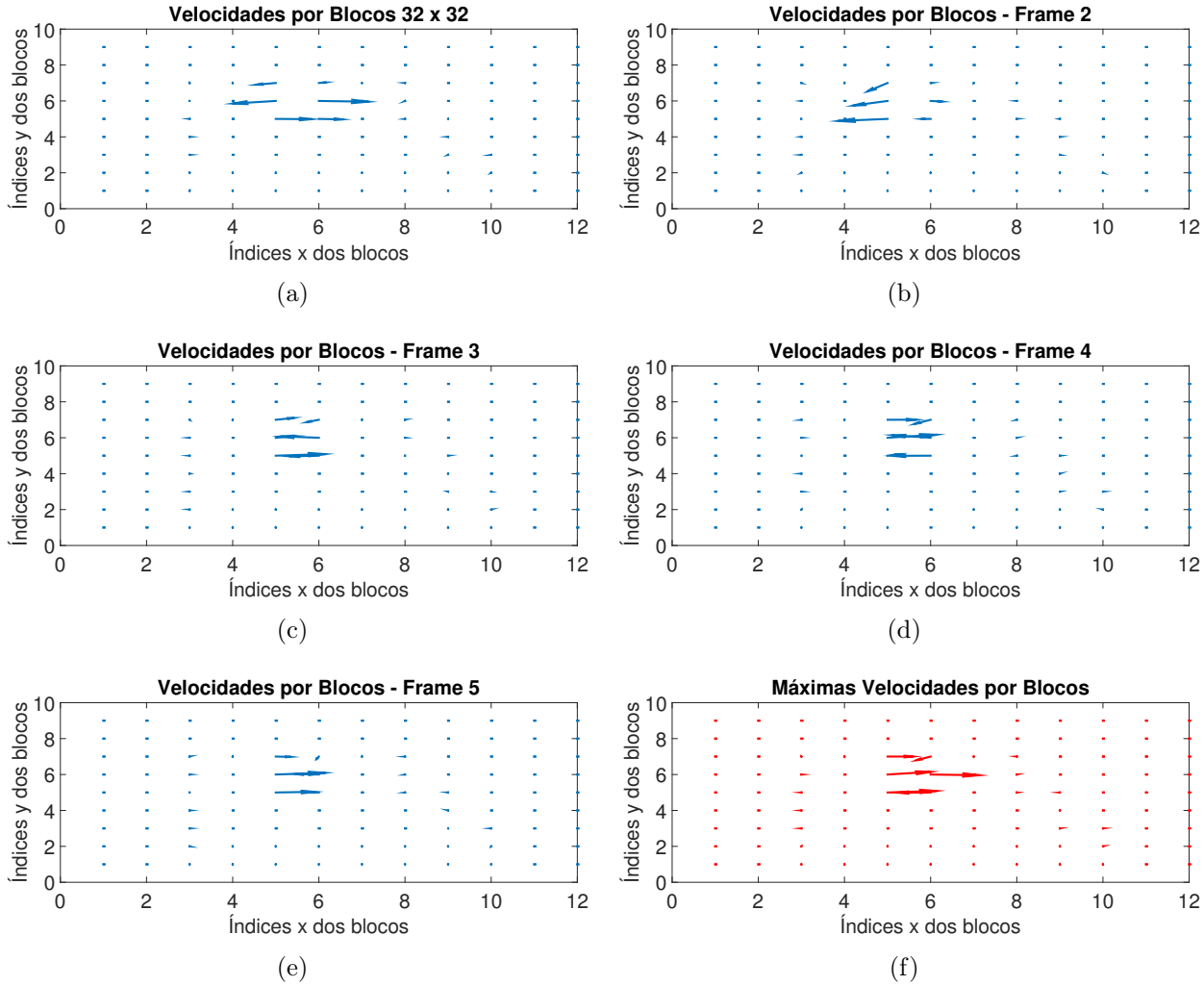


Figura 4.8: Exemplo de Filtragem Temporal: (a)-(e) Velocidades por Blocos dos frames 1 ao 5 (f) frame resultante com cada bloco contendo a maior velocidade dentre os 5 frames

Essa fase tem como resultado um mapa de atividade espaço-temporal, o que é muito importante, porque se apenas o fluxo óptico instantâneo entre dois frames fosse considerado, a informação de movimento no tempo seria descontínua devido à oclusão e aos movimentos do objeto. Portanto, essa etapa age como uma janela deslizante ao longo do eixo do tempo [2] e tenta contornar os problemas de oclusão.

4.4 Agrupamento Hierárquico

Esta seção apresenta os passos do Agrupamento Hierárquico e tem por objetivo principal realizar a clusterização das velocidades a fim de agrupar as velocidades que correspondem a um mesmo objeto em um mesmo grupo.

Esta etapa tem como entrada os frames com informação espaço-temporal vindo da etapa anterior. Assim, analisando os blocos dentro da região de interesse, podem-se perceber atividades de movimento pertencentes aos objetos com alguns blocos indicando altas atividades e outros blocos indicando baixa atividade. De acordo com o trabalho de A. S. Rao *et al* [2], o centro dos objetos possui valores de velocidades de pico e conforme se distancia do centro, estes valores decrescem. Analogamente, se há vários objetos na região de interesse, há vários picos separados por valores baixos.

Dessa forma, é utilizado um agrupamento hierárquico para agrupar tais valores de velocidade, porque aproximadamente cada pico corresponde a um único objeto. Para isso, é necessário visualizar cada velocidade como um ponto, a fim de realizar o agrupamento como descrito no Capítulo 2. Por exemplo, se existe uma velocidade $\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ em um bloco qualquer, ou seja, uma velocidade de componente horizontal u igual a 2 e componente vertical v igual a 3, deve-se entender essa velocidade como um ponto localizado em $(2, 3)$, a fim de que se possa agrupá-la com outras velocidades que possuam componentes similares. Tal agrupamento possui três fases principais:

1. encontrar a distância euclidiana entre os valores das velocidades, representadas como pontos, como descrito acima;
2. agrupar as velocidades (duas a duas) com distâncias similares, por meio de *Single-linkage Clustering* (ou Método Mínimo);
3. continuar até que todas as velocidades estejam agrupadas, criando uma árvore binária hierárquica de agrupamento chamada dendrograma.

Um fluxograma com os passos do Agrupamento Hierárquico é apresentado na Figura 4.9:

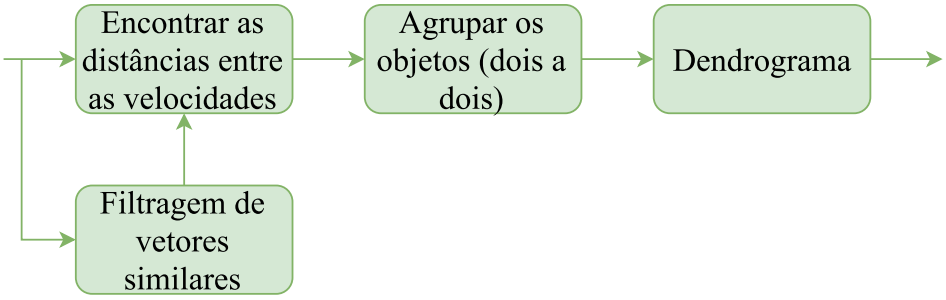
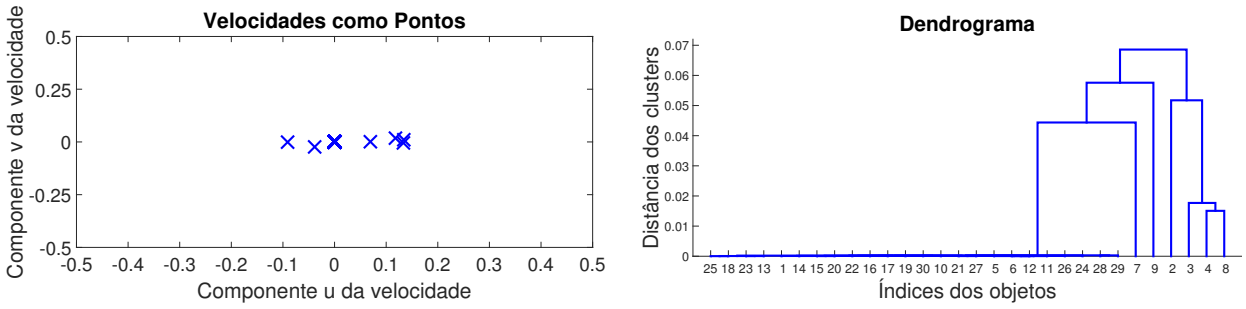


Figura 4.9: Diagrama com os passos do Agrupamento Hierárquico

A Figura 4.10 ilustra as velocidades como pontos, a fim de que possa se calcular a distância entre elas, e o dendrograma formado pelo agrupamento dois a dois desses pontos. O bloco de Filtragem de Vetores Similares é explicado na Subseção 4.4.1.



(a) (b)

Figura 4.10: Exemplo de Agrupamento Hierárquico: (a) Velocidades como Pontos; (b) Dendrograma.

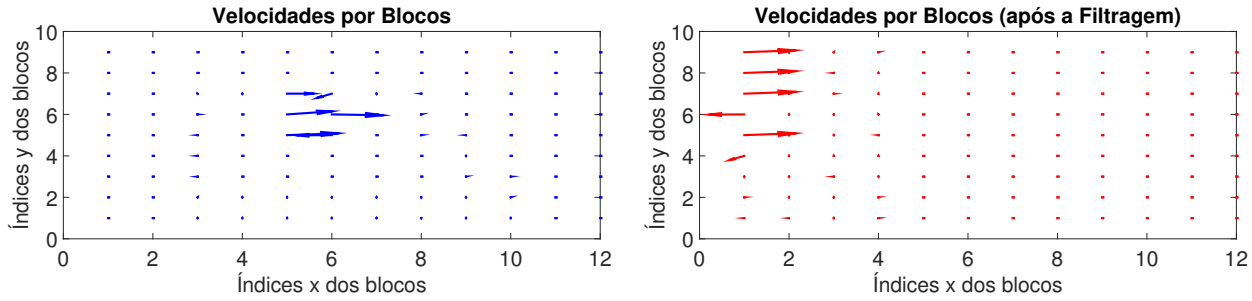
4.4.1 Filtragem de vetores similares antes do agrupamento

Esta fase não estava presente no artigo de A. S. Rao *et al* [2], mas foi proposta com a tentativa de amenizar um problema que houve durante os testes. Foi observado que em alguns casos alguns vetores similares estavam resultando em contagem de diferentes objetos. Assim, propôs-se esta filtragem de vetores similares antes do Agrupamento Hierárquico a fim de tentar contornar essa situação.

Dessa forma, verificam-se quais vetores podem ser considerados similares e então se igualam os seus valores, para que na contagem após o agrupamento hierárquico eles sejam contados como o mesmo objeto. Aqui foram considerados similares os vetores que possuíam até 30 graus de diferença na fase e até determinada diferença de porcentagem na magnitude. Esses valores foram determinados empiricamente. Aqui não é considerada a posição dos blocos em que as velocidades estão, mas sim a própria natureza das velocidades, como magnitude e fase. Depois dessa "filtragem", repete-se o Agrupamento Hierárquico para esse novo conjunto de vetores, e este resultado será passado também para a próxima fase, a Estimação do Número de Pessoas. Dessa forma, haverá dois resultados: um resultado sem a filtragem dos vetores similares e outro resultado com tal filtragem.

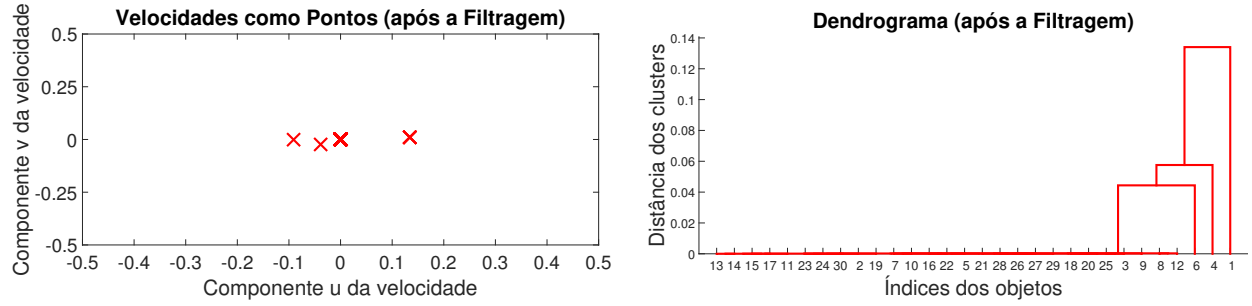
A Figura 4.11 mostra um exemplo dos vetores velocidades antes e depois dessa filtragem proposta. A Figura 4.11(a) mostra as velocidades antes da filtragem e a sua localização reflete a real posição do vetor no frame. A Figura 4.11(b) mostra as velocidades após a filtragem e a sua localização não representa a real posição do vetor no frame, porque tais vetores foram reordenados durante o processo da filtragem. Essa reordenação não terá nenhum impacto nos passos posteriores.

A Figura 4.12 ilustra as velocidades como pontos e o dendrograma após a filtragem dos vetores similares.



(a) (b)

Figura 4.11: Exemplo dos Vetores Velocidade antes e depois da filtragem de vetores similares: (a) Vetores Velocidade antes da Filtragem; (b) Vetores Velocidade após a Filtragem.



(a) (b)

Figura 4.12: Exemplo de Agrupamento Hierárquico após filtragem de vetores similares: (a) Velocidades como Pontos após a filtragem; (b) Dendrograma após a filtragem

4.5 Estimação do Número de Pessoas

Esta seção apresenta os passos para a Estimação do Número de Pessoas e tem por finalidade contar o número de clusters distintos vindos do Agrupamento Hierárquico e estimar o número de pessoas.

Como os picos distintos correspondem a objetos distintos, mapeia-se o o número de pessoas em uma determinada área pelo número de agrupamentos distintos obtidos do Agrupamento Hierárquico.

Faz-se um gráfico $dist \times dist$, em que dist é a distância entre os agrupamentos do dendrograma, para determinar os diferentes objetos. Se a maior distância for ínfima, é porque há apenas ruídos nesse conjunto de frames e o número de pessoas é considerado como zero. Se as distâncias forem consideráveis, ainda assim haverá um pouco de ruído. Então, primeiramente, desprezam-se valores muito pequenos, causados por ruídos, menores que determinada porcentagem da maior distância. Tal porcentagem é passada por parâmetro, sendo igual a 15% para o Resultado com Filtragem e 35% para o caso sem Filtragem de vetores similares. Em seguida, deve-se contar o número de clusters distintos. A princípio, cada ponto desse gráfico representa um objeto. Porém, foi observado que pontos muito próximos representam o mesmo objeto. Assim, verifica-se o quão próximas essas distâncias estão, para definir se podem ou não ser consideradas como o mesmo objeto (sendo considerados como mesmo objeto quando a distância entre eles for até 10%

da maior distância), e conta-se tal número de clusters distintos, resultando na contagem dos objetos para esse conjunto de frames. Com isso, o número de objetos nessa área de interesse pode ser determinado utilizando informação espacial e temporal.

Um fluxograma com os passos da Estimação do Número de Pessoas é apresentado na Figura 4.13:

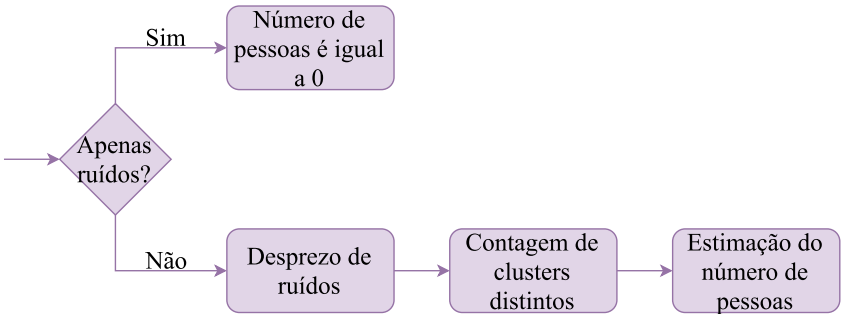
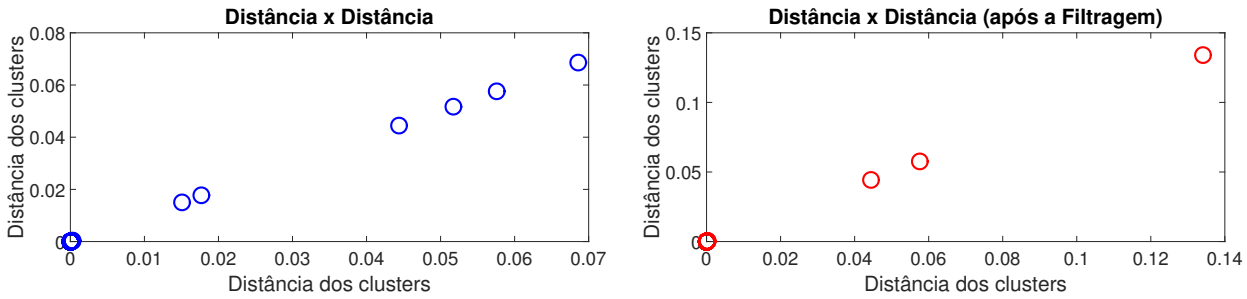


Figura 4.13: Diagrama com os passos da estimação do número de pessoas

A seguir, apresentam-se dois exemplos de gráfico $dist \times dist$ na Figura 4.14. O primeiro deles vem do dendrograma sem a filtragem de vetores similares. Foram contados 3 clusters distintos nesse caso e havia 2 pessoas no conjunto de frames considerado. O segundo vem do dendrograma após a filtragem de vetores similares. Foram contados 2 clusters distintos e havia 2 pessoas. A filtragem de vetores similares ajuda em alguns casos, porém em outros não, como é apresentado no capítulo de Resultados.



(a) (b)

Figura 4.14: Exemplo de Estimação do número de pessoas: (a) Gráfico $dist \times dist$ sem filtragem de vetores similares. Estimação: 3, Realidade: 2 ; (b) Gráfico $dist \times dist$ após filtragem de vetores similares. Estimação: 2, Realidade: 2.

4.6 Exemplos

Para ilustrar melhor, são apresentados três exemplos com os principais passos do algoritmo (notar que foram utilizados blocos de 32×32 pixels na implementação):

- um exemplo de caso ótimo, em que ambos os resultados, com e sem filtragem de vetores similares, acertaram em 100%;

- um exemplo de caso médio, em que o resultado após a filtragem de vetores similares acertou 100% e em que houve taxa de erro no resultado sem tal filtragem;
- um exemplo de caso ruim, em que houve uma taxa de erro em ambos os resultados;

4.6.1 Exemplo de Caso Ótimo

A Figura 4.15 representa um exemplo de caso ótimo, em cujo conjunto de frames há 2 pessoas andando. O resultado da estimação sem a filtragem foi de 2 pessoas e o resultado para a estimação com a filtragem de vetores similares foi de 2 pessoas.

4.6.2 Exemplo de Caso Médio

A Figura 4.16 representa um exemplo de caso médio, em cujo conjunto de frames há 2 pessoas andando. O resultado da estimação sem a filtragem foi de 3 pessoas e o resultado para a estimação com a filtragem de vetores similares foi de 2 pessoas.

4.6.3 Exemplo de Caso Ruim

A Figura 4.17 representa um exemplo de caso ruim, em cujo conjunto de frames há 3 pessoas andando. O resultado da estimação sem a filtragem foi de 4 pessoas e o resultado para a estimação com a filtragem de vetores similares foi de 2 pessoas.

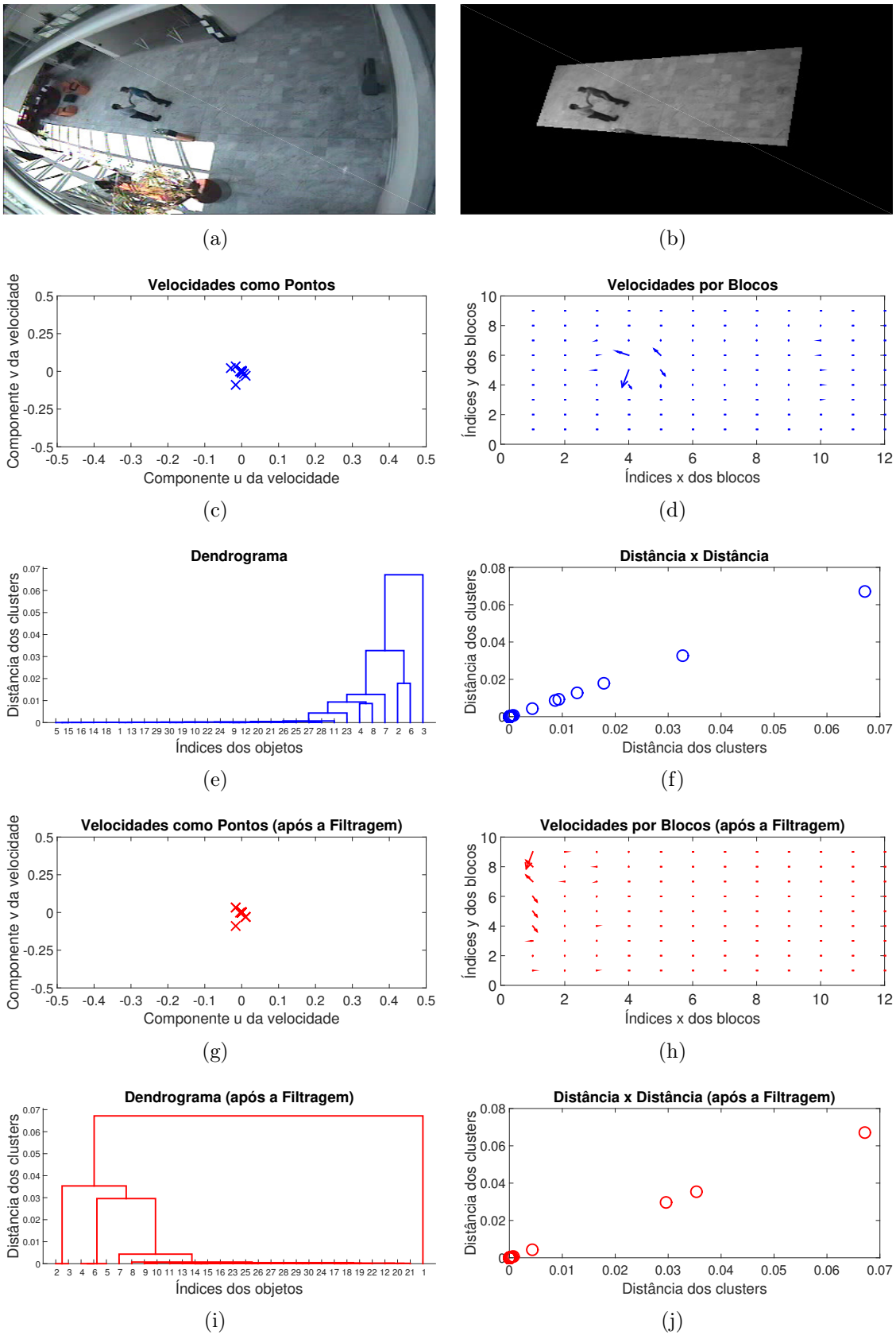


Figura 4.15: Exemplo de Caso Ótimo: (a) Frame RGB; (b) Frame após pré processamento; (c) Velocidades do fluxo óptico após filtragem espaço-temporal; (d) Velocidades de c por blocos; (e) dendrograma; (f) distâncias; (g)-(j) mesmo que (c)-(f) mas com a aplicação da filtragem dos vetores similares

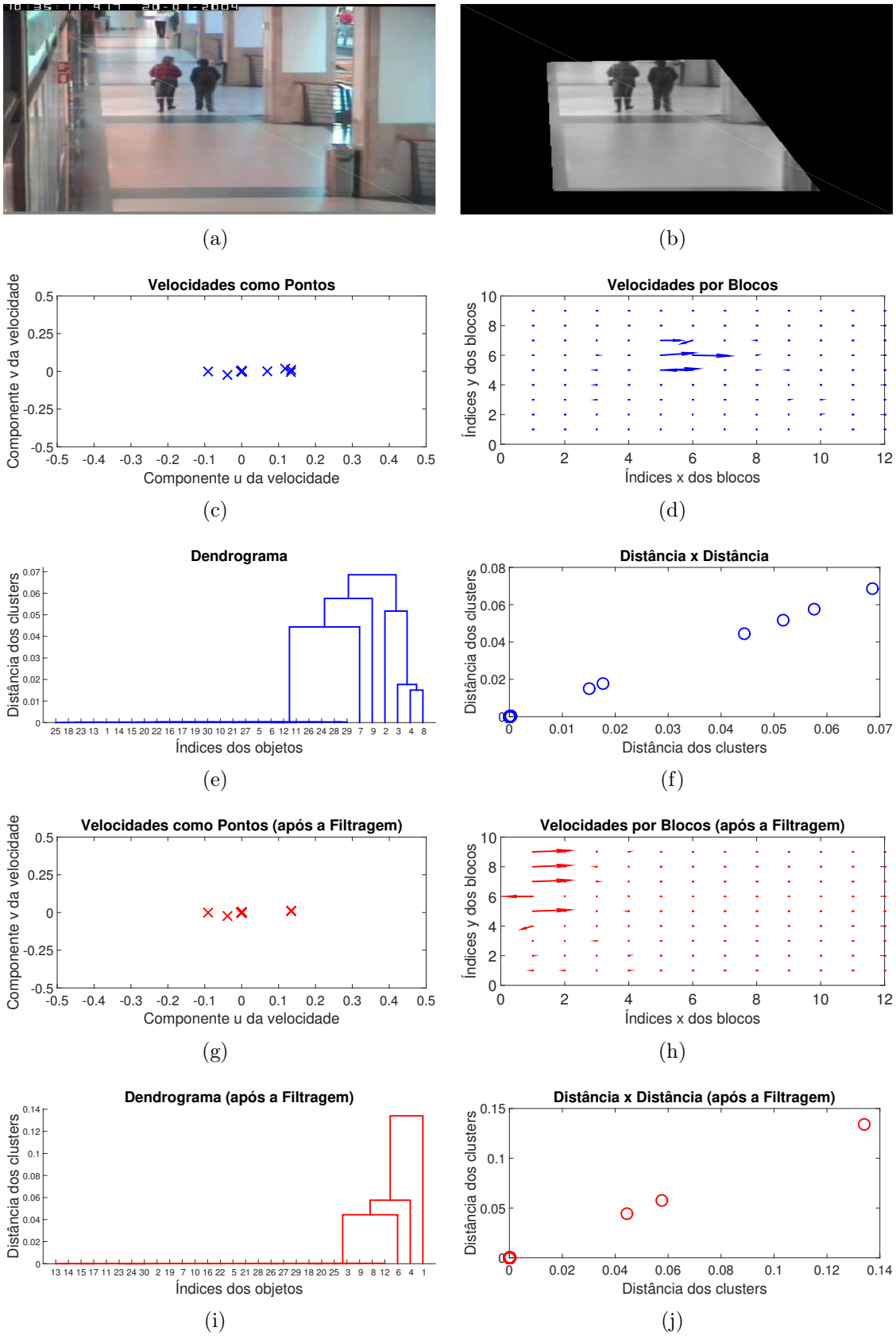


Figura 4.16: Exemplo de Caso Médio: (a) Frame RGB; (b) Frame após pré processamento; (c) Velocidades do fluxo óptico após filtragem espaço-temporal; (d) Velocidades de c por blocos; (e) dendrograma; (f) distâncias; (g)-(j) mesmo que (c)-(f) mas com a aplicação da filtragem dos vetores similares

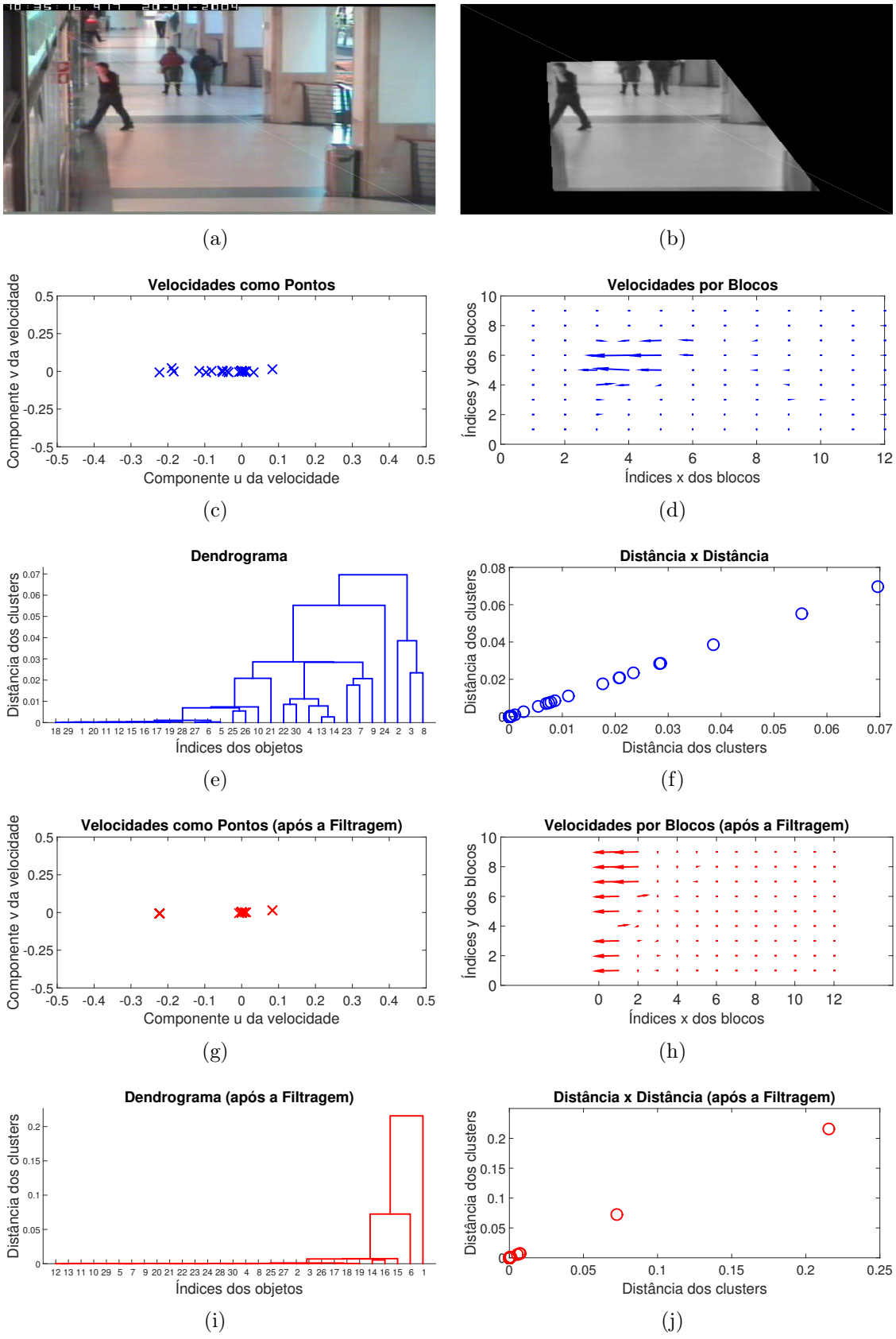


Figura 4.17: Exemplo de Caso Ruim : (a) Frame RGB; (b) Frame após pré processamento; (c) Velocidades do fluxo óptico após filtragem espaço-temporal; (d) Velocidades de c por blocos; (e) dendrograma; (f) distâncias; (g)-(j) mesmo que (c)-(f) mas com a aplicação da filtragem dos vetores similares

Capítulo 5

Resultados

Este capítulo tem por intuito expor os Resultados obtidos nos casos com e sem Filtragem de Vetores Similares para cada vídeo e também apresentar algumas métricas de avaliação, tais como Média do Número de Pessoas Estimadas, Média do Número de Pessoas do Valor de Referência, Média de Acurácia e Média de Erro Absoluto.

O algoritmo foi testado em 5 vídeos diferentes. Tais vídeos fazem parte do Projeto CAVIAR e podem ser encontrados na página CAVIAR Test Scenarios [7]. Possuem padrão half-resolution PAL, com 384 x 288 pixels, 25 frames por segundo e método de compressão MPEG2 [45]. Quatro vídeos foram filmados em um corredor de um shopping em Lisboa e um vídeo foi filmado em um lobby de entrada do INRIA Labs em Grenoble, na França. Todos estão disponíveis publicamente e podem ser baixados no formato MPEG2 [45] ou separados em JPEGS [49].

As implementações foram feitas no MATLAB 2015b [46] e testadas em um computador com Windows 10, 64 bits, equipado com processador Intel(R) Core(TM) i7-4500U CPU em 1.8GHz e placa de Vídeo Intel(R) HD Graphics 4400.

Para cada vídeo é apresentado um gráfico de Resultados. Nele, pode-se ver o número de pessoas que realmente estavam no vídeo - o "Valor de Referência", o número de pessoas estimadas pelo Algoritmo Sem Filtragem e o número de pessoas estimadas pelo Algoritmo com a Filtragem de Vetores Similares. Há um valor de número de pessoas a cada 5 frames devido à natureza da implementação, por causa da Filtragem Temporal que ocorre a cada 5 frames. Assim, o valor para o frame 5 representa o valor calculado para os frames 1 a 5, o valor para o frame 10 representa o valor calculado para os frames 6 a 10 e assim por diante.

Além disso, para cada vídeo são apresentadas quatro tipos de métricas, a saber:

- Média do Número de Pessoas Estimadas: média simples do número de pessoas estimadas, dada pela equação 5.1

$$m_{\mathrm{NPE}} = \frac{1}{N}\sum_{i=1}^{N} npe_i \tag{5.1}$$

  em que N é o número de amostras ou o número total de conjuntos de 5 frames.

- Média do Número de Pessoas do Valor de Referência: média simples do número de pessoas do Valor de Referência, dada pela equação 5.2

$$m_{\text{NPVR}} = \frac{1}{N}\sum_{i=1}^{N} npvr_i \tag{5.2}$$

em que N é o número de amostras ou o número total de conjuntos de 5 frames.

- Média de Acurácia: média simples das acurácias, dada pela equação 5.3

$$m_{\text{A}} = \frac{1}{N}\sum_{i=1}^{N} a_i \tag{5.3}$$

em que N é o número de amostras ou o número total de conjuntos de 5 frames e em que cada acurácia a [50] [51] é dada por:

$$a := \frac{VP + VN}{VP + VN + FP + FN} \tag{5.4}$$

em que

  - VP significa Verdadeiro Positivo: o elemento de entrada é Genuíno (Positivo) e o algoritmo o classifica como Positivo, ou seja, há uma pessoa e o algoritmo a contabilizou corretamente;
  - VN significa Verdadeiro Negativo: o elemento de entrada é Impostor (Negativo) e o algoritmo o classifica como Negativo. Não há equivalência direta com o nosso problema, então seu valor será sempre 0;
  - FP significa Falso Positivo: o elemento de entrada é Impostor (Negativo) e o algoritmo o classifica como Positivo, ou seja, uma pessoa foi contabilizada erroneamente, seja por movimentos dos membros ou falha nas filtragens;
  - FN significa Falso Negativo: o elemento de entrada é Genuíno (Positivo) e o algoritmo o classifica como Negativo, ou seja, uma pessoa deveria ter sido contabilizada, mas não foi.

- Média de Erro Absoluto: média simples dos erros absolutos, dada pela equação 5.5

$$m_{\text{EA}} = \frac{1}{N}\sum_{i=1}^{N} ea_i \tag{5.5}$$

em que cada erro absoluto ea é dado por:

$$ea := |ValordeReferencia - ValorEstimado| = FP + FN \tag{5.6}$$

e em que N é o número de amostras ou o número total de conjuntos de 5 frames.

Esta métrica é importante pois representa quantas pessoas o algoritmo erra em média por frame.

As Seções a seguir apresentam os Resultados para cada vídeo.

5.1 Vídeo 1

Este primeiro vídeo foi filmado de uma vista frontal do corredor de um shopping em Lisboa. Há duas pessoas caminhando e uma pessoa entra na loja e outra sai, de modo que o número de pessoas na Região de Interesse varia entre 2 e 3. A Figura 5.1 apresenta um frame retirado do Vídeo 1 como exemplo, enquanto a Figura 5.2 apresenta os Resultados e a Tabela 5.1 apresenta as principais métricas para o Vídeo 1.

Figura 5.1: Frame do Vídeo 1

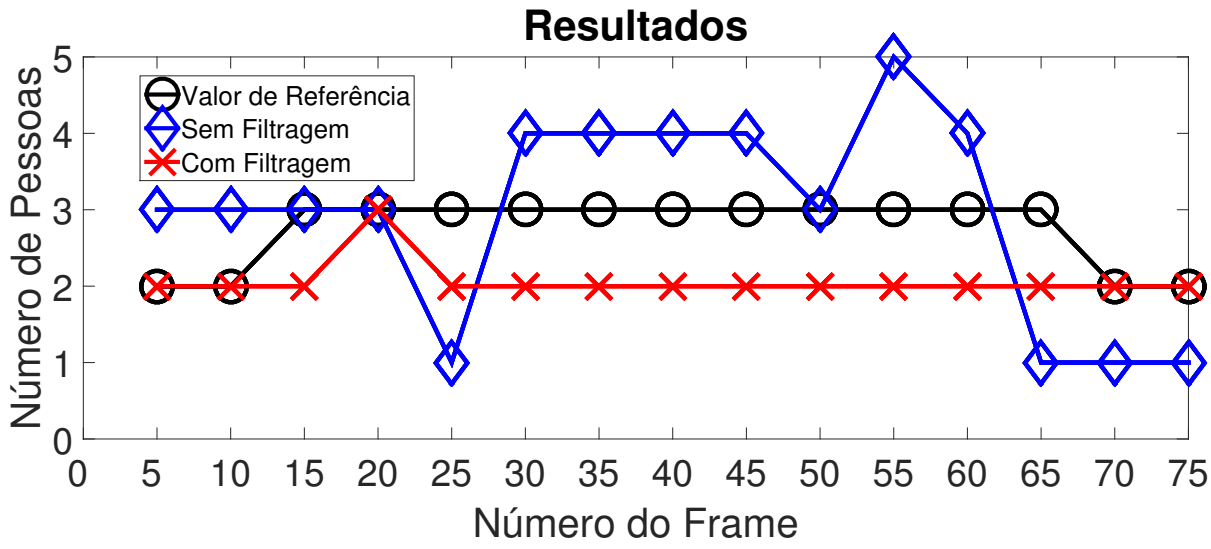


Figura 5.2: Resultados para Vídeo 1

Para este caso, o Resultado com a Filtragem foi melhor que o Resultado sem Filtragem, com menor média de Erro Absoluto e melhor média de acurácia. Porém, a Média do número de Pessoas Estimadas do Resultado Sem Filtragem foi mais próxima da média do Valor de Refêrencia.

No caso Sem Filtragem, alguns valores foram maiores que os de referência devido ao fato de que os membros das pessoas se movem com velocidades diferentes da velocidade do centro do corpo (como os braços) e isso algumas vezes é contado como outro objeto, gerando Falsos Positivos.

Médias	Sem Filtragem	Com Filtragem
Média do nº de Pessoas Estimadas	2,93	2,07
. Média do nº de Pessoas do Valor de Referência	2,73	2,73
Média de Erro Absoluto do nº de Pessoas	1	0,67
Média de Acurácias	69,0%	77,8%

Tabela 5.1: Métricas para Vídeo 1

No caso Com a Filtragem, após esses vetores similares terem sido igualados e não mais contados como objetos diferentes, temos o problema de que as duas pessoas estão andando juntas e aproximadamente com a mesma velocidade, sendo contadas como o mesmo objeto, gerando Falsos Negativos. Assim, quando há três pessoas - uma separada e duas juntas, elas são contadas apenas como duas pessoas.

5.2 Vídeo 2

O segundo vídeo foi filmado de uma vista lateral do corredor do shopping em Lisboa, de modo que é possível ver uma área dentro da loja. Há uma pessoa andando dentro da loja, depois vai para o corredor e volta, enquanto nos frames finais outra pessoa caminha no corredor. O número de pessoas na Região de Interesse varia entre 1 e 2. A Figura 5.3 apresenta um frame retirado do Vídeo 2 como exemplo, enquanto a Figura 5.4 apresenta os Resultados e a Tabela 5.2 apresenta as principais métricas para o Vídeo 2.

Figura 5.3: Frame do Vídeo 2

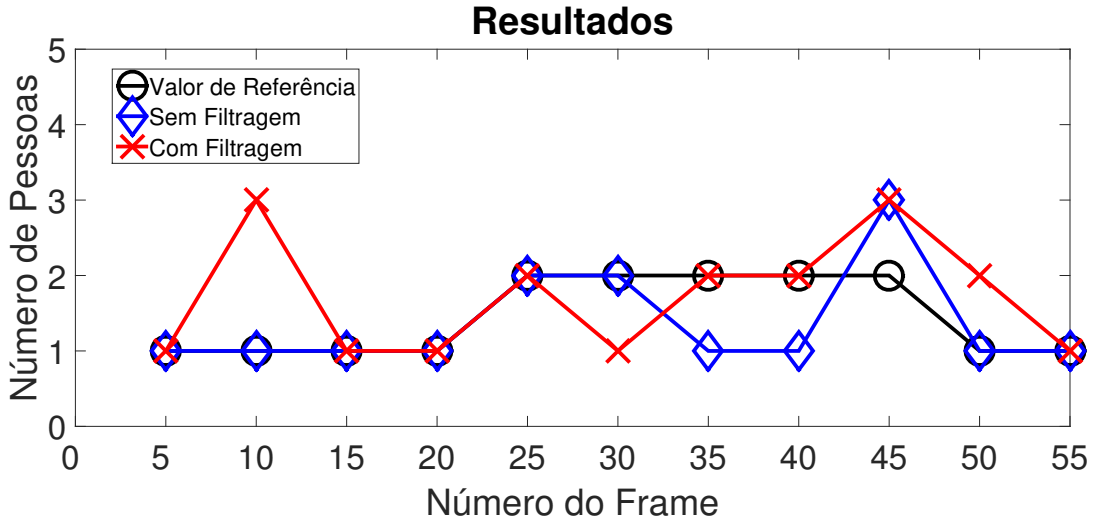


Figura 5.4: Resultados para Vídeo 2

Para este caso, os Resultados Com e Sem Filtragem foram muito bons, com a Média do Número de Pessoas Estimadas bem próximas à Média do Valor de Referência, baixo Erro Absoluto e boa Acurácia. O Resultado Sem a Filtragem foi um pouco melhor que o Resultado Com a Filtragem.

Esse vídeo tem uma peculiaridade por ter ser filmado da vista lateral do corredor, e as pessoas ficaram mais distantes da câmera do que nos casos de vista frontal, gerando

Médias	Sem Filtragem	Com Filtragem
Média do nº de Pessoas Estimadas	1,36	1,73
. Média do nº de Pessoas do Valor de Referência	1,45	1,45
Média de Erro Absoluto do nº de Pessoas	0,27	0,45
Média de Acurácias	87,8%	81,8%

Tabela 5.2: Métricas para Vídeo 2

vetores velocidade de magnitude menor e gerando distâncias menores no Agrupamento Hierárquico.

Assim, o Resultado Com Filtragem teve alguns pontos com mais pessoas porque durante a Estimação do Número de Pessoas, alguns valores vindo do Agrupamento Hierárquico são considerados como ruído se estiverem abaixo de determinada porcentagem da maior distância. Como esse parâmetro da porcentagem é menor no caso Com Filtragem (porque já houve a filtragem) e a maior distância é pequena, alguns valores de ruídos não foram desprezados e foram contabilizados como pessoas, causando Falsos Positivos.

5.3 Vídeo 3

Este vídeo foi filmado em um lobby de entrada do INRIA Labs em Grenoble. Nele, duas pessoas vêm caminhando em sentidos opostos, se encontram e cumprimentam, depois seguem o caminho juntas. O número de pessoas na Região de Interesse varia entre 0 e 2. A Figura 5.5 apresenta um frame retirado do Vídeo 3 como exemplo, enquanto a Figura 5.6 apresenta os Resultados e a Tabela 5.3 apresenta as principais métricas para o Vídeo 3.

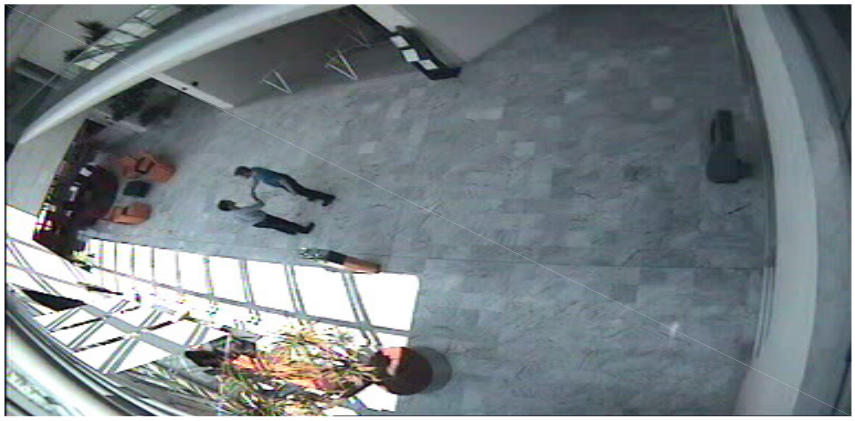


Figura 5.5: Frame do Vídeo 3

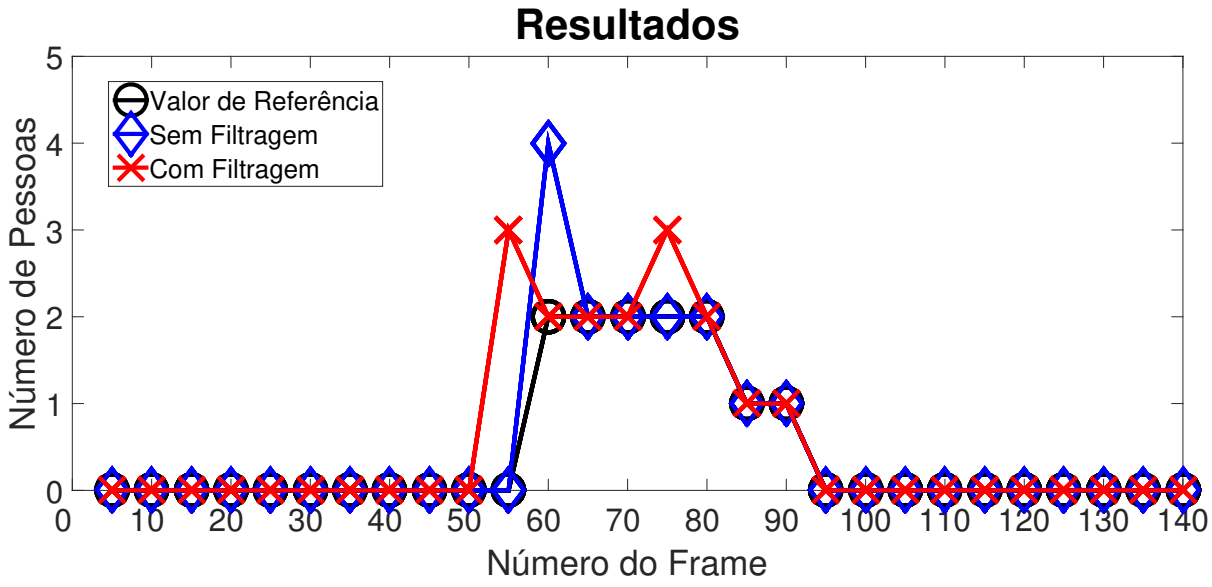


Figura 5.6: Resultados para Vídeo 3

Este caso apresentou o melhor resultado entre todos os vídeos, com a Média do Número de Pessoas Estimadas bem próximas à Média do Valor de Referência, Erro Absoluto médio muito baixo e alta média de Acurácias, tanto para o Resultado Sem Filtragem quanto para o Com Filtragem. Os pontos que apresentaram alguns picos ocorreram porque os ruídos naqueles frames não foram tão pequenos a ponto de ser desprezados, causando a contagem de Falsos Positivos.

Médias	Sem Filtragem	Com Filtragem
Média do nº de Pessoas Estimadas	0,50	0,57
. Média do nº de Pessoas do Valor de Referência	0,43	0,43
Média de Erro Absoluto do nº de Pessoas	0,07	0,14
Média de Acurácias	98,2%	95,2%

Tabela 5.3: Métricas para Vídeo 3

Este resultado também é importante porque mostra que o algoritmo, na grande maioria das vezes, não acusa que há pessoas se de fato não há. Isso ocorre porque se o passo de Agrupamento Hierárquico só gera distâncias muito pequenas (consideradas como ruídos), o passo de Estimação do Número de Pessoas já considera que não há ninguém naquele conjunto de frames.

5.4 Vídeo 4

Este vídeo foi filmado de uma vista frontal do corredor de um shopping em Lisboa e é mais longo que os vídeos anteriores. Primeiramente, há três pessoas andando perto da câmera, mas elas começam a andar em direções diferentes, olham as vitrines, saem da Região de Interesse e voltam. Posteriormente, um grupo de outras três pessoas vêm caminhando de longe e se aproximando da câmera. O número de pessoas na Região de Interesse varia entre 0 a 5. A Figura 5.7 apresenta um frame retirado do Vídeo 4 como exemplo, enquanto a Figura 5.8 apresenta os Resultados e a Tabela 5.4 apresenta as principais métricas para o Vídeo 4.

Figura 5.7: Frame do Vídeo 4

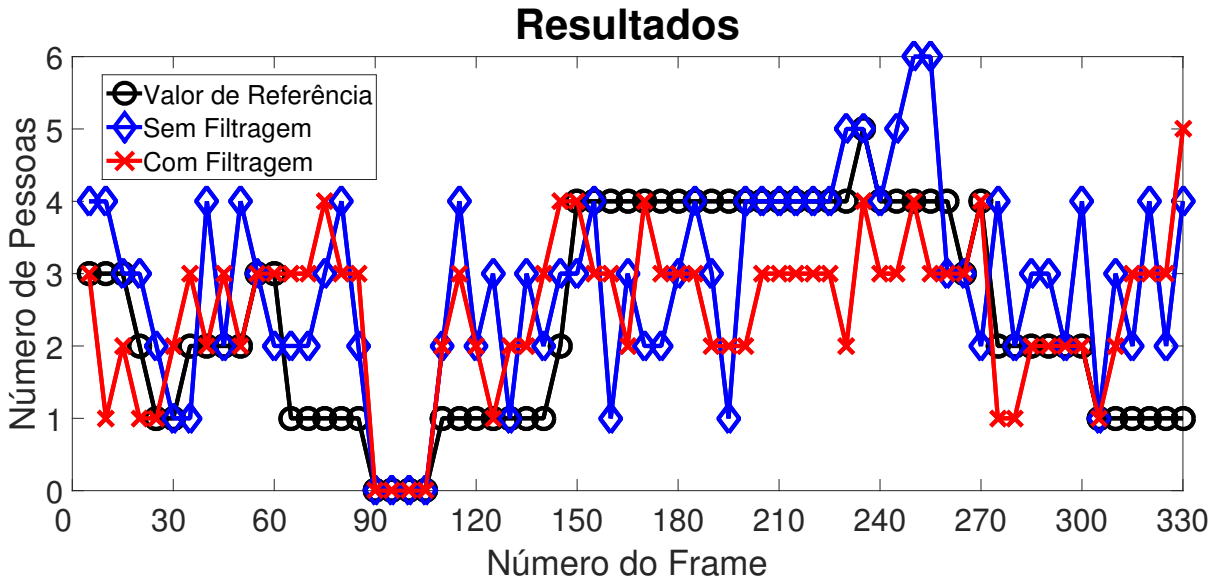


Figura 5.8: Resultados para Vídeo 4

Para este vídeo, os Resultados Com e Sem Filtragem foram praticamente equivalentes, com média de Erro Absoluto em torno de 1 pessoa de média de acurácia em torno de 69%. As médias do número de Pessoas Estimadas foram próximas à Média do Valor de Referência, mas a média do Resultado Com Filtragem foi a mais próxima.

Médias	Sem Filtragem	Com Filtragem
Média do nº de Pessoas Estimadas	2,86	2,47
. Média do nº de Pessoas do Valor de Referência	2,41	2,41
Média de Erro Absoluto do nº de Pessoas	1,03	1,00
Média de Acurácias	69,5%	69,1%

Tabela 5.4: Métricas para Vídeo 4

Alguns valores foram maiores que os de referência devido ao fato de que os membros das pessoas se movem com velocidades diferentes da velocidade do centro do corpo (como os braços) e isso algumas vezes é contado como outro objeto, gerando Falsos Positivos.

Alguns valores foram menores que os de referência, principalmente entre os frames 150 a 270, devido ao problema de que quando as três pessoas estão andando juntas e aproximadamente com a mesma velocidade, são contadas menos pessoas, porque se confundem duas delas ou mesmo as três como o mesmo objeto, gerando Falsos Negativos.

5.5 Vídeo 5

Este vídeo foi filmado de uma vista frontal do corredor de um shopping em Lisboa e é o mais longo do nosso conjunto de dados. No começo, não há ninguém na Região de Interesse, depois uma moça caminha pelo corredor se distanciando da câmera. Posteriormente, um casal vêm caminhando de longe se aproximando da câmera, enquanto algumas pessoas entram e saem das lojas. O número de pessoas na Região de Interesse varia entre 0 e 4. A Figura 5.9 apresenta um frame retirado do Vídeo 5 como exemplo, enquanto a Figura 5.10 apresenta os Resultados e a Tabela 5.5 apresenta as principais métricas para o Vídeo 5.

Figura 5.9: Frame do Vídeo 5

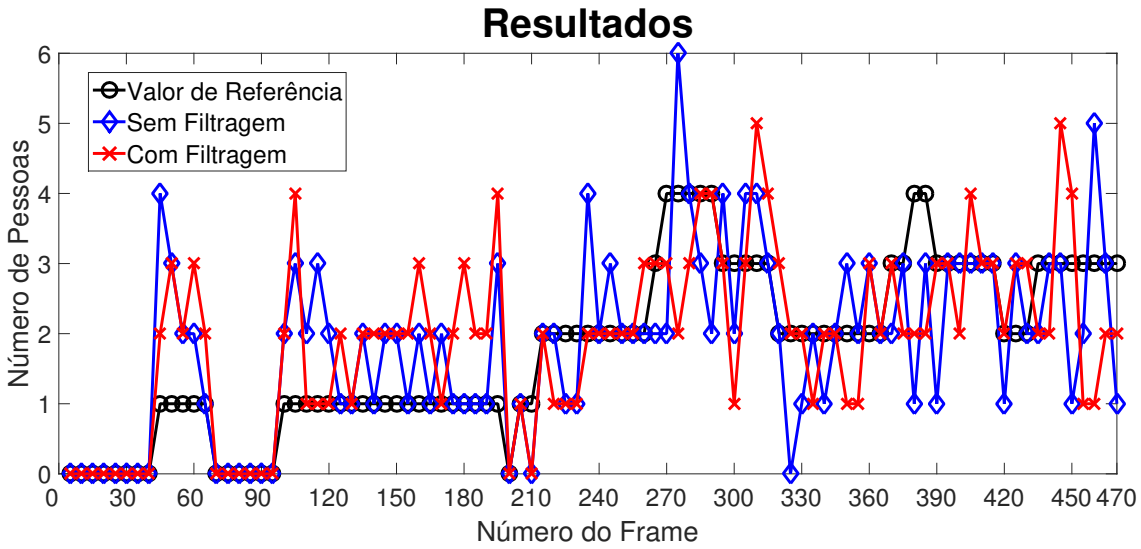


Figura 5.10: Resultados para Vídeo 5

Neste caso, o Resultado Sem Filtragem foi ligeiramente melhor, mas ambos os Resultados foram praticamente equivalentes, com Média de Erro Absoluto em torno de 0,7 pessoas e Média de Acurácia em torno de 75%. As Médias do Número de Pessoas Estimadas foram bem próximas à Média do Valor de Referência, mas a Média do Resultado Sem Filtragem foi a mais próxima.

Médias	Sem Filtragem	Com Filtragem
Média do nº de Pessoas Estimadas	1,84	1,91
. Média do nº de Pessoas do Valor de Referência	1,78	1,78
Média de Erro Absoluto do nº de Pessoas	0,66	0,71
Média de Acurácias	76,4%	74,1%

Tabela 5.5: Métricas para Vídeo 5

Alguns valores foram maiores que os de Referência devido ao fato de que os membros das pessoas se movem com velocidades diferentes da velocidade do centro do corpo (como os braços ou as pernas) e isso algumas vezes é contado como outro objeto, gerando Falsos Positivos.

Alguns valores foram menores que os de Referência devido ao problema de que quando as duas pessoas estão andando juntas e aproximadamente com a mesma velocidade, são contadas como o mesmo objeto, gerando Falsos Negativos.

Capítulo 6

Conclusão e Trabalhos Futuros

Com o aumento populacional e a urbanização, aliados à grande disponibilidade de câmeras, cresce um interesse em saber quantas pessoas transitam pelos espaços públicos, aumentando assim a relevância e a importância da Estimação de Número de Pessoas. Dessa forma, a Estimação do Número de Pessoas baseada em imagens de vídeo vindas de sistemas de vigilância e de segurança tem grande utilidade para espaços públicos, possuindo aplicações comerciais, para saber qual o horário em que uma loja é mais movimentada, por exemplo, e aplicações de segurança, como na detecção de acidentes e auxílio de seguranças e policiais.

Este trabalho apresentou uma Implementação para a Estimação do Número de Pessoas em Vídeos baseado no artigo de Rao *et al* [2], utilizando contagem por agrupamento. O Fluxo Óptico de Lukas-Kanade [6] foi empregado para gerar os vetores velocidade dos objetos, que passaram por uma Filtragem Espacial ou blocagem e por uma Filtragem Temporal para computar o máximo valor em cada bloco dentro de um intervalo de tempo, e posteriormente o Agrupamento Hierárquico também foi utilizado para agrupar tais velocidades em clusters e estimar o número de pessoas. Além disso, um passo de Filtragem de Vetores Similares foi proposto antes do Agrupamento Hierárquico a fim de evitar a contagem de diferentes objetos para vetores similares, de modo que cada Estimação possui um Resultado Sem Filtragem e um Resultado Com Filtragem.

Foram testados 5 vídeos, com número de pessoas variando entre 0 e 5. Os Resultados foram apresentados por vídeo, totalizando a análise de 1040 frames e apresentando 214 estimações de números de pessoas (1 resultado a cada 5 frames devido à Filtragem Temporal). Os valores dos Resultados foram bons mas não é possível uma comparação direta com o artigo de Rao *et al* [2], porque nele são apresentados os valores para apenas poucos frames. Um resultado importante observado nos testes mostra que o algoritmo, na grande maioria das vezes, não acusa que há pessoas se de fato não há. Isso ocorre porque se o passo de Agrupamento Hierárquico só gera distâncias muito pequenas (consideradas como ruídos), o passo de Estimação do Número de Pessoas já considera que não há ninguém naquele conjunto de frames.

Quatro métricas são apresentadas para cada vídeo: Média do Número de Pessoas Estimadas, Média do Valor de Referência, Média do Erro Absoluto e Média das Acurácias. Ambos os Resultados Com e Sem Filtragem apresentaram resultados similares. As Médias do Número de Pessoas Estimadas foram próximas às Médias do Valor de Referência. Para os Resultados Sem Filtragem, a Média do Erro Absoluto no número de Pessoas para cada

vídeo variou de 0,07 a 1,03 pessoas enquanto a Média de Acurácia variou de 69,0% a 98,2%. Já para os Resultados Com Filtragem, a Média do Erro Absoluto no número de Pessoas para cada vídeo variou de 0,14 a 1,00 pessoa enquanto a Média de Acurácia variou de 69,1% a 95,2%.

A principal causa de Falsos Positivos é o movimento dos membros, como os braços e as pernas, quando se movem ou se articulam em velocidades diferentes da velocidade do centro do corpo, acusando duas ou às vezes mais pessoas quando de fato há apenas uma pessoa. Por sua vez, a principal causa de Falsos Negativos ocorre devido ao fato de pessoas diferentes estarem andando na mesma trajetória com praticamente a mesma velocidade, implicando a contagem de só uma pessoa, quando na realidade há duas pessoas.

Possíveis trabalhos futuros poderiam tentar desprezar esses movimentos dos braços e pernas não coerentes com o movimento do centro do corpo, verificando que é um membro pela associação com o método de Correspondência de Formatos [1]. Também tentar diferenciar quando duas pessoas estiverem andando uma ao lado da outra a fim de evitar a contagem como uma só pessoa, utilizando juntamente a técnica de Correspondência de Formatos ou a técnica de contagem por regressão [1]. Além disso, outro trabalhos futuros poderiam testar outras Métricas de Dissimilaridade no cálculo do Agrupamento Hierárquico, ao invés da Distância Euclidiana, como a Distância de Manhatan ou Distância Máxima, e ainda outros Critérios de Ligação ao invés do *Single-Linkage clustering*, como o *Complete-linkage clustering* ou o *Average-linkage clustering* e comparar o resultado final com o Valor de Referência, analisando qual combinação de técnicas tem o melhor resultado. Outros trabalhos poderiam tentar calcular o melhor limiar de porcentagem para desprezar os valores pequenos na última fase de Estimação de Número de Pessoas, assim como calcular o melhor limiar de porcentagem que considera se os objetos são os mesmos ou não, ao invés de utilizar limiares fixos, ou ainda utilizar uma abordagem de treinamento supervisionado para calcular esses limiares. Outros possíveis trabalhos futuros poderiam também utilizar múltiplas câmeras com diferentes pontos de vista, realizando a contagem para cada ponto de vista separadamente e analisando as possíveis discrepâncias, tentando resolver os problemas de oclusão inter-objeto e tornando o método mais robusto.

Apêndice A

Código Fonte

O Código fonte desenvolvido neste trabalho pode ser consultado em:
https://github.com/MarinaMartins/PeopleCounting

Referências

[1] Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. Crowd counting and profiling: Methodology and evaluation, 2013. 1, 2, 15, 16, 17, 18, 46

[2] A. S. Rao, J. Gubbi, S. Marusic, P. Stanley, and M. Palaniswami. Crowd density estimation based on optical flow and hierarchical clustering. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 494–499. IEEE, Aug 2013. 1, 2, 16, 19, 24, 25, 26, 45

[3] Chiao-Fe Shu, A. Hampapur, M. Lu, L. Brown, J. Connell, A. Senior, and Yingli Tian. Ibm smart surveillance system (s3): a open and extensible framework for event based surveillance. In *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.*, pages 318–323, Sept 2005. 1

[4] J. C. Silveira Jacques Junior, S. R. Musse, and C. R. Jung. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, 27(5):66–77, Sept 2010. 1, 2

[5] V. Rabaud and S. Belongie. Counting crowded moving objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 705–711, June 2006. 1, 16, 17

[6] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc. 2, 8, 9, 10, 19, 21, 45

[7] Robert Fisher EC Funded CAVIAR project/IST 2001 37540. Caviar test case scenarios. `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`, 2007. [Online; acessado em 15/10/2017]. 2, 19, 33

[8] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006. 3

[9] A. Murat Tekalp. *Digital Video Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995. 4

[10] Alan C. Bovik. *Handbook of Image and Video Processing (Communications, Networking and Multimedia)*. Academic Press, Inc., Orlando, FL, USA, 2005. 5

[11] Aaron Bobick. Cs 4495 computer vision motion and optic flow. https://www.cc.gatech.edu/~afb/classes/CS4495-Fall2014/slides/CS4495-OpticFlow.pdf, 2014. 5

[12] V. Argyriou, J.M.D. Rincon, B. Villarini, and A. Roche. *Image, Video and 3D Data Registration: Medical, Satellite and Video Processing Applications with Quality Metrics*. Wiley, 2015. 5, 7, 8, 9, 10

[13] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981. 7, 8, 10, 17

[14] J.J. Gibson. *The perception of the visual world*. Houghton Mifflin, 1950. 7

[15] S. Susan Young, Ronald G. Driggers, and Eddie L. Jacobs. *Signal Processing and Performance Analysis for Imaging Systems*. Artech House, Inc., Norwood, MA, USA, 2008. 7, 21

[16] Ricardo Linden. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, 4:18–36, 2004. 11, 13

[17] Michel Marie Deza and Elena Deza. *Encyclopedia of Distances*. Springer Berlin Heidelberg, 2009. 11

[18] Eugene F. Krause. Taxicab geometry. *Mathematics Teacher*, 66, 12 1973. 11

[19] Carl Hammer. Principles of mathematical analysis: by walter rudin. 227 pages, new york, mcgraw-hill book co., inc., 1953. *Journal of the Franklin Institute*, 256, 1953. 11

[20] Michiel Hazewinkel. Mahalanobis distance. *Encyclopedia of Mathematics, Springer Science+Business Media B.V. / Kluwer Academic Publishers, ISBN 978-1-55608-010-4*, ed. (2001) [1994]. 11

[21] Prof. Dr. Paulo A. V. de Miranda. Métodos de agrupamento (clustering). http://www.vision.ime.usp.br/~pmiranda/mac6903_2s12/aulas/aula18.pdf, 2012. 11, 13, 14

[22] Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, Sep 1967. 11

[23] Leonardo Marques Rocha, Fábio A. M. Cappabianco, and Alexandre Xavier Falcão. Data clustering as an optimum-path forest problem with applications in image analysis. *International Journal of Imaging Systems and Technology*, 19(2):50–68, 2009. 13

[24] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern classification, 2nd ed, 2001. 13

[25] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004. 15

[26] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005. 15

[27] Bo Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 90–97 Vol. 1, Oct 2005. 15

[28] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. 15

[29] Paul Viola, Michael J. Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, Jul 2005. 15

[30] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2188–2202, Nov 2011. 15

[31] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept 2010. 15

[32] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008. 15, 16

[33] W. Ge and R. T. Collins. Marked point processes for crowd counting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2913–2920, June 2009. 16

[34] D. B. Yang, H. H. Gonzalez-Banos, and L. J. Guibas. Counting people in crowds with a real-time network of simple image sensors. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 122–129 vol.1, Oct 2003. 16

[35] Weina Ge and Robert T. Collins. Crowd detection with a multiview sampler. In *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV'10, pages 324–337, Berlin, Heidelberg, 2010. Springer-Verlag. 16

[36] M. Wang, W. Li, and X. Wang. Transferring a generic pedestrian detector towards specific scenes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3274–3281, June 2012. 16

[37] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 878–885 vol. 1, June 2005. 16

[38] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1198–1211, July 2008. 16

[39] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 594–601, June 2006. 16, 17

[40] A. C. Davies, Jia Hong Yin, and S. A. Velastin. Crowd monitoring using image processing. *Electronics Communication Engineering Journal*, 7(1):37–47, Feb 1995. 17

[41] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, Nov 1973. 18

[42] W. Ma, L. Huang, and C. Liu. Crowd density analysis using co-occurrence texture features. In *5th International Conference on Computer Sciences and Convergence Information Technology*, pages 170–175, Nov 2010. 18

[43] Tsung-Yi Lin, Yen-Yu Lin, Ming-Fang Weng, Yu-Chiang Frank Wang, Yu-Feng Hsu, and Hong-yuan Liao. Cross camera people counting with perspective estimation and occlusion handling. In *IEEE International Workshop on Information Forensics and Security*, 11 2011. 18

[44] Ke Chen, Chen Change Loy, Shaogang Gong, and Tao Xiang. Feature mining for localised crowd counting. In *British Machine Vision Conference*, 01 2012. 18

[45] International Organization for Standardization. *ISO/IEC 13818-1:2000: Information technology — Generic coding of moving pictures and associated audio information: Systems*. 2000. 19, 33

[46] MATLAB R2015b. *(version 8.6)*. The MathWorks Inc., Natick, Massachusetts, 2015. 19, 33

[47] MATLAB. *Computer Vision System Toolbox (R2015b)*. The MathWorks Inc., Natick, Massachusetts, 2015. 19

[48] MATLAB. *Statistics and Machine Learning Toolbox (R2015b)*. The MathWorks Inc., Natick, Massachusetts, 2015. 19

[49] International Organization for Standardization. *ISO/IEC 10918-1:1994: Information technology — Digital compression and coding of continuous-tone still images: Requirements and guidelines*. 1994. 33

[50] Baratloo A, Hosseini M, Negida A, and El Ashal G. Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Emergency*, 3(2):48–49, 2015. 34

[51] Adilson Gonzaga. Métodos de avaliação de classificadores. `http://iris.sel.eesc.usp.br/sel886/Aula_9.pdf`, 2017. 34