UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Jorge Leonid Aching Samatelo

Uma Técnica de Subtração de Fundo em Vídeos de Monitoramento

VITÓRIA
2012

Jorge Leonid Aching Samatelo

Uma Técnica de Subtração de Fundo em Vídeos de Monitoramento

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Doutor em Engenharia Elétrica.
Orientador: Prof. Dr. Evandro Ottoni Teatini Salles.

VITÓRIA
2012

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

Samatelo, Jorge Leonid Aching, 1975-

S187t Uma técnica de subtração de fundo em vídeos de monitoramento / Jorge Leonid Aching Samatelo. - 2012.

197 f. : il.

Orientador: Evandro Ottoni Teatini Salles.

Tese (Doutorado em Engenharia Elétrica) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Processamento de imagens. 2. Visão por computador. 3. Sistemas de reconhecimento de padrões. 4. Vigilância eletrônica. 5. Câmaras de vídeo. I. Salles, Evandro Ottoni Teatini. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 621.3

Jorge Leonid Aching Samatelo

Uma Técnica de Subtração de Fundo em Vídeos de Monitoramento

Tese submetida ao programa de Pós-Graduação em Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para a obtenção do Grau de Doutor em Engenharia Elétrica.

Aprovada em 14 de dezembro do 2012.

COMISSÃO EXAMINADORA

Prof. Dr. Evandro Ottoni Teatini Salles
Universidade Federal do Espírito Santo
Orientador

Prof. Dr. Mário Sarcinelli Filho
Universidade Federal do Espírito Santo

Prof. Dr. Hans Jorg Andreas Schneebeli
Universidade Federal do Espírito Santo

Prof. Dr. Klaus Fabian Côco
Universidade Federal do Espírito Santo

Prof. Dr. Lee Luan Ling
Universidade Estadual de Campinas

A mis amigos de toda una vida:
Cesar, Marlene y Cesar Domingo.

A mis amigos que se fueron pero me dejaron un ejemplo a seguir:
Ursula y Temistocles.

A mis amigos que no obstante la distancia están siempre presentes:
Enith y Jorge.

A mi compañera en este caminar:
Gabriela.

A mi mas nuevo amigo:
Alessandro.

Agradecimentos

Agradeço à minha família, que sempre esteve presente, não obstante a distância, a meus pais Cesar e Marlene, pelo amor e apoio sempre oferecido, a meu irmão Cesar, pela confiança e fortaleza que me brinda sempre.

A Gabriela Callo Quinte, pelo apoio incondicional para o meu desenvolvimento pessoal e profissional, pela ajuda com os textos em inglês e por compartilhar sua história com a minha.

A meu orientador, Prof. Dr. Evandro Ottoni Teatini Salles pela oportunidade oferecida, e por ser um exemplo de dedicação, e compromisso com a pesquisa.

Aos docentes do Programa de Pós-Graduação em Engenharia Eletrica da UFES, os quais com muita dedicação contribuíram para minha formação.

A meus colegas do laboratório de Computação e Sistemas Neurais (CISNE), em especial Fernando Kentaro, Karin Komati, Patrick Marques, Alex Brandão, Felipe Pedroso, Janayna Passarinho e Anibal Cotrina, que dividiram comigo o desafio do conhecimento e crescimento profissional.

A meus amigos, em especial, Patric-*Bahiano*, Gregorio, Cesar, Richard, Zui, Jhon, Carlos, Camilo, Willian, Cristopher, Dennis, Chie, David, Talita, Beatrice e Heidy, pelos inúmeros momentos de solidariedade e alegria.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro.

E a todas as outras pessoas que direta ou indiretamente colaboraram com o sucesso deste trabalho.

Sumário

Lista de Tabelas

[1]As percentagens entre parênteses indicam a proporção dos quadros (quer seja para teste ou treino) em relação ao número total de quadros do vídeo em questão.

Lista de Figuras

Glossário

abertura do primeiro plano é um dos desafios tratados por uma técnica de subtração de fundo. O problema ocorre quando um objeto do primeiro plano de cores homogêneos se desloca e as variações entre quadros na cor dos píxeis internos do objeto em movimento não são detectadas. Assim, todo o objeto pode ser excluído do primeiro plano, causando uma detecção de falsos negativos. 10, 95, 100

brilho é um atributo da percepção visual, na qual uma área parece emitir mais ou menos luz. Note-se que o conceito de brilho não depende da cor da luz. Assim, cores de igual brilho no cubo RGB são aqueles cujas três componentes de cor somam o mesmo valor. Portanto, um determinado nível de brilho é representado no cubo RGB como um plano perpendicular ao eixo neutro ($R + G + B =$ constante). xix, xxi, 39, 41, 44, 112, 138–142, 144–146, 149, 150, 152

camuflagem é um dos desafios tratados por uma técnica de subtração de fundo. Nesta situação, alguns objetos do primeiro plano diferem pouco da aparência do fundo, tornando difícil uma correta classificação, produzindo uma detecção de falsos negativos. 10, 82, 95, 100, 105

critério de discrepância é o critério utilizado na avaliação de uma técnica de subtração de fundo, que define as características usadas para o cálculo das métricas de desempenho. Assim, pode ser definido como critério o número de píxeis classificados como primeiro plano ou o número de objetos detectados em movimento numa sequência de vídeo. xxiii, xxv, 52, 55, 56, 58

cromaticidade é a propriedade que a média das pessoas consideram como a cor da luz, sendo definida pelas propriedades de matiz e saturação. Também existem outros pares de propriedades (diferentes de matiz e saturação) que podem ser usados para descrever a cromaticidade, permitindo assim diferentes modelos de cor. xxi, 39, 112, 138–140, 149, 152, 153

curva de precisão-sensibilidade é o gráfico da sensibilidade (m_{TVP}) versus a precisão (m_{VPP}). xviii, xix, 63, 93, 96, 97, 99, 101, 102, 104, 105, 113

curva de exatidão é o gráfico da métrica de exatidão (m_E) em relação ao parâmetro de sobreposição ($\tau_{\text{casamento}}$). 63, 113

curva ROC é o gráfico da taxa de falsos positivos (m_{TFP}) versus a taxa de verdadeiros positivos (m_{TVP}). 62, 63, 113

detecção de falsos negativos é o caso onde uma uma entidade do primeiro plano é identificada como parte do fundo. xxi, 9, 10

detecção de falsos positivos é o caso onde uma região do fundo é identificada como uma entidade do primeiro plano. xxiv, 9–11, 14, 41, 112

detecção de variações também conhecida como detecção do primeiro plano, identifica píxeis nos quadros que não podem ser explicados pelo modelo do fundo. O resultado deste passo é a máscara do primeiro plano. xx, xxii, xxxiii, 6, 19, 20, 29, 32–34, 38–42, 50, 112, 152, 153

distância Euclidiana num espaço euclidiano é a distância entre dois pontos, calculada através da norma L_2. xiii, xxx, xxxiii, 7, 38–41, 44, 45, 50, 93, 94, 99, 108, 112, 114, 119, 120, 123, 139

fundo região estática de uma cena que se encontra atrás dos objetos em movimento. xv, xxi–xxiv, xxvi, xxix, 2, 3, 5, 8–16, 18–20, 22–24, 26, 27, 29, 36, 41, 44, 50, 51, 54, 56, 57, 60, 74, 77, 78, 82, 89, 95, 99

fundo dinâmico é um dos desafios tratados por uma técnica de subtração de fundo. Nesta situação, algumas partes do cenário apresentam um certo grau de movimento, devendo estas ser consideradas como parte do fundo. Por exemplo, as luzes de um semáforo, as folhas das árvores movidas pelo vento. 9, 14, 15, 24, 27, 30, 77, 82, 86, 89, 94, 99, 104, 107, 111, 113

ground-truth a tradução deste termo para o português é *levantamento de campo* ou *padrão-ouro*, e para o caso da avaliação das técnicas de subtração de fundo, faz-se referência ao conjunto de imagens segmentadas manualmente/automaticamente por um especialista/procedimento, servindo de padrão de comparação em relação aos quadros segmentados por uma determinada técnica de subtração de fundo a ser avaliada. Posto que na literatura especializada é usado o termo em inglês, este é mantido ao longo do texto. xvi–xix, xxiii, xxv, xxxii, 52, 53, 56–63, 65–67, 69, 71, 73–75, 77–79, 82, 84, 88, 90–92, 98, 103, 104, 106, 113, 114

máscara binária é uma imagem binária que indica as regiões de interesse numa imagem a cores ou em escala de cinza. Assim, na máscara binária as regiões sem interesse são rotuladas com 0 e as regiões de interesse são rotuladas com 1 . xxiii, 11, 19, 20, 53, 62

máscara do primeiro plano é uma máscara binária relacionada a um quadro, onde as regiões de interesse são definidas como aquelas áreas do quadro que contém um objeto em movimento. Assim, na máscara do primeiro plano as regiões do fundo são rotuladas com 0 e as regiões do primeiro plano são rotuladas com 1. xv, xxii, xxviii, 20, 29, 33, 34, 37, 38, 40, 42, 44, 50, 89, 91, 112, 126

matiz refere-se à tonalidade específica da cor. Assim, é a propriedade que distingue vermelho de laranja, de azul, e assim por diante. A matiz de um ponto no cubo RGB é definida em relação à sua posição angular em torno do eixo neutro. xix–xxi, xxiii, 39, 41, 44, 112, 138–140, 142–145, 149, 150, 152, 153

matriz de confusão é de utilidade quando se pretende determinar o desempenho de um problema de duas classes. xi, xiii, 53, 54, 56–59, 63, 64, 119–125

média móvel gaussiana é um tipo de média móvel, semelhante a uma média móvel simples, excepto que dá um maior peso aos dados mais recentes da série temporal tratada. Na literatura, é também conhecida como media móvel exponencial. xi, xii, xvi–xviii, xxix, 13, 21, 24, 27, 82–85, 93–100, 107, 108, 111

medida F é a medida que combina as métricas m_{VPP} e m_{TVP} numa única medida, através do cálculo da média harmônica dos dois valores. xi, xii, xviii, xix, 55, 57, 59, 91–97, 99–102, 104–108, 110, 111

métodos de discrepância empíricos também conhecidos como métodos de avaliação supervisionados, são métodos que permitem realizar a a avaliação das técnicas de subtração de fundo. Aqui, a avaliação é realizada pela comparação entre as regiões vinculadas ao primeiro plano detectado e as regiões vinculadas ao *ground-truth*. Como resultado desta comparação, uma medida que indica o grau de sobreposição das duas regiões é obtida. A definição desta medida implica considerar o critério de discrepância usado na avaliação. 52

métrica de desempenho é uma medida que permite avaliar o rendimento de uma técnica de subtração de fundo considerando o processo de avaliação como um problema de classificação binária ou de múltiplas classes. xxi, 53–55, 57, 59, 63

modelo do fundo é uma estimação do fundo de uma cena. xv, xxii–xxiv, xxvi, xxviii, xxx, xxxiii, 2, 3, 5, 6, 9–11, 13, 16–20, 23–25, 29–34, 37, 41, 44, 50, 82, 86, 94, 99, 111, 113, 114

mudanças graduais da iluminação é um dos desafios tratados por uma técnica de subtração de fundo. Este desafio descreve a capacidade de adaptação do modelo do fundo às mudanças graduais do ambiente. Por exemplo, em cenas externas, a intensidade da

luz tipicamente varia durante o dia e, portanto, a evolução da iluminação provocarÃ¡ uma mudança global da aparência do fundo. 9, 13, 22, 30

mudanças repentinas da iluminação é um dos desafios tratados por uma técnica de subtração de fundo. Este desafio faz referência às repentinas mudanças da iluminação que não são cobertas pelo modelo do fundo, afetando fortemente a aparência do fundo, e originam detecção de falsos positivos. Por exemplo, num cenário exterior, quando a iluminação numa rua é ligada. 9, 11–13, 18, 24, 27

nível de significância é a probabilidade de rejeitar a hipótese nula quando esta é verdadeira, uma decisão conhecida como: erro de tipo I, falso positivo ou falso alarme, é denotado por α, e a partir dele é definido o nível de confiança, o qual é igual a $1 - \alpha$. Assim, se é definido como hipótese nula que um píxel pertence ao fundo (primeiro plano) o fato de classificá-lo como parte do primeiro plano (fundo) gera um falso positivo, e sua probabilidade de ocorrência é estabelecida pelo valor de α. Portanto, um valor baixo para α sempre é recomendável. xviii, xxxiii, 36, 42, 50, 93, 94, 96, 97, 99, 101, 102, 112

objeto do primeiro plano faz referência a um objeto em movimento da cena. xxi, xxiv, xxvi, 9, 10, 13, 17, 19, 27, 78, 99, 111, 114

parâmetro de sobreposição é um parâmetro do processo de avaliação que estabelece qual é o grau de sobreposição entre dois retângulos para que eles sejam casados. Tal parâmetro é denotado pela variável $\tau_{\text{casamento}}$. xiii, xvi, xvii, xxii, xxxii, 59, 61, 82, 83, 86, 87, 90, 119–125

precisão também denominada como valor preditivo positivo (m_{VPP}), esta métrica indica a proporção de instâncias positivas que são realmente positivas num problema de classificação binária. xi, xii, xxi, xxxi, 55, 57, 59, 63, 91–97, 99–102, 104–106, 108, 112

primeiro plano região de uma cena que contém os objetos em movimento. xv, xxi–xxv, 2, 3, 5, 8–13, 16, 18–20, 23, 24, 27, 32, 36, 37, 41, 51–54, 56–58, 60, 61, 63, 77, 78, 82, 92, 94, 95, 99, 104, 112

primeiro plano adormecido é um dos desafios tratados por uma técnica de subtração de fundo. Este desafio faz referência à situação onde um objeto do primeiro plano que se torna imóvel, ao passar o tempo, passa a ser visto como parte do fundo. 10, 89, 107, 111, 113

problema de classificação binária faz referência ao problema de classificação de instâncias em duas classes. xi, xxiii–xxv, 53, 57, 59–61

problema de classificação de múltiplas classes faz referência ao problema de classificação de instâncias em mais de duas classes. xi, 62–64

procedimento de avaliação baseado em métricas orientadas a objetos abordagem de avaliação que usa como critério de discrepância o número de objetos detectados em movimento numa sequência de vídeo. xi, 58, 59, 61–65, 71, 73, 78, 112–114, 127

procedimento de avaliação baseado em métricas orientadas a píxeis abordagem de avaliação que usa como critério de discrepância os píxeis classificados como primeiro plano. xi, 56, 57, 61, 76, 91

processo de um píxel intuitivamente, são os valores que assume um píxel numa localização particular de uma imagem considerando um conjunto de quadros pertencentes a uma sequência de vídeo. Formalmente, é um processo estocástico discreto, ou série temporal vinculada a cada píxel de uma sequência de vídeo. xiv, xxviii, xxix, 12–14, 21–23, 30, 85, 86, 89, 104

quadro uma única imagem de um conjunto de imagens que formam uma sequência de vídeo. xii, xv, xvi, xxi–xxiii, xxv, xxvi, xxviii–xxx, xxxii, xxxiii, 2, 3, 5, 8, 10–13, 15, 17–27, 30, 32–34, 37, 38, 40, 42, 44, 48, 50, 52, 53, 56–58, 60, 61, 64, 67, 69, 71, 73–79, 91, 94, 95, 108, 110, 113, 116, 117, 126, 127, 130–133, 135–137

reflexões é um dos desafios tratados por uma técnica de subtração de fundo. Este desafio descreve o fato que uma cena pode refletir instâncias do primeiro plano, devido a superfícies molhadas ou refletoras, como a estrada, janelas, vidros, etc, e essas entidades refletidas não devem ser classificadas como parte do primeiro plano. Por exemplo, uma estrada molhada com o sol brilhando provocarÃ¡ a reflexão dos veículos que passam. 10

retângulo forma geométrica específica que contém um objeto a detectar no *ground-truth*. Por sua simplicidade, usam-se formas retangulares, sendo estas apropriadas para objetos tais como indivíduos, faces, texto, entre outros. Assim, para o caso de técnicas de subtração de fundo tem-se um retângulo para cada objeto em movimento no *ground-truth*. xvi–xviii, xxiv, xxv, xxxii, 58–67, 69, 73–75, 84, 88, 90, 113, 114

saturação refere-se à pureza da cor. Assim, é a propriedade que distingue vermelho de rosa. Uma cor muito saturada é viva e intensa, enquanto uma cor menos saturada é descolorida e cinzenta. Sem saturação uma cor torna-se um tom de cinza. xix, xxi, xxv, 39, 41, 138–140, 142–145, 150, 152

sensibilidade também denominada como a taxa de acerto (m_{TVP}), indica a proporção de instâncias positivas que são corretamente classificadas como positivas num problema

de classificação binária. xi, xii, xxi, xxxi, 54, 57, 59, 63, 91–94, 96, 97, 99–102, 104–106, 108, 112

sequência de vídeo conjunto de imagens estáticas, denominadas quadros, representando instantâneos de uma cena, tomadas a intervalos de tempo espaçados regularmente. Por exemplo, os principais sistemas de vídeo trabalham com cadências entre 25 e 30 quadros por segundo. xxi, xxv, xxvi, xxviii, xxxiii, 5, 12, 13, 15, 20, 22, 33, 53, 57, 58, 60–62, 77, 78

sistema de vigilância automatizado é um sistema que deve ser capaz de detectar a presença de objetos em movimento no campo de visão estabelecido por uma ou um conjunto de câmeras, classificá-los em diversas categorias, e rastrear estes objetos ao longo do tempo. Também deve ser capaz de gerar uma descrição dos eventos que ocorrem dentro do seu campo de visão. A definição de um objeto de interesse é dependente do contexto, mas para um sistema de vigilância automatizado, qualquer objeto deslocando-se de forma independente, por exemplo um pedestre ou um veículo, são considerados de interesse. xiv, 1–3, 8, 52

subtração de fundo procedimento que permite separar os objetos em movimento do resto da cena no campo visual de uma câmera (na maioria de casos estática), através da determinação e adaptação de um modelo do fundo, tal que os objetos em movimento são todos aqueles grupos de píxeis que são diferentes do modelo do fundo. 1, 2, 5, 8, 10, 11, 19, 23, 53

técnica de detecção de variações técnica que permite identificar os píxeis de um quadro que correspondem com os objetos do primeiro plano. xi–xiii, xvi–xviii, 40, 50, 82–84, 86–88, 94–103, 107, 112, 114, 119–121

técnica de subtração de fundo técnica que permite a identificação de objetos em movimento numa sequência de vídeo, onde cada quadro do vídeo é comparado a um modelo de referência do fundo.. xi–xiv, xvi–xix, xxi–xxv, xxxiii, 2, 3, 5–12, 15, 19, 21, 39–41, 50, 52, 53, 55–59, 61–63, 65, 67, 69, 71, 72, 75–80, 82–93, 96–103, 105–108, 112–114, 119–127

Siglas

CD Correta Detecção. 60, 63, 64

FA Falso Alarme. 60, 63, 64

FD Falha na Detecção. 60, 63, 64

HMM *Hidden Markov Model*. 17

KDE *Kernel Density Estimation*. 14

LBP *Local Binary Patterns*. 16

MACE *Minimal Average Correlation Energy*. 18

MATLAB *MATrix LABoratory*. 72, 73

MOG *Mixture Of Gaussians*. 14

PCA *Principal Component Analysis*. 17

PETS2004 *Performance Evaluation of Tracking and Surveillance 2004 Dataset*. xi, xii, xvi, 73, 75, 78–80, 107, 111–113, 116

RGB *Red-Green-Blue*. 16, 39

S Separação. 60, 63, 64

SABS *Stuttgart Artificial Background Subtraction Dataset*. xi, xiii, 73, 91–94, 99, 104, 107, 111, 112, 117

SU Separação - União. 60, 63, 64

U União. 60, 63, 64

YUV *Luminance* (Y), *blue-luminance* (U), *red-luminance* (V). 13

Lista de Símbolos

t indica o índice do quadro atual em relação ao conjunto de quadros de uma sequência de vídeo. xxix, xxx, 20, 22, 23, 25, 26, 30, 32, 33, 42

l sub-índice da localização do pixel (x_l, y_l) numa imagem de tamanho $N_{\text{fil}} \times N_{\text{col}}$, onde $l = 1, \cdots, N_{\text{fil}} \times N_{\text{col}}$. 20, 56

$\mathbf{I}(t)$ quadro no instante t de uma sequência de vídeo a cores, onde, $\mathbf{I}(t) \in [0, 255]^{N_{\text{fil}} \times N_{\text{col}} \times 3}$. 20

$\mathbf{I}_l(t)$ valor de um pixel na localização (x_l, y_l) no quadro $\mathbf{I}(t)$, onde, $\mathbf{I}_l(t) \in [0, 255]^3$. xiv, xix, xx, xxix, xxx, 20, 23–28, 32, 38–41, 44, 45, 139, 149, 150, 152, 153

$\mathbf{B}(t)$ modelo do fundo a cores calculado/atualizado no instante t, onde $\mathbf{B}(t) \in [0, 255]^{N_{\text{fil}} \times N_{\text{col}} \times 3}$. xv, 20, 32

$\mathbf{B}_l(t)$ valor de um pixel na localização (x_l, y_l) no modelo do fundo $\mathbf{B}(t)$, onde $\mathbf{B}_l(t) \in [0, 255]^3$. xix, xx, xxx, 20, 23, 24, 30, 32, 38–41, 44, 45, 139, 149, 150, 152, 153

$\hat{\mathbf{B}}_l(t)$ valor de um pixel estimado/atualizado na localização (x_l, y_l) do modelo do fundo, onde $\hat{\mathbf{B}}_l(t) \in [0, 255]^3$. xxviii, 24

$\mathbf{B}_C(l, t)$ valor de um pixel no canal $C \in \{R, G, B\}$ na localização (x_l, y_l) no modelo do fundo $\mathbf{B}(t)$, onde $\mathbf{B}_C(l, t) \in [0, 255]$. xxix, 31, 32

$\mathbf{M}(t)$ máscara do primeiro plano calculada no instante t, onde $\mathbf{M}(t) \in [0, 1]^{N_{\text{fil}} \times N_{\text{col}}}$. xv, 20, 43, 44, 50

$\mathbf{M}_l(t)$ valor de um rótulo na localização (x_l, y_l) na máscara do primeiro plano $\mathbf{M}(t)$, onde $\mathbf{M}_l(t) \in [0, 1]$. xxx, 20, 29, 33, 34, 36–38, 44, 50

$\mathcal{I}_l(\tau, s)$ processo de um píxel na localização (x_l, y_l) de uma determinada sequência de vídeo s para todo $\tau \in \mathbb{N}$. xiv, xxix, 22, 23

$\mathcal{I}_l(t)$ série temporal observada, vinculada ao processo de um píxel $\mathcal{I}_l(\tau, s)$, considerando que a parte observada corresponde aos últimos T quadros. em relação ao quadro atual t. xiv, 22–24, 30

T número de quadros que correspondem à parte observada de um processo de um píxel. xxix, 22

α_{svt} coeficiente de suavização da média móvel gaussiana, onde $\alpha_{svt} \in [0, 1]$. 24, 80, 94

K_{mg} número de Gaussianas da mistura. 25–27, 29, 80, 94

$\omega_{l,k}(t)$ peso da k^{th} Gaussiana da mistura no quadro t. xiv, xxix, 25–28

$\boldsymbol{\mu}_{l,k}(t)$ valor médio da k^{th} Gaussiana da mistura no quadro t. xv, 25–28

$\boldsymbol{\Sigma}_{l,k}(t)$ matriz de covariância da k^{th} Gaussiana da mistura no quadro t. É assumido que tenha a forma $\boldsymbol{\Sigma}_{l,k}(t) = \sigma^2_{l,k}(t)\mathbf{I}$. xxix, 25, 26

D_{vz} parâmetro que permite dimiuir o raio da vizinhança centrada em $\mathbf{I}_l(t)$, onde, $D_{vz} \in [0, 1]$. xxix, 25, 27, 80, 94

$Z_{l,k}(t)$ variável binária que é 1 se o valor $\mathbf{I}_l(t)$ é próximo à média $\boldsymbol{\mu}_{l,k}(t-1)$, e caso contrario é 0. xiv, xv, 25, 26, 28, 29

α_{apr} taxa de aprendizagem dos pesos $\omega_{l,k}(t)$, onde $\alpha_{apr} \in [0, 1]$. xiv, xix, 26, 28, 80, 93, 94, 104, 105, 113

b_{mg} número de componentes da mistura que definem o fundo em relação ao limiar T_{fundo} aplicado sobre a soma dos pesos. xxix, 29

$T_{\textbf{fundo}}$ limiar aplicado sobre a soma dos pesos que define o número de componentes da mistura b_{mg} que representam o fundo. xxix, 29, 80, 94

$h^C_l(b)$ número de ocorrências da intensidade b na posição (x_l, y_l) do canal $C \in \{R, G, B\}$, considerando-se 256 níveis de intensidade (escala de cinza). xv, xxix, 30–32

$\hat{h}^C_l(b)$ histograma $h^C_l(b)$ filtrado, usando um filtro de ponderação de tamanho $2N_{\text{filtro}} + 1$. 31

$\sigma^C_l(b)$ medida de dispersão vinculada ao histograma $h^C_l(b)$. xxix, 31, 32

σ_{ds} parâmetro que define o tamanho da vizinhança centrada em $\mathbf{B}_C(l, t)$ na qual é calculada $\sigma^C_l(b)$. 31, 32, 80, 94

$\mathbf{D}_l(t)$ diferença entre o modelo do fundo e cada quadro do vídeo, para cada pixel l, num instante t, calculado na escala de cinza. 33, 34

$\mathcal{H}_{\mathbf{var}}$ ação de decidir por $\mathbf{M}_l(t) = 1$. 34–37

$\mathcal{H}_{\mathbf{invar}}$ ação de decidir por $\mathbf{M}_l(t) = 0$. 34–37

$\sigma_{\mathbf{var}}$ variância da distribuição Gaussiana com média igual a zero para a probabilidade condicional $p(\mathbf{D}_r(t)|\mathcal{H}_{\mathrm{var}})$. 35, 36

$\sigma_{\mathbf{invar}}$ variância da distribuição Gaussiana com média igual a zero para a probabilidade condicional $p(\mathbf{D}_r(t)|\mathcal{H}_{\mathrm{invar}})$. 35, 36

$\mathcal{D}_l(t)$ vizinhança $\mathcal{N}_l$ centrada no pixel de análise l. xxx, 35, 36

$N_{\mathcal{D}}$ cardinalidade da vizinhança $\mathcal{D}_l(t)$. 35, 36, 80

$s_l(t)$ suficiência estatística vinculada à vizinhança $\mathcal{D}_l(t)$. 36

τ_σ limiar de decisão dependente dos parâmetros σ_{var} e σ_{invar}. 36

$\tau_{\mathbf{prior}}$ limiar de decisão dependente das probabilidades a priori das decisões $\mathcal{H}_{\mathrm{var}}$ e $\mathcal{H}_{\mathrm{invar}}$. 36, 37

α nível de significância estatística. xviii, xxiv, 36, 42, 44, 50, 80, 93, 94, 96, 97, 99, 101, 102, 112

θ ângulo entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$. xxx, 39–42, 80, 94, 150–153

λ medida angular entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$. 39–41, 44–46, 149

$\tilde{\lambda}$ medida angular entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ quando θ é limiarizado. 41, 42

$d_l(t)$ distância Euclidiana entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$. xv, 38–42, 44–47

d_{0^o} distância Euclidiana entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ quando $\theta = 0^o$. 40–42, 44, 149, 150

d_{90^o} distância Euclidiana entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ quando $\theta = 90^o$. 40, 41, 44, 149, 150

$\rho_l(t)$ medida de similaridade das cores entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$. xv, 41, 42, 44

κ variável que quantifica a variação em relação à referência $||\mathbf{B}_l(t)||$. 45

$d_P(l,t)$ distância Euclidiana entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ quando estes têm a mesma magnitude, porém diferentes orientações. xv, xxxi, 45–47

$d_M(l,t)$ distância Euclidiana entre os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ quando estes têm a mesma orientação, porém diferentes magnitudes. xv, xxxi, 45–47

$P(l,t)$ distância $d_P(l,t)$ projetada sobre o plano $D(l,t) = M(l,t) + P(l,t)$. xv, xxxi, 47–49

$M(l,t)$ distância $d_M(l,t)$ projetada sobre o plano $D(l,t) = M(l,t) + P(l,t)$. xv, xxxi, 47–49

$P''_F(l,t)$ distância $P(l,t)$ projetada sobre o plano $D''_F(l,t) = 0$. 48

$M''_F(l,t)$ distância $M(l,t)$ projetada sobre o plano $D''_F(l,t) = 0$. 48

$N_{\mathbf{MinArea}}$ número de píxeis dos objetos a ser filtrados na etapa de pós-processamento. 38, 80, 94, 108

$\mathcal{H}_0$ hipótese nula. 44, 50, 53, 54, 56–59

$\mathcal{H}_1$ hipótese alternativa. 53, 56, 58, 59

N_{VP} número de instâncias positivas que são classificadas como positivas. 53–57, 59

N_{FP} número de instâncias negativas que são classificadas como positivas. 53–55, 57, 59, 60

N_{FN} número de instâncias positivas que são classificadas como negativas. 53–55, 57, 59, 60

N_{VN} número de instâncias negativas que são classificadas como negativas. 53–55, 57, 59, 62

m_{TVP} taxa de verdadeiro positivo, esta métrica também é referida como a taxa de acerto ou sensibilidade, e indica a proporção de instâncias positivas que são corretamente classificadas como positivas. xxi–xxiii, xxv, 54, 55, 57, 59, 96, 97, 101, 102, 104–106

m_{TFP} taxa de falso positivo, esta métrica também é referida como a taxa de falso alarme, e indica a proporção de instâncias negativas que são erroneamente classificadas como positivas. xxii, 54, 59

m_{TFN} taxa de falso negativo, esta métrica indica a proporção de instâncias positivas que são erroneamente classificadas como negativas ($= 1 - m_{TVP}$). 54

m_{TVN} taxa de verdadeiro negativo, esta métrica também é referida como a especificidade, e indica a proporção de instâncias negativas que são corretamente classificadas como negativas. 55, 57

m_{VPP} valor preditivo positivo, esta métrica também é referida como precisão, e indica a proporção de instâncias positivas que são realmente positivas. xxi, xxiii, xxiv, 55, 57, 59, 96, 97, 101, 102, 104–106

m_F medida F, é uma medida que combina as métricas m_{VPP} e m_{TVP} numa única medida, através do cálculo da média harmônica dos dois valores. 55, 57, 96, 97, 101, 102, 105, 106

m_E exatidão, esta métrica também é referida como exatidão preditiva, e é a proporção de instâncias que são corretamente classificadas. xvi, xxii, xxxii, 55, 57, 63, 65–68, 79

m_{ERR} taxa de erro, é a proporção de instâncias que são incorretamente classificadas. 55

m_{JAC} coeficiente de Jaccard, é a proporção de instâncias que são corretamente classificadas, desconsiderando as instâncias negativas. 55, 57

$\tau_{\textbf{casamento}}$ parâmetro de sobreposição, estabelece qual é o grau de sobreposição entre dois retângulos para que eles sejam casados, onde $\tau_{\text{casamento}} \in [0, 1]$. xvi, xxii, xxiv, xxxii, 61, 63, 65–67, 69, 70, 79, 113

$\mathcal{R}_{gt_i}$ i-ésimo retângulo, referente ao i-ésimo objeto presente no *ground-truth*. 58, 60, 61

$\bar{\mathcal{R}}_{gt}$ o conjunto de retângulos do *ground-truth*. 58, 59, 67, 69

$\mathcal{R}_{dt_j}$ j-ésimo retângulo, referente ao j-ésimo objeto detectado. 58, 60, 61

$\bar{\mathcal{R}}_{dt}$ conjunto de retângulos do quadro atual. 58, 59, 67, 69

$N_{\textbf{quadros}}$ número de quadros de um vídeo específico. 64, 69

$N_{\textbf{vídeos}}$ número de vídeos presentes num determinado banco de dados. 64, 69

$\mathbf{D}_{\textbf{decs}}$ matriz de decisão, indica os casamentos entre os retângulos presentes no quadro atual e no *ground-truth*. 60, 61, 65

N_{FD} número de ocorrências de um casamento um a zero. 60–65

N_{FA} número de ocorrências de um casamento zero a um. 60–65

N_{CD} número de ocorrências de um casamento um a um. 60–65

N_S número de ocorrências de um casamento um a muitos. 60–65

N_U número de ocorrências de um casamento muitos a um. 60–65

N_{SU} número de ocorrências de um casamento muitos a muitos. 60–65

τ_L limiar localizado no lado esquerdo da curva $m_E$vs $\tau_{\text{casamento}}$, chamado de limiar liberal. xiii, xxxii, 66, 67, 81–83, 85–87, 89, 90, 119–125

τ_C limiar localizado no lado direito da curva $m_E$vs $\tau_{\text{casamento}}$, chamado de limiar conservador. xxxii, 66, 67, 82, 83, 86, 87, 90

τ_R limiar definido como a média dos limiares τ_L e τ_C, chamado de limiar razoável. 67, 81–83, 85–87, 89, 90

Resumo

Aplicações de vigilância baseadas em vídeo, tais como monitoramento de atividades, rastreamento e identificação de objetos, requerem como etapa inicial a detecção dos objetos em movimento existentes numa sequência de vídeo. A detecção é um problema complexo, devido às diferentes variáveis a considerar, tais como a natureza da captura dos vídeos, podendo ser imagens capturadas ao ar livre ou em interiores, variações na iluminação e ruído, entre outros. As abordagens que permitem detectar os objetos em movimento a partir de uma cena estacionária são denominadas técnicas de subtração de fundo, e sua implementação requer a elaboração de um conjunto de etapas, denominadas pré-processamento, modelamento do fundo, detecção de variações e pós-processamento. Neste trabalho é implementada uma técnica de subtração de fundo envolvendo a elaboração de cada uma das etapas indicadas. Consequentemente: (a) diversas técnicas são realizadas para a etapa de modelamento do fundo, permitindo explorar as diferentes perspectivas de solução e como elas lidam com os desafios da área; (b) duas técnicas são propostas para a etapa de detecção de variações, que operam no nível de cores e estão baseadas na distância Euclidiana calculada entre cada quadro do vídeo e o modelo do fundo, e numa operação de limiarização, tal que o limiar é definido em relação a um nível de significância estatístico; (c) um procedimento de avaliação que permite determinar o desempenho de uma técnica com estas características é proposto, baseado no cálculo de uma métrica de desempenho, que representa quão exata é uma técnica de subtração de fundo em encontrar os objetos em movimento ao longo dos quadros numa sequência de vídeo, podendo ser determinada para cada vídeo ou para todo o banco de dados. Finalmente, para testes são utilizados dois bancos de dados: o primeiro próprio para o desenvolvimento e teste de um sistema de vigilância em espaço público, e o segundo especializado em analisar individualmente alguns dos desafios de que deve tratar uma técnica de subtração de fundo. Ao final desta tese são apresentados os resultados obtidos e são feitas as considerações finais.

Abstract

Video surveillance applications, such as activity monitoring, tracking and identification of objects, etc, require as an initial step the detection of the moving objects existing in a video sequence. The detection step is a challenging task, because of the different variables to consider, such as the nature of the video capture, images captured in indoor or outdoor environments, variations in lighting and noise, etc. The approaches that can detect moving objects from a stationary scene are called as background subtraction techniques, and their implementation requires the development of a set of steps, namely pre-processing, background modeling, change detection and postprocessing. So, here it is implemented a background subtraction technique that includes in the development of each steps after mentioned. Consequently: (a) various techniques are performed for the background modeling step, allowing to explore different solutions and how they handle the typical challenges; (b) two techniques, that working in the color level are proposed for the change detection step, based on the Euclidean distance calculated between each video frame and the model of the background, and a thresholding operation, such that the threshold is defined in relation to a statistical significance level; (c) an evaluation procedure for determining the performance of a technique with these features is proposed, based on the calculation of a performance metric that represents how accurate is the technique of background subtraction to find the moving objects over the frame in a video sequence, and can be determined for each video or the entire database. Finally, we used two databases: the first is suitable for the development and testing of a monitoring system in public space and the second is specialized in analyzing some challenges that a background subtraction technique should handle. In the end of this thesis the results obtained are presented and some concluding remarks are highlighted.

Capítulo 1

Introdução

1.1 Motivação

Um sistema de vigilância urbana em geral emprega uma ou mais câmeras que capturam um grande número de imagens por dia (nos Estados Unidos mais de 30 milhões de câmeras produzem cerca de 600 milhões de horas de vídeo por dia [69]). Armazenar e permitir uma busca por eventos automatizada é prioritário, pois o volume de informação a processar ultrapassa a capacidade humana. Assim, um sistema de vigilância automatizado é atualmente de grande interesse, sendo as aplicações com maiores expectativas o monitoramento de rodovias e auto-estradas, o reconhecimento de ações humanas, a classificação e o seguimento de objetos.

A implementação de sistemas de vigilância automatizados implica na elaboração de técnicas de análise de vídeo que precisam ser robustas em relação às condições típicas do ambiente urbano, tais como mudanças climáticas (por exemplo, chuva, neve, fortes efeitos de sombra), variações na iluminação devido ao dia e à noite, e cenas com muitos alvos. De maneira geral, um sistema de vigilância automatizado é composto de três etapas principais (ver [22]):

- subtração de fundo: um dos problemas atuais em visão computacional é segmentar uma imagem em regiões de *interesse* e *sem interesse* para futuros processamentos. As regiões de *interesse* podem ser definidas em função das propriedades físicas dos objetos, como variação de posição ou orientação, caracterizando a dinâmica dos objetos em movimento. Abordagens que permitem atacar este problema são baseadas na subtração de fundo, procedimento esse que separa os objetos em movimento do resto

da cena no campo visual de uma câmera (na maioria dos casos estática) através da determinação e adaptação de um modelo do fundo. Deste modo, os objetos em movimento são todos aqueles grupos de píxeis que são diferentes do modelo do fundo. Estes conjuntos de píxeis podem ser agrupados como objetos nas fases posteriores do processamento. Essencialmente, a subtração de fundo pode ser vista como um problema de classificação, sendo o objetivo classificar cada píxel de um quadro de uma sequência de vídeo ou como *primeiro plano* ou como *fundo*, onde *primeiro plano* significa que este píxel pertence a um objeto em movimento (de *interesse*) e *fundo* significa que o píxel pertence a uma região estática do quadro (*sem interesse*), estando portanto, bem representado no modelo do fundo;

- rastreamento de objetos: uma vez que os objetos em movimento foram detectados, o próximo passo é determinar as trajetórias percorridas por eles em relação a uma sequência de quadros do vídeo. Assim, para cada quadro é criada uma lista de objetos em movimento achados e, a partir dos quadros precedentes, é criada uma lista das trajetórias percorridas pelos objetos. Portanto, o problema consiste na atualização das trajetórias em função da detecção de objetos em movimento no quadro atual, sendo esta, essencialmente, uma tarefa de associação. Aqui, o objetivo é achar um casamento entre cada objeto detectado e as trajetória prévias já definidas. Essa operação de associação entre todo objeto em movimento e as trajetórias existentes permite estabelecer um único identificador para cada objeto presente no vídeo, o qual é um pré-requisito para outras tarefas de alto nível, como são classificação de objetos e sistemas de alerta em tempo real [22];

- classificação de objetos: dado um objeto na cena, o objetivo é classificá-lo, por exemplo, como um veículo ou como um pedestre. O problema de classificação apresenta vários desafios e tem como objetivo satisfazer aos seguintes requisitos: bom desempenho, funcionamento para diferentes localizações da câmera, capacidade de distinguir objetos semelhantes (como veículos e grupos de pedestres) ou classificá-los com informação incompleta (um objeto que acabou de entrar ou está saindo da cena), entre outros.

Essas etapas compõem o diagrama de fluxo de um sistema de vigilância automatizado, ilustrado na Figura 1.1. Observe-se que alguns sistemas podem não conter todas as etapas listadas e/ou conter etapas diferentes, mas, de maneira geral, o diagrama de fluxo segue esse padrão.

Neste trabalho é estudada em detalhe a etapa de subtração de fundo, já que o rendimento das etapas de rastreamento e classificação são dependentes do resultado da detecção dos objetos em movimento. Embora muitas técnicas de subtração de fundo sejam detalhadas

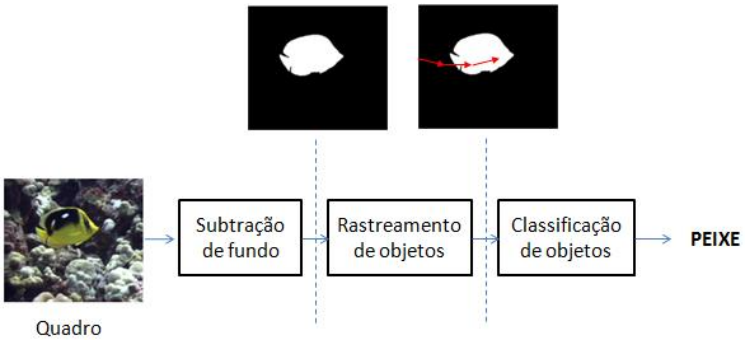


Figura 1.1: Diagrama de fluxo de um sistema de vigilância automatizado padrão.

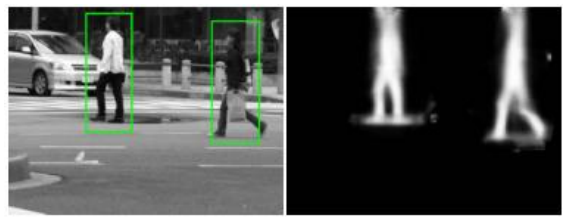
na literatura, modelar o fundo, especialmente para cenas ao ar livre, é um problema que oferece muitos desafios. Na literatura existem várias técnicas propostas para o modelamento do fundo. Isto ocorre, principalmente, porque nenhuma técnica é capaz de lidar com todos os desafios presentes na área.

Em relação à importância das técnicas de subtração de fundo para as etapas subsequentes de um sistema de vigilância automatizado, tem-se os seguintes exemplos: considerando-se o rastreamento de objetos, pode-se concentrar atenção unicamente nas áreas da cena definidas pelo primeiro plano [11][14][48]; analogamente, a classificação de objetos pode ser otimizada ao se restringir a busca dos objetos a classificar somente nas áreas do primeiro plano. Além disso, métodos de reconhecimento que trabalham com as formas definidas através das silhuetas do primeiro plano também estão presentes na literatura [42].

Ademais, técnicas de subtração de fundo são de interesse em diferentes aplicações de visão computacional, como detecção de pedestres [40] (ver Figura 1.2.a), detecção de tráfego nas rodovias [61](ver Figura 1.2.b) e rastreamento de indivíduos na multidão [41] (ver Figura 1.2.c). Em todas estas aplicações a subtração de fundo é a etapa inicial que permite determinar a presença dos objetos em movimento, porém não sua classe (um pedestre, veículos ou multidões). Assim, a subtração de fundo é um problema chave na área de visão computacional, possuindo diversas aplicações com potencial para melhorar a qualidade de vida da sociedade.

1.2 Caracterização do Problema

O problema que aparece nas técnicas de subtração de fundo pode ser caracterizado como: dada uma sequência de quadros, a partir de uma câmera fixa, determinar todos os objetos pertencentes ao primeiro plano. Neste ponto, surgem outros problemas, como:



(a)

(b)

(c)

Figura 1.2: Aplicações de visão computacional onde são usadas técnicas de subtração de fundo. Assim tem-se: (a) detecção de pedestres (figura obtida de [40]); (b) detecção de tráfego nas rodovias (figura obtida de [61]); (c) rastreamento de indivíduos na multidão (figura obtida de [41]). Nestas figuras a primeira coluna contém o quadro a analisar e na segunda coluna está a máscara binária gerada pela correspondente técnica de subtração de fundo usada em cada aplicação.

- a estimação e adaptação do modelo do fundo, tal que ele seja o menos sensível às mudanças dinâmicas da cena, como são as mudanças na iluminação, oscilações de movimento, entre outras. Todos estes problemas a serem tratados são referidos na literatura como os *desafios de uma técnica de subtração de fundo*, e são explicados em detalhes na seção 2.2 desta tese;

- a definição do que constitui uma diferença significativa do fundo em relação ao primeiro plano, ou seja, a detecção de variações, e, portanto a determinação dos objetos em movimento. Comumente, esta tarefa é efetuada por um processo de limiarização, sendo a principal dificuldade, fruto da inexatidão do modelo do fundo, a classificação (pelo limiar) de algumas áreas do fundo como primeiro plano, gerando a detecção de objetos em movimento inexistentes;

- a determinação do desempenho de uma técnica de subtração de fundo, não sendo essa uma tarefa trivial devido: (a) ao alto tempo de avaliação[1], fruto de se ter um grande espaço amostral[2]; (b) à complexidade na geração das amostras de referência utilizadas na comparação com os resultados da técnica analisada; (c) à falta de uma métrica que quantifique objetivamente o desempenho obtido por uma técnica de subtração de fundo.

Por estes motivos, e por ter aplicação real no mundo atual, o problema da subtração de fundo tem despertado o interesse de vários grupos de pesquisa durante os últimos anos, como é indicado nos estudos feitos sobre a evolução da área encontrados na literatura (ver [12][21][55][56]).

1.3 Objetivos

O objetivo geral desta Tese de Doutorado é implementar e testar um sistema para detecção de objetos em movimento numa sequência de vídeo, considerando-se uma cena capturada com uma câmera fixa.

O objetivo específico é a implementação e avaliação de uma técnica de subtração de fundo, para o que é necessário tratar os problemas de: (a) modelamento do fundo, em que serão

[1]Para ter uma ideia deste problema pode-se considerar a relação temporal teórica proposta em [25]: teoria : implementação : teste = 1 : 10 : 100.

[2]Os bancos de dados utilizados neste tipo de aplicação são compostos por vídeos, e cada quadro de um vídeo é considerado como uma amostra.

utilizadas distintas técnicas para o cálculo do modelo do fundo, permitindo assim explorar diferentes perspectivas de solução e como elas tratam os desafios da área; (*b*) detecção de variações, em que serão propostas técnicas que sejam capazes de operar no nível de cores e fazer a detecção de variações através de um limiar automatizado, com o intuito de dar uma maior capacidade de adaptação às técnicas através da sintonia contínua do limiar; (*c*) determinação do desempenho, para o que será proposta uma técnica para quantificar a exatidão obtida na detecção de objetos em movimento utilizando um banco de dados especializado para aplicações de monitoramento. Finalmente, também será utilizado um banco de dados especializado para analisar individualmente alguns dos desafios que deve tratar uma técnica de subtração de fundo.

1.4 Metodologia

A elaboração da Tese de Doutorado é feita de acordo com a metodologia:

1. Revisão da bibliografia relacionada as técnicas de subtração de fundo.
2. Implementação de cada um dos componentes de uma técnica de subtração de fundo levando-se em conta as técnicas de maior interesse presentes na literatura. Especificamente, (*a*) diversas técnicas são realizadas para a etapa de modelamento do fundo, e (*b*) duas técnicas são propostas para a etapa de detecção de variações.
3. Teste e avaliação dos algoritmos modificados em função de banco de dados e métricas especializadas para aplicações de vigilância. Especificamente, (*a*) é proposto um procedimento de avaliação que permite determinar o desempenho de uma técnica de subtração de fundo; (*b*) para testes são utilizados dois bancos de dados, (*i*) o primeiro próprio para o desenvolvimento e teste de um sistema de vigilância em espaço público, e (*ii*) o segundo especializado em analisar individualmente alguns dos desafios de que deve tratar uma técnica de subtração de fundo. Ambas situações possibilitam fazer uma avaliação completa das técnicas de subtração de fundo propostas.
4. Elaboração do documento final.

1.5 Organização da Tese

Este trabalho está organizado da seguinte maneira:

- No Capítulo 2 se detalham: os desafios a serem vencidos pelas técnicas de subtração de fundo; uma taxonomia das diferentes abordagens presentes na literatura; uma explicação de cada uma de suas partes constitutivas, e onde também são descritas as técnicas implementadas para cada uma delas, destacando-se o passo de detecção de variações, no qual são propostas duas abordagens baseadas na distância Euclidiana e numa operação de limiarização automática, as quais são validadas experimentalmente.

- No Capítulo 3 se apresentam: as perspectivas na avaliação das técnicas de subtração de fundo, e é proposto um procedimento que permite calcular uma métrica de desempenho, seja para cada vídeo ou para todo o banco de dados.

- No Capítulo 4 são apresentados os resultados experimentais para a avaliação das técnicas de subtração de fundo implementadas, usando-se dois bancos de dados especializados da área.

- No Capítulo 5 se apresentam as conclusões deste trabalho, junto com algumas propostas de trabalhos futuros.

- Na parte final da tese, além das Referências que foram consultadas para o desenvolvimento deste trabalho, são apresentados quatro Apêndices: no Apêndice A são apresentadas as tabelas relacionadas aos bancos de dados usados na parte experimental; no Apêndice B são apresentadas as tabelas relacionadas aos resultados experimentais do Capítulo 4; no Apêndice C é exposto o problema de rastreamento de objetos, e como o mesmo é solucionado através do filtro de Kalman, iniciando-se a explicação com o caso do rastreamento de um único objeto e logo analisando-se o caso geral de rastreamento de múltiplos objetos; e no Apêndice D são expostos os conceitos relacionados à teoria de cores, as conversões entre espaços de cores, e, por fim, é feita uma exposição das características da distância Euclidiana em função dos atributos que descrevem as cores.

Capítulo 2

Técnicas de Subtração de Fundo

2.1 Introdução

A área de sistemas de vigilância automatizados baseados em vídeo é, na atualidade, de grande interesse devido a suas implicações no campo da segurança como, por exemplo, vigilância de tráfego veicular e monitoramento de pessoas e suas atividades [13]. Tipicamente, estes sistemas têm uma ou várias câmeras fixas focadas em regiões de interesse, tais como pontos de controle em rodovias de alta velocidade, estacionamentos, prédios públicos ou residências. Em todos estes casos os sistemas de vigilância automatizados requerem algum mecanismo para detectar o movimento de objetos no campo de visão do sensor. Tal mecanismo serve como uma forma de focalizar a atenção. Uma vez que os objetos são focalizados, o processamento posterior de rastreamento é limitado na correspondente região da imagem.

A identificação de objetos em movimento é baseada nas técnicas de subtração de fundo, onde cada quadro do vídeo é comparado a um modelo de referência do fundo. Píxeis no quadro atual que apresentam uma variação significativa em relação ao fundo são considerados como objetos em movimento, e definidos como *píxeis do primeiro plano*. Assim, a subtração de fundo é uma importante tarefa em todo sistema de vigilância automatizado, e é objeto de estudo deste capitulo.

Sendo assim, o capítulo se inicia com uma descrição dos desafios encarados por toda técnica de subtração de fundo. Na segunda seção são relatadas as taxonomias de classificação das diferentes técnicas de subtração de fundo encontradas na literatura, e, considerando o nível de análise em que pode operar cada técnica, é efetuada uma apresentação das abordagens

mais representativas da área. Na terceira seção são indicadas as partes constitutivas de uma técnica de subtração de fundo. Na quarta seção são descritas três técnicas representativas da literatura para o modelamento do fundo, e também são descritas as técnicas implementadas para as outras etapas da técnica de subtração de fundo. Finalmente, na última seção, é apresentado um resumo do capítulo.

2.2 Desafios

Uma técnica de subtração de fundo tem que lidar com vários desafios dependentes das características próprias do ambiente/cenário considerado. Assim, na literatura, vários desafios são indicados (ver [7] para obter uma lista abrangente), descrevendo problemas vinculados à variação da iluminação numa cena, comportamentos conflitivos dos objetos do primeiro plano, problemas decorrentes de uma técnica específica, entre outros. Aqui se analisará essas questões seguindo as pautas adotadas em [9][17]. Por uma questão de clareza, é definido aqui: *a*) a detecção de falsos positivos como o caso onde uma região do fundo é identificada como uma entidade do primeiro plano; *b*) a detecção de falsos negativos como o caso onde uma entidade do primeiro plano é identificada como parte do fundo (ambos conceitos serão explicados com maior detalhe no capítulo 3). A seguir, são explicados os desafios de maior importância relatados em [9][17].

- Mudanças graduais da iluminação: é desejável que o modelo do fundo se adapte às mudanças graduais do ambiente. Por exemplo, em cenas externas, a intensidade da luz varia durante o dia, e tal variação provocará uma mudança global da aparência do fundo;

- mudanças repentinas da iluminação: improvistas mudanças da iluminação que não são cobertas pelo modelo do fundo afetam fortemente a sua aparência, e originam detecção de falsos positivos. Elas ocorrem, global ou localmente, por exemplo, num cenário exterior, quando a iluminação numa rua é ligada; num cenário interior, quando a iluminação num quarto varia devido a diferentes fontes de luz, etc.;

- fundo dinâmico: algumas partes do cenário podem apresentar um certo grau de movimento, devendo estas ser consideradas como parte do fundo. Tal movimento pode ser periódico ou irregular, como, por exemplo, no caso das luzes de um semáforo ou das folhas das árvores movidas pelo vento. Tal comportamento oscilante infringe a característica principal do fundo, ou seja, de ser estático, fazendo com que essas partes do

cenário sejam rotuladas como um objeto do primeiro plano, produzindo, assim, uma detecção de falsos positivos;

- camuflagem: intencionalmente ou não, alguns objetos do primeiro plano podem diferir pouco da aparência do fundo, tornando difícil uma correta classificação, produzindo uma detecção de falsos negativos;

- primeiro plano adormecido: um objeto do primeiro plano que se torna imóvel, ao passar o tempo, passará a ser visto como parte do fundo, este problema ocorre muitas vezes, em situações de vigilância. Por exemplo, uma estrada de cruzamento com semáforos, onde os veículos se deslocam e param. Em tal caso, os veículos não devem ser rotulados como parte do fundo ao passar o tempo;

- abertura do primeiro plano: quando um objeto do primeiro plano de cores homogêneos se desloca, variações entre quadros na cor dos píxeis internos do objeto em movimento podem não ser detectadas. Assim, todo o objeto pode ser excluído do primeiro plano, causando uma detecção de falsos negativos. Por exemplo, um veículo pequeno pintado com uma única cor, num ambiente de iluminação constante, pode não ser detectado;

- sombras: sombras projetadas por objetos do primeiro plano muitas vezes complicam as etapas de processamento posteriores à subtração de fundo. Por exemplo, a sobreposição de sombras sobre regiões do primeiro plano dificultará sua classificação. Por isso, é preferível ignorar estas regiões irrelevantes (as sombras são simplesmente alterações irregulares e locais na iluminação da cena, e assim não devem ser consideradas entidades do primeiro plano);

- reflexões: a cena pode refletir instâncias do primeiro plano, devido a superfícies molhadas ou refletoras, como uma estrada, janelas, vidros, etc, e essas entidades refletidas não devem ser classificadas como parte do primeiro plano. Por exemplo, uma estrada molhada com o sol brilhando provocara a reflexão dos veículos que passam;

- *bootstrapping*: na etapa de treinamento, se não estiverem disponíveis quadros livres de objetos do primeiro plano, o modelo do fundo terá que ser inicializado utilizando uma estratégia de inicialização que tenha a capacidade de criar adaptativamente um modelo adequado do fundo, mesmo se os dados de treinamento contiverem objetos do primeiro plano;

- ruído de vídeo: o sinal de vídeo contém ruído. Assim, técnicas de subtração de fundo para vídeo de vigilância têm que lidar com tais sinais afetadas por diferentes tipos de ruído, como o ruído do sensor ou artefatos da compressão.

2.3 Estado da Arte

Considerando o uso de uma câmera fixa, uma técnica de subtração de fundo terá como entrada uma matriz de píxeis, que representa o quadro adquirido pela câmara (seja na escala de cinza ou a cores) e gera como saída uma máscara binária que indica os píxeis pertencentes ao primeiro plano. Assim, a subtração de fundo consiste em comparar o quadro atual com o modelo do fundo, definindo os píxeis do primeiro plano como aqueles que não podem ser explicados pelo modelo do fundo. Segundo esta perspectiva, diferentes classificações das técnicas de subtração de fundo, considerando uma câmera fixa, têm sido propostas na literatura.

- Em [12][53], as técnicas são divididas em recursivas e não-recursivas, sendo (a) as técnicas recursivas as que mantêm um único modelo do fundo, o qual é atualizado usando cada novo quadro do vídeo e (b) as técnicas não-recursivas aquelas que mantêm um *buffer* que armazena uma certa quantidade de quadros do vídeo e estima um modelo do fundo baseado apenas nas propriedades estatísticas de tais quadros.

- Em [45], as técnicas são divididas em preditivas e não-preditivas, sendo (a) as técnicas preditivas as que modelam a cena como uma série temporal e desenvolvem um modelo dinâmico para avaliar a entrada atual baseada nas observações anteriores e (b) as técnicas não-preditivas as que negligenciam a ordem das observações de entrada e criam uma representação probabilística das observações num píxel particular.

- Em [17][67] as técnicas são classificadas como orientadas a píxeis, regiões e quadros, onde (a) as técnicas orientadas a píxeis realizam a discriminação entre o primeiro plano e o fundo considerando cada píxel como um processo independente, sendo essa a abordagem mais utilizada hoje em dia, devido ao baixo esforço computacional requerido; (b) as técnicas orientadas a regiões são aquelas que relaxam a suposição de independência dos píxeis, permitindo, assim, considerar relações espaciais locais entre píxeis, com a intenção de minimizar a detecção de falsos positivos, considerando a informação local de alta ordem (como as bordas), complementando assim a análise por píxel; (c) as técnicas orientadas a quadros aquelas que em comparação com as técnicas orientadas a regiões, consideram as relações espaciais não de uma vizinhança, mas de todo o quadro, baseando-se num procedimento aplicado sobre um quadro inteiro, como forma de tratar os problemas globais como as mudanças repentinas da iluminação. Muitas vezes estas técnicas são utilizadas como suporte das abordagens anteriores.

Das três taxonomias, a proposta em [17][67], que considera diferentes níveis espaciais (por píxel, por região e por quadro), é a que apresenta várias vantagens em relação às demais, tais como: *a*) cada nível tomado isoladamente tem suas próprias vantagens e seus problemas principais bem definidos, o que permite especificar os problemas que uma determinada técnica (classificada num determinado nível) pode resolver ou não[1]; *b*) este enfoque permite considerar soluções multi-camadas integradas, onde a discriminação entre o primeiro plano e o fundo é feita no nível de píxeis, sendo a classificação refinada ao nível de regiões ou quadros. Pelas considerações já expostas, neste trabalho é apresentado um resumo do estado da arte das técnicas de subtração de fundo considerando a taxonomia que assume diferentes níveis espaciais.

2.3.1 Técnicas Orientadas a Píxeis

Estas técnicas têm como suposição principal que cada píxel de um quadro é um processo estocástico independente. Portanto, cada processo pode ser modelado estatisticamente, parametricamente[2] ou não parametricamente[3]. Tal processo estocástico é denominado na literatura da área como um processo de um píxel, e é definido como: os valores que assume um píxel numa localização particular de uma imagem considerando um conjunto de quadros pertencentes a uma sequência de vídeo [63] (este conceito será explicado com maior detalhe na subseção 2.5.1). Sendo assim, as técnicas de subtração de fundo que fazem uso desta suposição podem ser agrupadas em duas categorias, técnicas unimodais e multimodais, as quais são detalhadas a seguir.

Técnicas Unimodais. Essas técnicas pressupõem que as características que definem o valor do fundo de um píxel são representadas estatisticamente por uma distribuição unimodal (não necessariamente centrada). A desvantagem destes modelos é que apenas o fundo unimodal é levado em consideração, ignorando assim todas as situações em que a multimodalidade do fundo está presente. A virtude é que a suposição de unimodalidade permite elaborar técnicas computacionalmente eficientes para aplicações *on-line*. As técnicas uni-

[1]Por exemplo, assumir que a evolução temporal de cada píxel é um processo independente (assim direcionado ao nível de píxeis), e não levar em conta as informações observadas em outros píxeis (não é feita operação alguma ao nível de regiões ou nível de quadros), são pautas inadequadas para tratar o problema das mudanças repentinas da iluminação.

[2]Uma forma particular da função de densidade é suposta, e somente os parâmetros dela necessitam ser estimados.

[3]Não se consideram suposições sobre a função de densidade. Entretanto, é necessário uma quantidade importante de amostras na parte de treinamento.

modais mais representativas são descritas a seguir:

Considerando uma representação paramétrica:

- Em [75], é proposto o sistema de vigilância *Pfinder*, onde os píxeis de cada quadro são representados no espaço de cores *Luminance* (Y), *blue-luminance* (U), *red-luminance* (V) (YUV), e o modelo do fundo é definido como um valor médio atualizado recursivamente. Esta técnica trabalha segundo a suposição de que a cena é menos dinâmica que os objetos do primeiro plano. Embora o *Pfinder* possa lidar com mudanças graduais da iluminação, ele falha quando o fundo sofre mudanças repentinas da iluminação.
- Em [33], uma técnica semelhante à proposta em [75] é apresentada. Aqui, cada processo de um píxel é definido por uma distribuição normal multivariável, e a estimação dos parâmetros da normal é realizada através da maximização da verossimilhança, obtendo-se um modelo do fundo representado como uma média móvel gaussiana.

Considerando uma representação não paramétrica:

- Em [44], cada píxel do modelo do fundo é determinado através do cálculo do valor da mediana, obtido a partir de um *buffer* que contém um conjunto de quadros das capturas anteriores, atingindo uma elevada eficiência computacional e robustez ao ruído. No entanto, uma restrição desta técnica é que ela não modela a variância associada a um valor do fundo.
- Em [30][32], é proposto o sistema de vigilância W^4, onde um píxel é rotulado como primeiro plano se o valor satisfaz um conjunto de desigualdades. Os parâmetros das desigualdades representam a mínima, máxima e a maior diferença absoluta entre quadros que contêm a cena do fundo. Estes parâmetros são inicialmente calculados a partir dos primeiros segundos de uma sequência de vídeo, e são periodicamente atualizados considerando as partes da cena que não contêm objetos do primeiro plano.
- Em [43][83] se considera que a função de densidade de cada processo de um píxel pode ser aproximada por um histograma, definindo o modelo do fundo como o valor da moda do histograma, já que, ao considerar uma sequência de vídeo capturada com uma câmera fixa, é de se esperar que a intensidade com o maior número de ocorrências represente um píxel do modelo do fundo.

Técnicas Multimodais. Aqui é suposto que o valor do fundo de um píxel é representado estatisticamente por uma distribuição multimodal. Neste caso, estamos diante do problema de ter um fundo dinâmico, por exemplo, com pequenos movimentos repetitivos, como a imagem do mar, onde cada píxel tem pelo menos uma distribuição bimodal de cores, com destaque para o mar e as reflexões do sol.

- Em [26] é apresentada uma das primeiras técnicas que lidam com a multimodalidade, onde uma mistura de distribuições normais multivariáveis (*Mixture Of Gaussians* (MOG)) é ajustada gradualmente para cada píxel. O cenário de aplicação é o monitoramento de uma estrada, e é proposto um conjunto de heurísticas para rotular os píxeis que representam a estrada, as sombras e os carros.
- Em [63], o processo de um píxel também é estatisticamente modelado usando uma mistura de Gaussianas, onde seus parâmetros são determinados através da técnica de ajuste conhecida como *expectation-maximization*[4]. Esta técnica é amplamente conhecida na literatura (em [7] é realizado um sumário de todos os trabalhos vinculados a esta proposta).

Considerando uma representação não paramétrica:

- Em [20][21], é desenvolvida uma técnica que modela o processo de um píxel através de um método de estimação da densidade baseada em *kernels* (*Kernel Density Estimation* (KDE)).

Embora as técnicas orientadas a píxeis sejam amplamente utilizadas por seu razoável compromisso entre a precisão e velocidade (em termos computacionais), estas técnicas apresentam algumas desvantagens, principalmente devido à suposição de independência entre píxeis. Assim, qualquer situação que precisa de uma visão global da cena, a fim de realizar uma rotulação correta do fundo, não é possível, causando, em geral, detecção de falsos positivos.

Quanto ao custo computacional das técnicas orientadas a píxeis, em [55] é feita uma análise, sendo levados em consideração a velocidade e o uso da memória de alguns algoritmos

[4]A técnica *expectation-maximization* é usada em estatística para a estimação de parâmetros usando a maximização da verossimilhança, quando se tem modelos probabilísticos que dependem de variáveis não observáveis. No caso de misturas finitas, as variáveis não observáveis ou ocultas constituem um conjunto de vetores binários associados a cada uma das observações, indicando qual componente da mistura produz cada observação.

amplamente utilizados. Essencialmente, as técnicas unimodais são geralmente mais rápidas, enquanto as técnicas multimodais e não paramétricas apresentam maior complexidade. Em relação ao uso de memória, abordagens não-paramétricas são as mais exigentes, porque precisam recolher para cada píxel uma estatística sobre os valores do passado.

2.3.2 Técnicas Orientadas a Regiões

Uma análise no nível de regiões tenta modelar as relações entre píxeis vizinhos de uma imagem. Sendo assim, estas técnicas definem uma vizinhança local em torno de cada píxel, tal que um conjunto de operações são efetuadas sobre tal vizinhança. As técnicas de subtração de fundo que fazem uso desta suposição podem ser agrupadas considerando o tipo de informação extraída da vizinhança, podendo ser, as intensidades dos píxeis, a informação de textura ou a informação de borda.

Técnicas Baseadas na Intensidade dos Píxeis. Estas técnicas utilizam diretamente as intensidades dos píxeis das vizinhanças.

- A técnica apresentada em [20], que inicialmente foi classificada como uma técnica orientada a píxeis, também pode ser considerada como uma técnica orientada a regiões, já que contém uma etapa que incorpora um procedimento inerentemente orientado a regiões. Em tal etapa, os valores dos píxeis que podem ser modelados pelas distribuições de píxeis vizinhos são reclassificados como fundo, permitindo uma maior robustez contra o problema de um fundo dinâmico.
- Em [45], é proposta uma técnica mais avançada, utilizando uma estimação da densidade baseada em *kernels*. Aqui, os *kernels* baseados nos histogramas da intensidade dos píxeis são gerados em relação à vizinhança centrada em cada píxel, e não exclusivamente com os valores anteriores deste píxel tomados ao longo da sequência de vídeo.

Técnicas Baseadas em Características das Texturas e Bordas. Estas técnicas utilizam a informação estrutural de uma imagem, como as bordas ou as características de textura.

- Em [49], cada quadro de uma sequência de vídeo é dividido numa grade de vizinhanças

quadradas sobrepostas, tal que *kernels* baseados nos histogramas dos gradientes[5] são gerados para cada uma das vizinhanças. Os autores indicam que esta proposta é robusta às mudanças da iluminação.

- Em [34] é apresentada uma técnica baseada nas características discriminativas das texturas, representadas pelo histograma dos padrões locais binários (*Local Binary Patterns* (LBP)) [6]. Os padrões locais binários têm várias propriedades que são de utilidade no modelamento do fundo. Por serem baseados numa operação diferencial limiarizada, são robustos às mudanças na iluminação, podendo assim contornar o problema das sombras, quando as áreas das sombras numa imagem não são muito pequenas e o raio do círculo escolhido para o cálculo dos padrões locais binários é pequeno. Entretanto, os padrões não podem detectar mudanças em regiões uniformes suficientemente grandes, se o primeiro plano for também uniforme.

- Em [78] é proposta uma técnica que utiliza tanto uma abordagem orientada a píxeis como a regiões. Assim, a informação de cor associada a um píxel é definida em um espaço fotométrico invariante, e a informação estrutural da região é representada através do histograma dos padrões locais binários. Este modelo é particularmente robusto para sombras.

- Em [37] é apresentada uma técnica que define, para cada píxel, um vetor de 13 características, dadas por: os valores de intensidade para os canais de cor R,G e B do píxel, a magnitude e a orientação do gradiente do píxel, e a resposta da vizinhança do píxel em relação à oito características de Haar (que representa a informação de textura). Cada uma destas 13 características são explicitamente modeladas como misturas Gaussianas adaptativas no tempo, e utilizando um método de fusão de informação é obtida uma medida que permite, através de uma operação de limiarização, determinar se o píxel é parte do fundo ou não.

Em relação ao custo computacional, as técnicas orientadas a regiões apresentam um maior custo, tanto no espaço de memória usado como no tempo de processamento, em comparação

[5] A informação do gradiente de uma imagem tem uma maior capacidade discriminativa nos contornos dos objetos de uma cena, mas não fornece uma clara diferença para objetos grandes sem textura, por exemplo, um ônibus branco numa estrada.

[6] Os LBP foram descritos inicialmente em [51], sendo um tipo de característica para texturas invariantes às mudanças da intensidade de um píxel numa imagem em escala de cinza, pudendo ser estendidos para imagens a cores ao calculá-los separadamente para cada canal de cor (considerando o espaço de cores *Red-Green-Blue* (RGB)). Os LBP são padrões binários de determinado comprimento. Assim, para calculá-los é necessário definir uma vizinhança circular sobre cada píxel, tal que cada valor binário do padrão é 1 se a diferença entre a intensidade do píxel central e a intensidade de um píxel particular que está sobre a vizinhança circular é maior que um limiar.

com as técnicas orientadas a píxeis. De qualquer forma, a maioria dos trabalhos reivindicam rendimentos em tempo real.

2.3.3 Técnicas Orientadas a Quadros

Estas técnicas estendem a área local de refinamento da análise por píxel para todo o quadro.

- Em [65], é proposto usar uma versão *on-line* de um modelo oculto de Markov (*Hidden Markov Model* (HMM)) para representar adequadamente as mudanças de iluminação de uma cena. Embora os resultados sejam promissores, é importante notar que a técnica não tem sido avaliada em sua versão *on-line*; ainda mais, a mudança na iluminação deve ser global e pré-classificada numa seção de treinamento.
- Em [52], é proposta uma técnica que captura as correlações espaciais, aplicando uma análise de componentes principais (*Principal Component Analysis* (PCA)), para um conjunto de quadros que não contêm objetos do primeiro plano. Isto resulta em um conjunto de funções base, onde as primeiras d são necessárias para capturar as características primárias da cena observada. Um novo quadro pode ser projetado no autoespaço definido por estas d funções base e então voltar a projetar no espaço original da imagem. Já que as funções base apenas modelam a parte estática da cena quando não estão presentes objetos do primeiro plano, a imagem projetada de volta não conterá objetos do primeiro plano. Como tal, pode ser utilizado como um modelo do fundo. A principal limitação dessa abordagem reside na hipótese original da ausência de objetos do primeiro plano para calcular as funções base, a qual nem sempre é possível. Além disso, também não é claro como as funções base podem ser atualizadas ao longo do tempo, se objetos do primeiro plano vão estar presentes na cena.

No que diz respeito ao custo computacional, as técnicas orientadas a quadros usualmente consideram uma etapa de treinamento e uma etapa de classificação. A etapa de treinamento é levada a cabo de uma maneira *off-line*, enquanto que a etapa de classificação é bastante adequada para uso em tempo real.

2.3.4 Técnicas em Várias Etapas

Estas técnicas não podem ser incluídas em qualquer uma das categorias vistas anteriormente, e em geral são implementadas pela composição de um conjunto de diferentes etapas.

- Em [67], é apresentada uma técnica de três etapas, que opera no nível de píxel, regiões e quadros. Assim, tem-se: (a) no nível de píxeis, dois modelos do fundo são determinados para cada píxel de forma independente, ambos baseados em filtros de Wiener de um passo, onde os valores (passados) para cada filtro (modelo) são para o primeiro filtro, os valores observados para cada píxel (valor atual do píxel), e, para o segundo filtro, os valores preditos pelo primeiro filtro (valor predito do píxel). Uma verificação dupla contra esses dois modelos, é feita a cada passo: o valor do píxel atual é considerado como fundo, se este está dentro de um nível de tolerância em relação ao erro quadrático de predição esperado calculado utilizando os dois modelos; (b) no nível da regiões, onde é aplicado um algoritmo de crescimento de regiões que, essencialmente, fecha os possíveis furos (falso negativos) no primeiro plano no caso em que os valores da intensidade dos píxeis nas localizações dos falsos negativos (furos) sejam similares aos valores dos píxeis circundantes do primeiro plano; (c) no nível de quadros, onde é mantido um conjunto de modelos do fundo determinados no nível de píxeis na etapa de treinamento utilizando o algoritmo *k-means*, e é selecionado aquele que minimiza a quantidade de píxeis do primeiro plano.

- Em [16] uma técnica similar aquela em [67] foi apresentada, onde o problema das mudanças repentinas da iluminação, seja local ou global, é levado em conta. A técnica reside em uma segmentação do fundo, que segmenta partes do fundo onde o aspecto cromático é homogêneo e evolui de forma uniforme. Quando uma região do fundo de repente muda sua aparência, considera-se como uma evolução do fundo, em vez do surgimento do primeiro plano. A técnica funciona bem quando as regiões segmentadas no fundo são poucas e grandes. Por outro lado, os desempenhos são pobres quando o fundo é sobre segmentado, o que, em geral, ocorre para cenas externas.

- Em [77], a cena é particionada usando uma estrutura *quadtree*, caracterizada por um nó pai e quatro nós filho, tal que, para cada partição são construídos filtros *Minimal Average Correlation Energy* (MACE). Começando com uma partição de 32×32 píxeis, até chegar a vizinhanças de 4×4 píxeis. A técnica proposta visa evitar falsos positivos, uma vez que quando um filtro detecta a presença do primeiro plano em mais de 50% da sua área, a análise é propagada para os 4 filhos pertencentes ao nível inferior, e, por sua vez, para a vizinhança 4-conectada de cada um dos filhos. Quando a análise atinge o nível mais baixo (4×4) e o primeiro plano ainda não é descoberto, o conjunto

de píxeis relacionados são marcados como primeiro plano. Cada filtro que modela uma partição do fundo é atualizado, a fim de lidar com uma mudança lenta do fundo. A técnica é lenta, e uma implementação em tempo real não é apresentada pelos autores, devido à necessidade de calcular os coeficientes dos filtros para cada uma das partições.

- Em [72], é feita uma estimação não paramétrica no nível de píxeis do primeiro plano, gerando-se uma máscara binária. Sobre tal máscara é aplicada uma série de operações morfológicas, a fim de resolver um conjunto de problemas comuns da subtração de fundo. Estas operações avaliam o comportamento conjunto de valores de píxeis semelhantes e próximos, por análise de componentes conectados. Deste modo, se vários píxeis são rotulados como primeiro plano, formando uma área conectada com possíveis furos na parte interna, os furos podem ser preenchidos. Se esta área é muito grande, se considera que o possível objeto do primeiro plano é causado por uma mudança rápida e global do fundo, e ele é rotulado como fundo.

Todas as técnicas baseadas em várias etapas exigem um elevado custo computacional, se comparadas com as categorias vistas anteriormente. De qualquer forma, em todos os trabalhos acima referidos as técnicas em várias etapas são reivindicadas para funcionar num ambiente em tempo real.

2.4 Arquitetura Padrão

Uma técnica de subtração de fundo trabalha tipicamente num ciclo de operação *on-line*, que geralmente é composto por duas fases: *a*) inicialização do fundo, onde o modelo do fundo é inicializado; *b*) atualização do fundo, onde os parâmetros que regulam o fundo têm que ser atualizados. Considerando tais etapas, a maioria das técnicas de subtração de fundo segue um diagrama de fluxo simples, definido inicialmente em [12] e apresentado na Figura 2.1. Nele, podem-se observar quatro passos principais:

- pré-processamento, que consiste de uma coleção de tarefas de processamento de imagens simples, aplicadas aos quadros que serão processados nos passos subsequentes;
- modelamento do fundo, que usa um novo quadro para calcular e atualizar o modelo do fundo. Este modelo do fundo fornece uma descrição estatística do fundo da cena+;
- detecção de variações, que também conhecido como detecção do primeiro plano, identifica píxeis nos quadros que não podem ser explicados pelo modelo do fundo. O

resultado deste passo é uma máscara binária, denominada máscara do primeiro plano, onde as regiões do fundo são rotuladas com 0 e as regiões do primeiro plano são rotuladas com 1;

- pós-processamento, que também conhecido como validação dos dados, examina a possível máscara do primeiro plano com a intenção de eliminar aqueles píxeis que não correspondem aos objetos em movimento, tendo como resultado a máscara do primeiro plano final.

Cabe indicar que existem técnicas que fusionam os passos de modelamento do fundo e detecção de variações, não gerando diretamente um modelo do fundo e sim uma máscara binária a ser processada. Na próxima seção são descritas as técnicas implementadas na literatura, considerando a arquitetura apresentada na Figura 2.1. No que resta deste capítulo a seguinte notação é utilizada, para um vídeo colorido (3 componentes) capturado com resolução $N_{\text{fil}} \times N_{\text{col}}$ píxeis por quadro:

- t: indica o índice do quadro atual em relação ao conjunto de quadros de uma sequência de vídeo;
- $\mathbf{I}(t)$: quadro no instante t de uma sequência de vídeo a cores, onde $\mathbf{I}(t) \in [0, 255]^{N_{\text{fil}} \times N_{\text{col}} \times 3}$;
- $\mathbf{B}(t)$: modelo do fundo a cores, calculado/atualizado no instante t, onde $\mathbf{B}(t) \in [0, 255]^{N_{\text{fil}} \times N_{\text{col}} \times 3}$;
- $\mathbf{M}(t)$: máscara do primeiro plano calculada no instante t, onde $\mathbf{M}(t) \in [0, 1]^{N_{\text{fil}} \times N_{\text{col}}}$.

Por motivos de simplificação notacional a localização do píxel (x_l, y_l) é unicamente representada com o sub-índice l. Assim, tem-se que $l = 1, \cdots, N_{\text{fil}} \times N_{\text{col}}$, onde:

- $\mathbf{I}_l(t)$: é o valor de um píxel na localização (x_l, y_l) no quadro $\mathbf{I}(t)$, onde $\mathbf{I}_l(t) \in [0, 255]^3$;
- $\mathbf{B}_l(t)$: é o valor de um píxel na localização (x_l, y_l) no modelo do fundo $\mathbf{B}(t)$, onde $\mathbf{B}_l(t) \in [0, 255]^3$;
- $\mathbf{M}_l(t)$: é o valor de um rótulo na localização (x_l, y_l) na máscara do primeiro plano $\mathbf{M}(t)$, onde $\mathbf{M}_l(t) \in [0, 1]$.

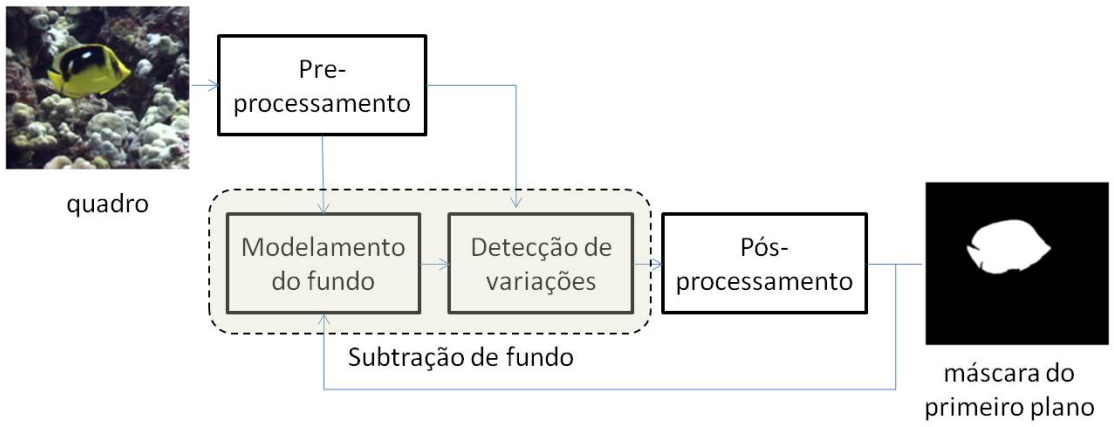


Figura 2.1: Diagrama de fluxo de uma técnica de subtração de fundo genérica [12].

2.5 Modelamento do Fundo

Considerando o compromisso apenas razoável entre precisão e velocidade indicado na literatura, em relação às outras abordagens orientadas a regiões, quadros ou em várias etapas, foram implementadas as técnicas orientadas a píxeis mais representativas da área. Dentre elas, tem-se: (a) baseada numa média móvel gaussiana, proposta em [33][75]; (b) baseada na mistura de Gaussianas, proposta em [63]; (c) baseada no histograma, proposta em [43][83].

A seguir, são detalhadas as implementações para cada uma destas abordagens, porém, por sua relevância, é explicado com maior detalhe o conceito de processo de um píxel.

2.5.1 O processo de um Píxel

Um processo estocástico é um modelo matemático de um experimento aleatório[7] que transcorre no tempo, gerando uma sequência de valores numéricos, onde cada valor numérico na sequência é modelado por uma variável aleatória[8]. Ou seja, um processo estocástico é simplesmente uma sequência (finita ou infinita) de variáveis aleatórias.

De uma maneira formal, seja um experimento aleatório com espaço amostral[9] $\mathcal{S}$. Para todo resultado $s \in \mathcal{S}$, é definida uma função temporal segundo a regra $X(\tau, s)$, tal que $\tau \in \mathcal{T}$,

[7]Um experimento aleatório é todo aquele cujos resultados não podem ser previstos (com certeza) antes da execução do mesmo.

[8]Variável aleatória é uma função que associa elementos do espaço amostral a valores numéricos.

[9]O espaço amostral de um experimento aleatório é o conjunto de todos os resultados possíveis desse experimento.

sendo $\mathcal{T}$ o conjunto de índices. Chama-se de processo aleatório ou processo estocástico a família indexada de variáveis aleatórias $\mathcal{X} = \{X(\tau, s), \tau \in \mathcal{T}, s \in \mathcal{S}\}$. Assim: *a*) para um $s \in \mathcal{S}$ fixo, o gráfico da função $X(\tau, s)$ versus τ é uma realização do processo aleatório; *b*) para um $\tau \in \mathcal{T}$ fixo, $X(\tau, s)$ é uma variável aleatória. No caso de considerar uma indexação discreta, o processo estocástico é uma coleção de variáveis aleatórias conhecidas como uma série temporal.

Levando-se em conta os conceitos apresentados, o processo de um píxel é definido como uma série temporal vinculada a cada píxel de uma sequência de vídeo. Assim, para imagens a cores a realização de um processo de um píxel na localização (x_l, y_l) de uma determinada sequência de vídeo s, é denotada como $\mathcal{I}_l(\tau, s)$ e definida pela equação

$$\begin{aligned} \mathcal{I}_l(\tau, s) &= \left\{ \begin{bmatrix} I_R(x_l, y_l, \tau) \\ I_G(x_l, y_l, \tau) \\ I_B(x_l, y_l, \tau) \end{bmatrix} ; \tau \in \mathbb{N} \right\} \\ &= \{\mathbf{I}_l(0), \mathbf{I}_l(1), \mathbf{I}_l(2), \cdots, \mathbf{I}_l(t-T), \mathbf{I}_l(t-(T-1)), \cdots, \mathbf{I}_l(t), \cdots\}, \end{aligned} \tag{2.1}$$

onde as variáveis aleatórias $\{\mathbf{I}_l(\tau)\}_{\tau \in \mathbb{N}}$ que constituem o processo são consideradas independentes[10]. Da realização do processo de um píxel $\mathcal{I}_l(\tau, s)$, só uma parte é observada, e em função a essa parte são determinadas as estatísticas relacionadas a $\mathcal{I}_l(\tau, s)$[11]. Considerando que a parte observada corresponde aos últimos T quadros em relação ao quadro atual t (ver Figura 2.2), a série temporal observada vinculada ao processo de um píxel $\mathcal{I}_l(\tau, s)$ é definida pela equação

$$\mathcal{I}_l(t) = \{\mathbf{I}_l(t-T), \mathbf{I}_l(t-(T-1)), \cdots, \mathbf{I}_l(t)\}. \tag{2.2}$$

Sumarizando, uma sequência de vídeo é constituída por um conjunto de realizações de processos de píxeis, tantas realizações quanto são os píxeis de um quadro. As estatísticas de cada uma destas realizações são estudadas através de suas correspondentes séries temporais observadas. Tais séries, dependendo da natureza da cena, podem apresentar um comportamento não estacionário (suas estatísticas variam ao longo do tempo) devido a: mudanças graduais da iluminação, objetos que podem diferir pouco da aparência do fundo, presença

[10] Independência entre variáveis aleatórias significa que a partir do resultado de uma delas não é possível inferir nenhuma conclusão sobre a outra.

[11] Não implicando a estacionariedade do processo.

de sombras, ruído do sistema de aquisição, entre outros. Portanto, pode-se entender o problema da subtração de fundo como a elaboração de técnicas que modelam o comportamento de cada uma destas séries temporais, tal que, a partir do modelo, é estabelecido se o píxel pertence ao fundo ou ao primeiro plano.

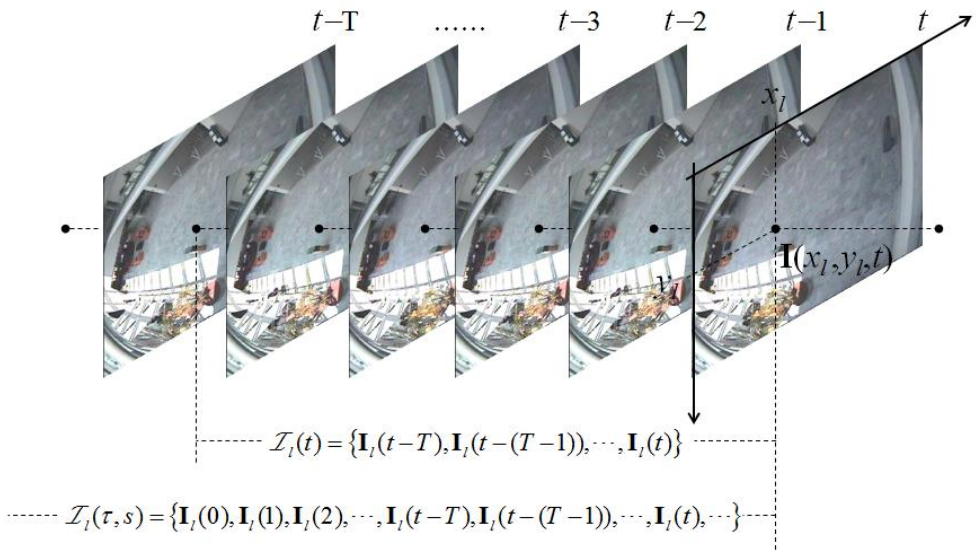


Figura 2.2: Relação entre a série temporal observada $\mathcal{I}_l(t)$ vinculada ao processo de um píxel $\mathcal{I}_l(\tau, s)$.

2.5.2 Modelo do Fundo Baseado na Média Móvel Gaussiana

Em [33][75] foi proposto um modelo do fundo baseado numa distribuição normal multivariável, tal como é definido em

$$\begin{aligned} p(\mathbf{I}_l(t)|\mathbf{B}_l(t), \sigma_l(t)) &= \mathcal{N}\left(\mathbf{B}_l(t), \sigma_l^2(t)\mathbf{I}\right) \\ &= \frac{1}{\sqrt{(2\pi)^3\sigma_l^3(t)}} \exp\left(-\frac{||\mathbf{I}_l(t) - \mathbf{B}_l(t)||^2}{2\sigma_l^2(t)}\right), \end{aligned} \tag{2.3}$$

onde o parâmetro $\sigma_l^2(t)$ é a componente de variância da normal multivariável no quadro t (assumindo que o valor dos píxeis nos canais das cores vermelha, verde e azul são independentes e têm a mesma variância). Um método padrão como a maximização da verossimilhança da série temporal observada $\mathcal{I}_l(t)$ é usado para calcular os valores de $\mathbf{B}_l(t)$ e $\sigma_l(t)$, através de

$$\left\{\hat{\mathbf{B}}_l(t), \hat{\sigma}_l(t)\right\} = \underset{\mathbf{B}_l(t),\sigma_l(t)}{\operatorname{argmax}} \left\{p(\mathcal{I}_l(t)|\mathbf{B}_l(t), \sigma_l(t))\right\}. \tag{2.4}$$

A solução sequencial para $\mathbf{B}_l(t)$ que se ajusta com a Equação (2.4) é o vetor médio das observações de $\mathcal{I}_l(t)$, calculado através da média móvel definida por

$$\hat{\mathbf{B}}_l(t) = \frac{1}{t}\sum_{k=1}^{t} \mathbf{I}_l(k) = \hat{\mathbf{B}}_l(t-1) + \frac{1}{t}(\mathbf{I}_l(t) - \hat{\mathbf{B}}_l(t-1)). \tag{2.5}$$

Em [38] a atualização do modelo do fundo é feita através de uma média móvel gaussiana, de modo a permitir uma suavização do modelo ao longo do tempo. Assim, tem-se

$$\hat{\mathbf{B}}_l(t) = \hat{\mathbf{B}}_l(t-1) + \alpha_{svt}(\mathbf{I}_l(t) - \hat{\mathbf{B}}_l(t-1)), \tag{2.6}$$

onde $\alpha_{svt} \in [0, 1]$ é o coeficiente de suavização. Valores de α_{svt} próximos de 1 têm um menor efeito de suavização e dão maior peso às mudanças recentes na cena[12], enquanto que os valores de α_{svt} próximos de 0 têm um maior efeito de suavização e são menos sensíveis às alterações recentes da cena. A média móvel gaussiana não requer um mínimo número de observações a serem feitas antes de começar a produzir resultados. Na prática, no entanto, uma *boa média* não será estimada antes de que várias amostras sejam ponderadas. Por exemplo, um sinal constante levará aproximadamente $3/\alpha_{svt}$ iterações para atingir 95% do valor real.

Para evitar que o modelo do fundo estimado através da média móvel seja perturbado com dados atípicos[13], é crucial selecionar cuidadosamente os píxeis a serem utilizados na estimação. A solução é que somente aqueles píxeis classificados como fundo no instante $t-1$ sejam contabilizados na estimação do $\hat{\mathbf{B}}_l(t)$ através da Equação (2.6). Em [33] são considerados outros casos particulares na estimação do modelo do fundo: (a) se um píxel é rotulado como primeiro plano mais de m vezes nos últimos M quadros, a estimação do modelo do fundo $\hat{\mathbf{B}}_l(t)$, também é efetuada, já que, possivelmente, se está diante do problema das mudanças repentinas da iluminação; (b) com a finalidade de superar o problema do fundo dinâmico, todo píxel que muda frequentemente de rótulo (de primeiro plano para fundo), é classificado como um píxel do fundo de altas frequências, e não é considerado parte do primeiro plano.

[12]No caso limite, com $\alpha_{svt} = 1$ a estimação do modelo do fundo é igual ao quadro atual, sendo, neste caso, evidente que a suavização é nula.

[13]Também chamados na literatura como *outliers*, que para esta estimação do modelo do fundo são representados pelos píxeis pertencentes ao primeiro plano.

2.5.3 Modelamento do Fundo Baseado na Mistura de Gaussianas

Em [63] foi proposto um modelo do fundo baseado numa mistura de Gaussianas para cada píxel, onde as componentes das misturas Gaussianas têm pesos normalizados calculados a partir das observações passadas. Assim, a mistura Gaussiana para cada píxel é definida por

$$p(\mathbf{I}_l(t)|\{\omega_{l,k}(t)\}_{k=1}^{K_{mg}}, \{\boldsymbol{\mu}_{l,k}(t)\}_{k=1}^{K_{mg}}, \{\boldsymbol{\Sigma}_{l,k}(t)\}_{k=1}^{K_{mg}}) = \sum_{k=1}^{K_{mg}} \omega_{l,k}(t)\mathcal{N}\left(\boldsymbol{\mu}_{l,k}(t), \boldsymbol{\Sigma}_{l,k}(t)\right), \quad (2.7)$$

onde K_{mg} é o número de Gaussianas da mistura, enquanto $\omega_{l,k}(t)$, $\boldsymbol{\mu}_{l,k}(t)$ e $\boldsymbol{\Sigma}_{l,k}(t)$ são o peso, o valor médio e a matriz de covariância da k^{th} Gaussiana da mistura no quadro t, respectivamente. K_{mg} é estabelecido em relação à memória disponível[14] (em [63] é recomendado de 3 a 5). Também, considera-se que a matriz de covariância tenha a forma $\boldsymbol{\Sigma}_{l,k}(t) = \sigma_{l,k}^2(t)\mathbf{I}$. Esta simplificação implica em que os canais de cores vermelho, verde e azul sejam independentes e tenham a mesma variância. Apesar de não ser o caso, tal suposição permite evitar a inversão da matriz de covariância[15].

Para cada píxel de entrada $\mathbf{I}_l(t)$, o primeiro passo é identificar a componente da mistura cuja média é próxima a $\mathbf{I}_l(t)$. Assim, é determinado que a média $\boldsymbol{\mu}_{l,k}(t-1)$ é próxima à píxel $\mathbf{I}_l(t)$ se $||\mathbf{I}_l(t) - \boldsymbol{\mu}_{l,k}(t-1)||^2 \leq 3{D_{vz}}^2\sigma_{l,k}^2(t-1)$, onde $D_{vz} \in [0,1]$ é um parâmetro de modelamento que permite diminuir o raio da vizinhança centrada em $\mathbf{I}_l(t)$. Esta operação de seleção é representada pelos valores assumidos pela variável binária $Z_{l,k}(t)$, a qual é definida por

$$Z_{l,k}(t) = \begin{cases} 1, & \text{se } ||\mathbf{I}_l(t) - \boldsymbol{\mu}_{l,k}(t-1)||^2 \leq 3{D_{vz}}^2\sigma_{l,k}^2(t-1) \\ 0, & \text{caso contrário} \end{cases}, \quad (2.8)$$

onde os parâmetros das componentes da mistura são atualizados como segue:

- os pesos $\omega_{l,k}(t)$ das K_{mg} distribuições no quadro t são calculados através da relação iterativa[16]

[14] A memória alocada para armazenar os parâmetros das misturas são $5 \times N_{\text{fil}} \times N_{\text{col}} \times K_{mg}$ posições de memória. Portanto, se é considerado que cada quadro é de tamanho 100×100 píxeis, então a quantidade de memória necessária a alocar seria de $0,38$ Mega para um $K_{mg} = 1$ e de $3,8$ Mega para $K_{mg} = 10$.

[15] Se é considerada uma matriz de covariância completa, sua inversão é necessária ao momento de calcular o valor do termo exponencial da distribuição normal multivariável para uma determinada observação.

[16] Esta regra é facilmente interpretada como a interpolação entre dois pontos.

$$\omega_{l,k}(t) = (1 - \alpha_{apr})\omega_{l,k}(t-1) + \alpha_{apr} Z_{l,k}(t), \tag{2.9}$$

onde $\alpha_{apr} \in [0, 1]$ é a taxa de aprendizagem. Depois desta aproximação, todos os pesos são normalizados, para somar 1. O comportamento de $\omega_{l,k}(t)$ é dependente do valor assumido por $Z_{l,k}(t)$. Assim,

- no caso que $Z_{l,k}(t) = 1$, a regra de atualização de $\omega_{l,k}(t)$ se comporta como uma exponencial crescente, implicando que $\omega_{l,k}(t)$ toma um maior valor à medida que a média $\boldsymbol{\mu}_{l,k}(t-1)$ está na vizinhança do píxel $\mathbf{I}_l(t)$ (ver Figura 2.3.a);
- quando $Z_{l,k}(t) = 0$, a regra de atualização de $\omega_{l,k}(t)$ se comporta como uma exponencial decrescente, implicando que $\omega_{l,k}(t)$ assuma um peso de menor valor à medida que a média $\boldsymbol{\mu}_{l,k}(t-1)$ está fora da vizinhança do píxel $\mathbf{I}_l(t)$ (ver Figura 2.3.b).

Este comportamento é equivalente a um sistema de carga e descarga, onde $Z_{l,k}(t)$ atua como uma sinal com dois possíveis estados ($Z_{l,k}(t) = 0$ ou $Z_{l,k}(t) = 1$), e a regra de atualização de $\omega_{l,k}(t)$ atua como um sistema RC discretizado (ver Figura 2.3.c e Figura 2.3.d), permitindo priorizar aquelas componentes da mistura cuja média concorda com os valores do fundo. É importante notar que quanto maior é a taxa de aprendizagem α_{apr}, mais rápido o modelo se adaptada às mudanças nos valores dos píxeis.

- os parâmetros $\boldsymbol{\mu}_{l,k}(t)$ e $\sigma^2_{l,k}(t)$ das K_{mg} distribuições no quadro t são calculados através das equações

$$\boldsymbol{\mu}_{l,k}(t) = \boldsymbol{\mu}_{l,k}(t-1) + \rho(\mathbf{I}_l(t) - \boldsymbol{\mu}_{l,k}(t-1))Z_{l,k}(t), \tag{2.10}$$

$$\sigma^2_{l,k}(t) = \sigma^2_{l,k}(t-1) + \rho\left(\frac{||\mathbf{I}_l(t) - \boldsymbol{\mu}_{l,k}(t)||^2}{3} - \sigma^2_{l,k}(t-1)\right) Z_{l,k}(t), \tag{2.11}$$

onde

$$\rho = \alpha_{apr}\mathcal{N}\left(\boldsymbol{\mu}_{l,k}(t), \boldsymbol{\Sigma}_{l,k}(t)\right). \tag{2.12}$$

Aqui, $\boldsymbol{\mu}_{l,k}(0)$ é inicializado com valores aleatórios entre 0 e 255, $\omega_{l,k}(0)$ e $\sigma_{l,k}(0)$ são inicializados com valores constantes ω_0, σ_0, tal como é indicado a seguir:

$$\boldsymbol{\mu}_{l,k}(0) = 255\mathcal{U}(0, 1), \tag{2.13}$$

$$\omega_{l,k}(0) = \omega_0, \tag{2.14}$$

$$\sigma^2_{l,k}(0) = \sigma^2_0. \tag{2.15}$$

Para o caso onde se tem uma média próxima ao píxel $\mathbf{I}_l(t)$, a Equação (2.10) será definida por uma exponencial crescente ou decrescente, dependendo se a diferença $\mathbf{I}_l(t) - \boldsymbol{\mu}_{l,k}(t-1)$ é positiva ou não (ver Figura 2.4.a). Por outro lado, a Equação (2.11) sempre será definida por uma exponencial decrescente, posto que o fato de considerar um valor para D_{vz} no intervalo $[0, 1]$ assegura que $||\mathbf{I}_l(t) - \boldsymbol{\mu}_{l,k}(t)||^2 - 3\sigma^2_{l,k}(t-1) < 0$ (ver Figura 2.4.b). Ambos comportamentos asseguram que quando se está diante de uma coincidência, $\boldsymbol{\mu}_{l,k}(t)$, partindo de sua condição inicial, converge para $\mathbf{I}_l(t)$ incrementando/decrementando seu valor inicial de uma maneira exponencial até chegar ao valor de $\mathbf{I}_l(t)$, e à medida que se sucede essa aproximação, a variância $\sigma^2_{l,k}(t)$ vai decrescendo, também de uma maneira exponencial, onde a taxa de variação das exponenciais é definida por ρ. Portanto, a convergência é alcançada num número de quadros maior ou igual a $5/\rho$ (nas Figuras 2.4.a e 2.4.b os tempos correspondes à taxa de crescimento $1/\rho$ são indicadas pela linhas vermelhas).

Em [71] os autores propõem tratar o problema das mudanças repentinas da iluminação e do fundo dinâmico, de forma similar à proposta em [33] para o caso da técnica baseada numa média móvel gaussiana. Assim, a atualização dos parâmetros das componentes da mistura (Equações 2.9,2.10 e 2.11) é feita somente sobre aqueles píxeis classificados como fundo ou quando um píxel é rotulado como primeiro plano mais de m vezes nos últimos M quadros e em caso de que mude frequentemente de rótulo, ele é rotulado como fundo.

Em [63] se determinou heuristicamente que a gaussiana de cada mistura que tem a maior probabilidade de representar o fundo é aquela que apresenta o maior peso ω e a menor variância σ^2. Para entender esta escolha, são apresentados dois exemplos: em (a) tem-se um objeto do fundo, onde as componentes das misturas gaussianas que o representam como parte do fundo serão atualizadas através de uma exponencial crescente. Portanto, apresentam um maior peso e, à medida que o valor médio das componentes das misturas ficam próximas do valor dos píxeis objeto do fundo a variância das componentes sofre um decaimento exponencial; em (b) tem-se um objeto do primeiro plano, onde, em geral, seus valores não correspondem a uma das distribuições já existentes, o que irá resultar na criação de uma nova distribuição ou no aumento da variância de uma distribuição já existente. Um procedimento que permite decidir que componentes das misturas representam melhor o fundo é descrito a seguir:

- define-se a medida de confiança β_k como

$$\beta_k = \omega_{l,k}(t)/\sigma_{l,k}(t), \tag{2.16}$$

onde β_k incrementa seu valor quando a componente da mistura K_{mg}-ésima ganha um maior peso $\omega_{l,k}(t)$, ou quando sua variância $\sigma_{l,k}(t)$ diminui;

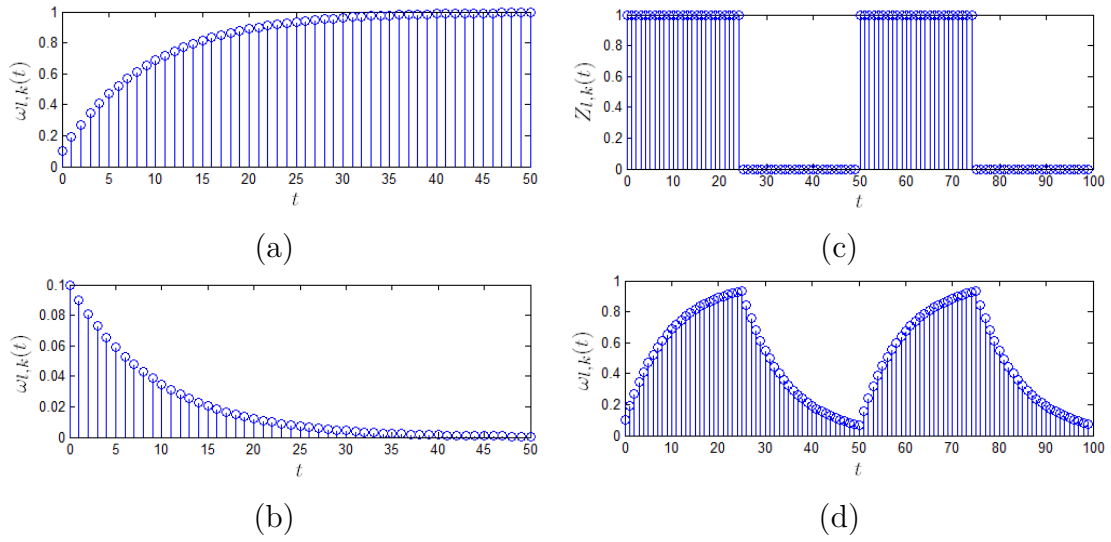


Figura 2.3: Comportamento da função de atualização dos pesos $\omega_{l,k}(t)$. (a) Considerando que a média $\boldsymbol{\mu}_{l,k}(t-1)$ está na vizinhança do píxel $\mathbf{I}_l(t)$, então $Z_{l,k}(t) = 1$ e portanto $\omega_{l,k}(t) = (1-\alpha_{apr})\omega_{l,k}(t-1) + \alpha_{apr}$. (b) Considerando que a média $\boldsymbol{\mu}_{l,k}(t-1)$ está fora da vizinhança do píxel $\mathbf{I}_l(t)$ e portanto $\omega_{l,k}(t) = (1-\alpha_{apr})\omega_{l,k}(t-1)$. (c) Considerando que $Z_{l,k}(t)$ é definida como uma sinal de dois possiveis estados ($Z_{l,k}(t) = 0$ ou $Z_{l,k}(t) = 1$), (d) a saída definida pela Equação (2.9). Para todos os gráficos foi assumido um $\alpha_{apr} = 0,1$ e $\omega_{l,k}(0) = 0,1$.

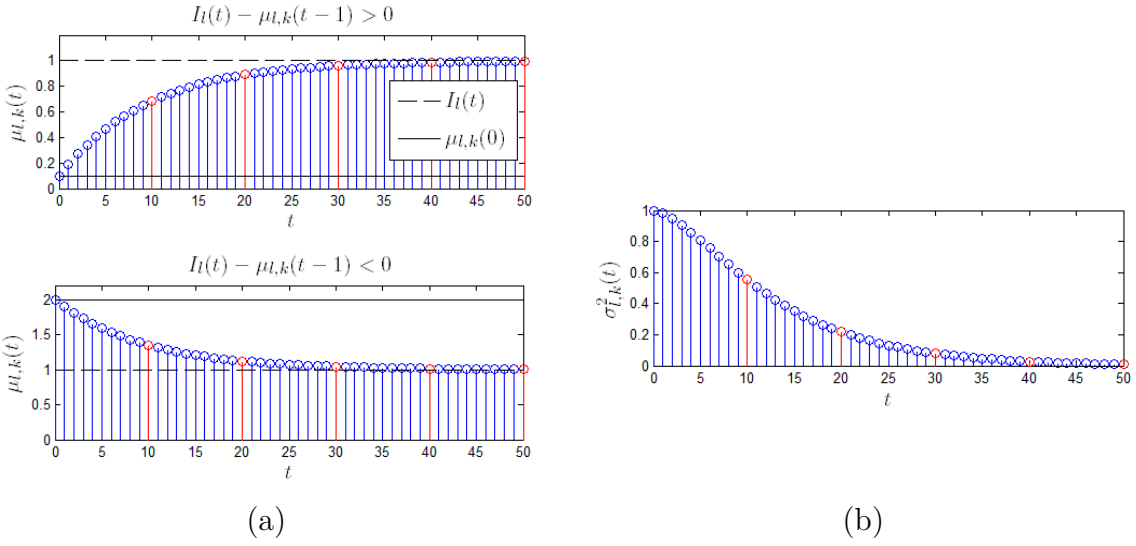


Figura 2.4: Comportamento das funções de atualização da média $\boldsymbol{\mu}_{l,k}(t)$ e a variância $\sigma^2_{l,k}(t)$, quando $Z_{l,k}(t) = 1$, (a) $\boldsymbol{\mu}_{l,k}(t)$ é definida por uma exponencial crescente (gráfico superior) ou decrescente (gráfico inferior). Para o primeiro caso se considerou como condição inicial $\boldsymbol{\mu}_{l,k}(0) = 0,2$ e como valor de referência $\mathbf{I}_l(t) = 1$ e para o segundo caso se considerou $\boldsymbol{\mu}_{l,k}(0) = 2$ e $\mathbf{I}_l(t) = 1$, (b) em quanto que $\sigma^2_{l,k}(t)$ sempre será uma exponencial decrescente.

- considerando que as componentes da mistura que apresentem uma alta medida de confiança representam o fundo, é necessário determinar a sequência de ordenamento de β_k, ou seja

$$\{\kappa_1, \kappa_2, \cdots, \kappa_{K_{mg}}\} \leftarrow \text{argsort}\left\{\{\beta_k\}_{k=1}^{K_{mg}}\right\}. \tag{2.17}$$

  A operação argsort $\{\bullet\}$ retorna uma sequência de índices que podem ser usados para indexar a sequência de entrada, de modo que o resultado corresponda com a sequência de entrada ordenada de maior a menor, ou seja, $\beta_{\kappa_1} > \beta_{\kappa_2} > \cdots > \beta_{\kappa_{K_{mg}}}$. Esta ordenação permite que as componentes da mistura com maior probabilidade de representar o fundo permaneçam no topo, e as menos prováveis ficam por último;

- considerando a soma acumulada dos pesos ordenados em relação à sequência $\{\beta_k\}_{k=1}^{K_{mg}}$, é determinada a variável binária $O_{l,k}(t)$ que contém os candidatos das componentes da mistura que podem definir o fundo, segundo

$$O_{l,\kappa_b}(t) = \begin{cases} 1, & \text{se } \text{argmin}_{b_{mg}} \left\{ \sum_{i=1}^{b_{mg}} \omega_{l,\kappa_i}(t) > T_{\text{fundo}} \right\} \\ 0, & \text{caso contrário} \end{cases}, \tag{2.18}$$

  onde b_{mg} é o número de componentes da mistura que definem o fundo em relação ao limiar T_{fundo} aplicado sobre a soma dos pesos. O valor que assume T_{fundo} permite tomar as melhores componentes da mistura até certa quantidade. Por exemplo, se é selecionado um pequeno valor para T_{fundo}, então o modelo do fundo tem uma distribuição unimodal. Se T_{fundo} apresenta um grande valor, então o modelo do fundo tem uma distribuição multimodal;

- finalmente, a máscara do primeiro plano é definida por

$$\mathbf{M}_l(t) = \begin{cases} 1, & \text{se existe um } k, \text{tal que: } O_{l,k}(t) \neq Z_{l,k}(t) \\ 0, & \text{caso contrário} \end{cases}. \tag{2.19}$$

Note-se que esta técnica unifica os passos de modelamento do fundo e detecção de variações, ao determinar a máscara do primeiro plano $\mathbf{M}_l(t)$ a partir da limiarização da soma acumulada dos pesos $\omega_{l,\kappa_i}(t)$.

2.5.4 Modelo do Fundo Baseado no Histograma

Os métodos baseados no histograma consideram que para uma sequência de imagens capturadas por uma câmera fixa o fundo é quase estático, implicando que as cores de seus

píxeis sejam aproximadamente similares para todo quadro t. Por conseguinte, quanto mais frequente um píxel toma um valor de um determinado intervalo de intensidades, é de se esperar que a intensidade com o maior número de ocorrências represente um píxel do modelo do fundo. Formalmente, pode-se expressar isto como

$$\mathbf{B}_l(t) = \text{Moda}\left(\mathcal{I}_l(t)\right), \tag{2.20}$$

onde, o operador Moda $(\bullet)$ retorna o valor que detém o maior número de observações, isto é, o valor mais comum. Partindo desse pressuposto, são usados histogramas vinculados a cada um dos canais de cores de um processo de um píxel para obter as cores do píxel correspondente no modelo do fundo. No histograma, o número de ocorrências de cada intensidade é proporcional à probabilidade que ela tenha. Assim, a moda de um processo de um píxel é decidida em função da intensidade que apresente o máximo número de ocorrências no histograma, ou seja

$$\mathbf{B}_l(t) = Moda\left(\mathcal{I}_l(t)\right) = \begin{bmatrix} \underset{b=0,\cdots,255}{\text{argmax}} \left\{h_l^R(b)\right\} \\ \underset{b=0,\cdots,255}{\text{argmax}} \left\{h_l^G(b)\right\} \\ \underset{b=0,\cdots,255}{\text{argmax}} \left\{h_l^B(b)\right\} \end{bmatrix}, \tag{2.21}$$

onde $h_l^C(b)$ é o número de ocorrências da intensidade b na posição (x_l, y_l) do canal $C \in \{R, G, B\}$, considerando-se 256 níveis de intensidades (escala de cinza). Entretanto, devido a problemas como fundo dinâmico, ruído de vídeo, mudanças graduais da iluminação, entre outros, a Equação (2.20) não é totalmente válida. Também é bastante frequente que várias das intensidade tenham o mesmo número máximo de ocorrências no histograma de cada canal, sendo mais difícil determinar o valor da moda a partir dos histogramas. Para incrementar a robustez contra tais tipos de problemas é necessário gerar os histogramas com um número grande de quadros. O problema é que o custo computacional é alto, degradando o desempenho em tempo real.

Diferentes soluções alternativas que modificam o cálculo do histograma foram propostas na literatura: em [66] os histogramas são correlatados numa vizinhança, em [83] é realizada uma operação de agrupamento do número de ocorrências das intensidades dos histogramas e em [62] os histogramas calculados são filtrados usando um filtro de ponderação, ou seja, tem-se que

$$\hat{h}_l^C(b) = \sum_{k=-N_{\text{filtro}}}^{N_{\text{filtro}}} h_l^C(b+k), \tag{2.22}$$

onde N_{filtro} é uma variável que define o tamanho do filtro de ponderação. Na Figura 2.5 pode-se observar que o efeito da filtragem é a suavização das bordas dos grupos de ocorrências, atenuando, assim, o problema do máximo número de ocorrências repetidas. Esta proposta é a mais aceitável, considerando o custo computacional, já que só implica em um número de somas proporcional a N_{filtro}. Em [62] também é definida uma expressão que vincula a largura do filtro N_{filtro} com a medida de dispersão do histograma $\sigma_l^C(b)$, de forma que se tem

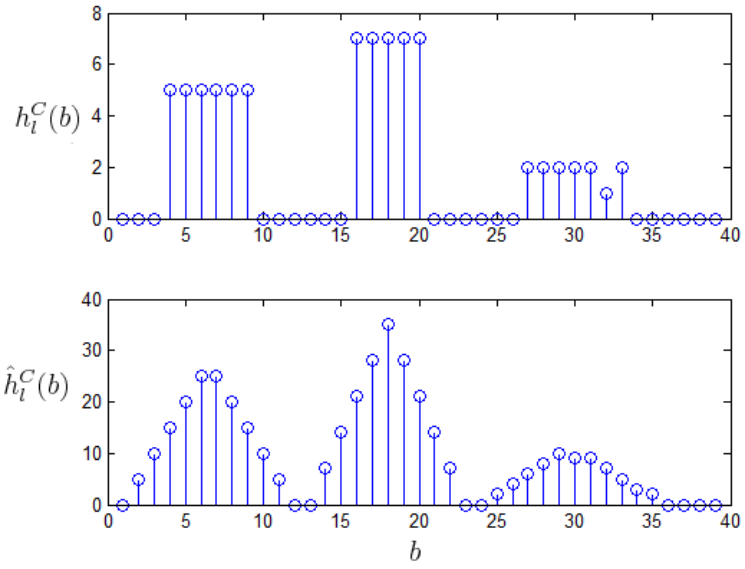


Figura 2.5: Efeito do filtro de ponderação sob $h_l^C(b)$, considerando $N_{\text{filtro}} = 5$.

$$N_{\text{filtro}} = \begin{cases} 3, & \sigma_l^C(b) \leq 7 \\ 5, & 8 \leq \sigma_l^C(b) \leq 10 \\ 7, & 11 \leq \sigma_l^C(b) \end{cases}, \tag{2.23}$$

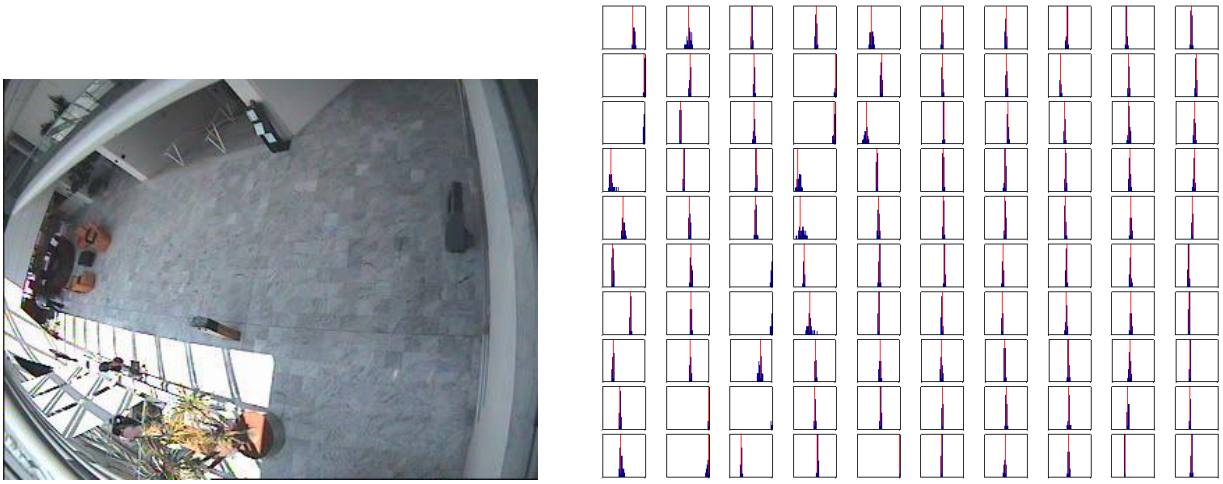
tal que

$$\sigma_l^C(b) = \sqrt{\frac{1}{\sum_{b=\mathbf{B}_C(l,t)-3\sigma_{ds}}^{\mathbf{B}_C(l,t)+3\sigma_{ds}} h_l^C(b)} \sum_{b=\mathbf{B}_C(l,t)-3\sigma_{ds}}^{\mathbf{B}_C(l,t)+3\sigma_{ds}} (b - \mathbf{B}_C(l,t))^2 h_l^C(b)}, \tag{2.24}$$

onde, $\mathbf{B}_C(l,t)$ é a cor do modelo do fundo na posição (x_l, y_l) no canal $C \in \{R, G, B\}$ no

instante t, e σ_{ds} é um parâmetro que define o tamanho da vizinhança centrada em $\mathbf{B}_C(l,t)$ na qual é calculada $\sigma_l^C(b)$.

Na Figura 2.6.a é apresentado o modelo do fundo obtido com esta técnica e na Figura 2.6.b os histogramas $h_l^C(b)$ vinculados ao canal R do modelo do fundo, para um conjunto de píxeis correspondentes aos centros de uma grade de 10×10 blocos feita somente para visualização dos histogramas (não é prático visualizar os histogramas de todos os píxeis do modelo do fundo). Pode-se observar que na maioria dos casos os histogramas apresentam um comportamento unimodal, não necessariamente centrado, sendo assim evidente a validade da Equação (2.20) para o cálculo do modelo do fundo.



(a) (b)

Figura 2.6: (a) O modelo do fundo $\mathbf{B}(t)$ determinado pela técnica baseada no histograma, vinculado ao vídeo *Browse1.mpg* do banco de dados PETS2004. (b) histogramas vinculados ao canal R do modelo do fundo, onde a linha vertical (na cor vermelha) indica o valor corresponde de $\mathbf{B}_R(l,t)$ no correspondente h_l^R.

2.6 Detecção de Variações

O método mais trivial para realizar a detecção de variações é baseado na diferença do modelo do fundo e cada quadro do vídeo, $\mathbf{I}_l(t) - \mathbf{B}_l(t)$. Têm-se várias técnicas baseadas nesta ideia, como é apresentado a seguir.

- Em [33] um píxel é marcado como primeiro plano se o valor absoluto da diferença entre os correspondentes píxeis em duas imagens em escala de cinza é superior a um limiar predefinido.

- Em [31] foi usada a diferença relativa normalizada. Uma versão para imagens a cores é apresentada em [19], onde é indicado obter um melhor desempenho na segmentação em comparação a imagens em escala de cinza, especialmente em áreas de baixo contraste, tais como objetos em áreas obscuras, onde usou-se a norma L_1[17] como uma métrica que quantifica a informação obtida pela diferença entre os correspondentes píxeis nas duas imagens a cores.
- Em [18] e [28] foi usada a norma L_∞[18], e em [75] foi usada a distância de Mahalanobis.
- Em [10] são usadas técnicas de detecção de bordas aplicadas à diferença entre a imagem atual e a referência, provendo uma maior robustez às variações de iluminação.

A seguir é explicada a técnica proposta em [4][15] para a realização da detecção de variações. Esta técnica utiliza a diferença entre o modelo do fundo e um quadro da sequência de vídeo para achar as regiões de troca de valores, trabalha na escala de cinza da diferença de imagens, e utiliza uma proposta estatística baseada tanto em teste de hipótese como em teoria de decisão Bayesiana.

Técnica Baseada no Teste de Significância

Define-se $\mathbf{D}_l(t)$ como a diferença entre cada quadro do vídeo e o modelo do fundo para cada píxel l num instante t, conforme

$$\begin{aligned}\mathbf{D}_l(t) &= |0,2989(\mathbf{I}_R(l,t)-\mathbf{B}_R(l,t))+0,5870(\mathbf{I}_G(l,t)-\mathbf{B}_G(l,t)) \\ &\quad +0,1140(\mathbf{I}_B(l,t)-\mathbf{B}_B(l,t))|. \end{aligned} \tag{2.25}$$

A máscara do primeiro plano $\mathbf{M}_l(t)$ é uma matriz binária, onde para cada píxel l um determinado valor binário é estabelecido, considerando-se que o valor assumido por $\mathbf{D}_l(t)$atende uma das seguintes hipóteses:

- $\mathbf{M}_l(t)=0$, se $\mathbf{D}_l(t)$ atende à hipótese que o valor da diferença é devido unicamente ao ruído presente, e, portanto, o píxel l corresponde a uma região sem variações entre o modelo do fundo e o quadro atual;

[17] A norma L_1 para o vetor $\mathbf{x}=[x_1,\cdots,x_n]^T$ é calculada pela expressão: $||\mathbf{x}||_1=|x_1|+\cdots+|x_n|$.
[18] A norma L_∞ para o vetor $\mathbf{x}=[x_1,\cdots,x_n]^T$ é calculada pela expressão: $||\mathbf{x}||_\infty=\max\{|x_1|,\cdots,|x_n|\}$.

- $\mathbf{M}_l(t) = 1$, se $\mathbf{D}_l(t)$ não atende à hipótese anterior, implicando que o píxel l corresponde a uma região onde o modelo do fundo varia em relação ao quadro atual.

Assim, o problema de detecção de variações pode ser formulado como um problema de classificação em duas categorias ou, equivalentemente, como um problema de avaliação de hipótese [29]. Se é assumido que o valor de $\mathbf{M}_l(t)$ não é conhecido, então o problema se reduz a decidir se $\mathbf{M}_l(t) = 1$ ou $\mathbf{M}_l(t) = 0$. O problema de decisão pode ser proposto em termos probabilísticos, supondo o conhecimento das probabilidades a posteriori vinculadas a cada decisão, dado que unicamente é conhecido o valor da diferença $\mathbf{D}_l(t)$. Ou seja, se é denotado por $\mathcal{H}_{\text{var}}$ a ação de decidir por $\mathbf{M}_l(t) = 1$ e aquela produzida por $\mathbf{M}_l(t) = 0$ é denotada como $\mathcal{H}_{\text{invar}}$, então pode-se estimar a máscara do primeiro plano $\mathbf{M}_l(t)$ tal que $p(\mathcal{H}_{\text{var}}|\mathbf{D}_l(t))$ é maximizada, tendo-se a regra de decisão

$$\mathbf{M}_l(t) = \begin{cases} 1, & \text{se } p(\mathcal{H}_{\text{var}}|\mathbf{D}_l(t)) > p(\mathcal{H}_{\text{invar}}|\mathbf{D}_l(t)) \\ 0, & \text{se } p(\mathcal{H}_{\text{var}}|\mathbf{D}_l(t)) < p(\mathcal{H}_{\text{invar}}|\mathbf{D}_l(t)) \end{cases}. \tag{2.26}$$

A expressão anterior é definida como regra de decisão de máximo a posteriori, e é expressa, em forma mais compacta, como

$$p(\mathcal{H}_{\text{var}}|\mathbf{D}_l(t)) \underset{0}{\overset{1}{\gtrless}} p(\mathcal{H}_{\text{invar}}|\mathbf{D}_l(t)). \tag{2.27}$$

Expressando-a em termos do logaritmo das probabilidades a posteriori, obtém-se

$$p(\mathcal{H}_{\text{var}}|\mathbf{D}_l(t)) \underset{0}{\overset{1}{\gtrless}} p(\mathcal{H}_{\text{invar}}|\mathbf{D}_l(t)) \tag{2.28}$$

$$\frac{p(\mathcal{H}_{\text{var}}|\mathbf{D}_l(t))}{p(\mathcal{H}_{\text{invar}}|\mathbf{D}_l(t))} \underset{0}{\overset{1}{\gtrless}} 1 \tag{2.29}$$

$$\ln\left(\frac{p(\mathcal{H}_{\text{var}}|\mathbf{D}_l(t))}{p(\mathcal{H}_{\text{invar}}|\mathbf{D}_l(t))}\right) \underset{0}{\overset{1}{\gtrless}} 0, \tag{2.30}$$

sendo que a razão do logaritmo das probabilidades a posteriori é chamada função discriminante, e é definida como

$$g(\mathbf{D}_l(t)) = \ln\left(\frac{p(\mathcal{H}_{\text{var}}|\mathbf{D}_l(t))}{p(\mathcal{H}_{\text{invar}}|\mathbf{D}_l(t))}\right) = \ln\left(\frac{p(\mathbf{D}_l(t)|\mathcal{H}_{\text{var}})}{p(\mathbf{D}_l(t)|\mathcal{H}_{\text{invar}})}\right) + \ln\left(\frac{p(\mathcal{H}_{\text{var}})}{p(\mathcal{H}_{\text{invar}})}\right). \tag{2.31}$$

Já que as posições onde os píxeis da diferença de imagens $\mathbf{D}(t)$ são frequentemente associadas com conjuntos de píxeis no lugar dos píxeis isolados, é comum que a função discriminante seja avaliada numa pequena vizinhança centrada no píxel de análise l [56]. Assim, define-se a vizinhança $\mathcal{D}_l(t) = \{\mathbf{D}_r(t)|(x_r, y_r) \in \mathcal{N}_l\}$, assumindo que a vizinhança tem uma cardinalidade igual a $N_\mathcal{D}$, e cada um de seus elementos são independentes [4]. Então a função discriminante toma a forma

$$g(\mathcal{D}_l(t)) = \sum_{(x_r,y_r)\in\mathcal{N}_l} \ln\left(\frac{p(\mathbf{D}_r(t)|\mathcal{H}_{\text{var}})}{p(\mathbf{D}_r(t)|\mathcal{H}_{\text{invar}})}\right) + \ln\left(\frac{p(\mathcal{H}_{\text{var}})}{p(\mathcal{H}_{\text{invar}})}\right). \tag{2.32}$$

Em [4] é proposto aplicar uma distribuição Gaussiana com média igual a zero para a probabilidade condicional $p(\mathbf{D}_r(t)|\mathcal{H})$, Assim, tem-se que

$$p(\mathbf{D}_r(t)|\mathcal{H}_{\text{var}}) = \frac{1}{\sqrt{2\pi}\sigma_{\text{var}}} \exp\left(-\frac{\mathbf{D}_r^2(t)}{{\sigma_{\text{var}}}^2}\right), \tag{2.33}$$

$$p(\mathbf{D}_r(t)|\mathcal{H}_{\text{invar}}) = \frac{1}{\sqrt{2\pi}\sigma_{\text{invar}}} \exp\left(-\frac{\mathbf{D}_r^2(t)}{{\sigma_{\text{invar}}}^2}\right). \tag{2.34}$$

Substituindo as Equações (2.33) e (2.34) na Equação (2.32) a função discriminante fica definida como

$$g(\mathcal{D}_l(t)) = \frac{1}{2\sigma_{\text{invar}}^2}\left(1 - \frac{\sigma_{\text{invar}}^2}{\sigma_{\text{var}}^2}\right) \sum_{(x_r,y_r)\in\mathcal{N}_l} \mathbf{D}_r^2(t) - N_\mathcal{D} \ln\left(\frac{\sigma_{\text{var}}}{\sigma_{\text{invar}}}\right) + \ln\left(\frac{p(\mathcal{H}_{\text{var}})}{p(\mathcal{H}_{\text{invar}})}\right). \tag{2.35}$$

Como as áreas vinculadas às variações normalmente apresentam diferenças de grande magnitude, o desvio padrão de σ_{var} é maior que σ_{invar} causado pelo ruído. Portanto, a relação anterior pode ser simplificada para

$$g(\mathcal{D}_l(t)) = \frac{1}{2\sigma_{\text{invar}}^2} \sum_{(x_r,y_r)\in\mathcal{N}_l} \mathbf{D}_r^2(t) - N_\mathcal{D} \ln\left(\frac{\sigma_{\text{var}}}{\sigma_{\text{invar}}}\right) + \ln\left(\frac{p(\mathcal{H}_{\text{var}})}{p(\mathcal{H}_{\text{invar}})}\right). \tag{2.36}$$

Portanto, a regra de decisão de máximo a posteriori é expressa como

$$\frac{1}{\sigma_{\text{invar}}^2} \sum_{(x_r,y_r)\in\mathcal{N}_l} \mathbf{D}_r^2(t) \underset{0}{\overset{1}{\gtrless}} 2N_\mathcal{D} \ln\left(\frac{\sigma_{\text{var}}}{\sigma_{\text{invar}}}\right) + 2\ln\left(\frac{p(\mathcal{H}_{\text{invar}})}{p(\mathcal{H}_{\text{var}})}\right). \tag{2.37}$$

Note-se que:

- O termo do lado esquerdo da Equação (2.37) indica que tudo que se necessita conhecer sobre as observações $\mathcal{D}_l(t)$ é sua soma. Esta quantidade é a suficiência estatística do problema, definida por
$$s_l(t) = \sum_{(x_r,y_r)\in\mathcal{N}_l} \mathbf{D}_r^2(t). \tag{2.38}$$

- O termo do lado direito da Equação (2.37) pode ser definido como a soma de dois limiares de decisão, o primeiro denotado por τ_σ, dependente dos parâmetros σ_{var} e σ_{invar}, e o segundo denotado por τ_{prior}, dependente das probabilidades a priori das decisões $\mathcal{H}_{\text{var}}$ e $\mathcal{H}_{\text{invar}}$. Assim
$$\tau_\sigma = 2N_{\mathcal{D}} \ln\left(\frac{\sigma_{\text{var}}}{\sigma_{\text{invar}}}\right) \tag{2.39}$$
$$\tau_{\text{prior}} = 2\ln\left(\frac{p(\mathcal{H}_{\text{invar}})}{p(\mathcal{H}_{\text{var}})}\right). \tag{2.40}$$

- Substituindo as Equações (2.38),(2.39) e (2.40) na Equação (2.37) obtém-se a regra de decisão simplificada, a saber,
$$\mathbf{M}_l(t) = \begin{cases} 1, & \text{se } \frac{1}{\sigma_{\text{invar}}^2} s_l(t) > \tau_\sigma + \tau_{\text{prior}} \\ 0, & \text{caso contrário} \end{cases}. \tag{2.41}$$

A determinação dos valores para ambos limiares é realizada de forma separada. Em [4][15] é utilizado um nível de significância α para determinar o valor de τ_σ, isto é, $\mathbb{P}(s_l(t) > \tau_\sigma) = 1 - \alpha$[19], onde é aproveitado o fato que $s_l(t)$, ao ser definido como uma soma de variáveis aleatórias gaussianas ao quadrado, seguirá uma distribuição chi quadrado com $N_{\mathcal{D}}$ graus de liberdade. O valor assumido para α deve ser suficientemente pequeno para assegurar um erro mínimo ao classificar como fundo um píxel que pertence ao primeiro plano, ou seja,

$$\begin{aligned} \mathbb{P}(\mathbf{M}_l(t) = 0|\mathcal{H}_{\text{var}}) &= 1 - \mathbb{P}(\mathbf{M}_l(t) = 1|\mathcal{H}_{\text{var}}) \\ &= 1 - \mathbb{P}(s_l(t) > \tau_\sigma) \\ &= \alpha. \end{aligned}$$

[19]Neste caso a notação $\mathbb{P}(\mathcal{A})$ faz referência à probabilidade do evento $\mathcal{A}$ acontecer.

Em [4][3][2] é considerado que a máscara do primeiro plano pode ser modelada através de um campo aleatório Markoviano, tal que as regiões detectadas tendem a ser compactas. A partir desta suposição as probabilidades $p(\mathcal{H}_{\text{invar}})$ e $p(\mathcal{H}_{\text{var}})$ podem ser definidas como função exponencial, implicando que o limiar τ_{prior} tome a forma[20]

$$\tau_{\text{prior}} = (4 - v_c)B. \tag{2.42}$$

O parâmetro v_c denota o número de píxeis que são rotulados como primeiro plano numa vizinhança 3×3 centrada no píxel l. Tais rótulos são conhecidos para aqueles píxeis vizinhos que já tenham sido processados. Para os píxeis da vizinhança que ainda não foram processados, simplesmente se considera os rótulos que assumiram na máscara do primeiro plano do quadro anterior. Claramente, τ_{prior} só pode ter nove valores diferentes, definidos por $v_c(= 0, 1, \cdots, 8)$, os quais podem ser pré-computados e armazenados em uma tabela de consulta. Assim, o limiar é menor ou maior em relação ao número de píxeis adjacentes rotulados como primeiro plano. Portanto, a decisão $\mathbf{M}_l(t) = 1$ é favorecida quanto mais píxeis que variam em seu rotulamento são encontrados na vizinhança centrada no píxel l. Claramente, este comportamento favorece o surgimento de regiões compactas e desencoraja erros de decisão devidos ao ruído.

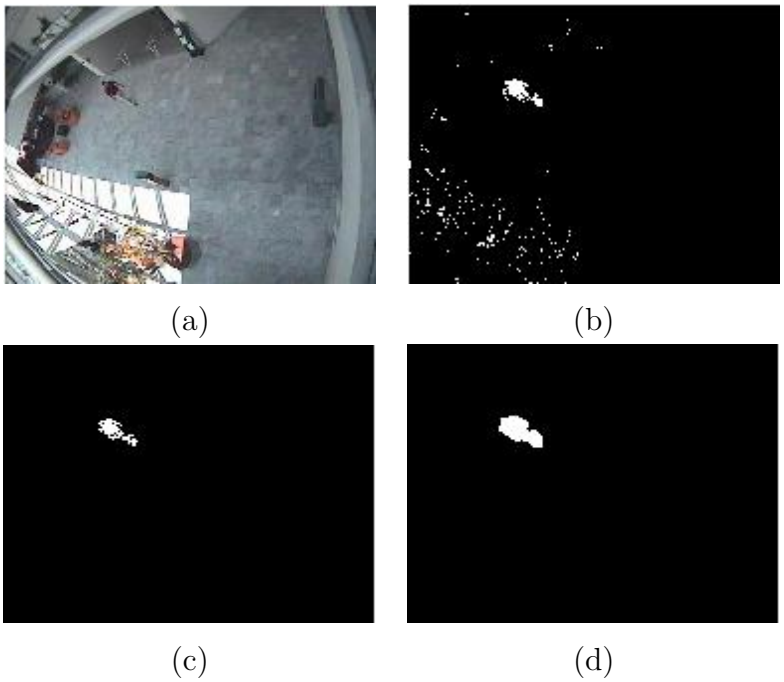
2.7 Pós-processamento

Devido às limitações do modelo do fundo, a máscara do primeiro plano $\mathbf{M}_l(t)$ tipicamente contém uma grande quantidade de pequenas manchas. Estas manchas podem ser removidas pela aplicação de algoritmos de filtragem em $\mathbf{M}_l(t)$. Neste trabalho, foi aplicado um algoritmo de agrupamento por componentes conectados para sua eliminação, considerando-se que as manchas têm um tamanho menor se comparadas aos objetos em movimento. À imagem resultante são aplicados filtros morfológicos para unir aquelas regiões pertencentes ao primeiro plano que são próximas porém não conectadas (é realizada uma operação de dilatação seguida de uma operação de erosão).

A Figura 2.7 apresenta cada fase do pós-processamento para um quadro específico (ver Figura 2.7.a). Inicialmente, a máscara do primeiro plano $\mathbf{M}_l(t)$ é calculada (ver Figura 2.7.b), então o algoritmo de agrupamento por componentes conectados é aplicado a $\mathbf{M}_l(t)$, e a eli-

[20]Os detalhes da demonstração de como obter a Equação (2.42) a partir da suposição de que a máscara do primeiro plano é um campo aleatório Markoviano podem ser encontrados em [4].

minação dos objetos com menos de N_{MinArea} píxeis é observada na Figura 2.7.c. Finalmente, a imagem resultante da aplicação dos filtros morfológicos pode ser vista na Figura 2.7.d.



(a) (b)

(c) (d)

Figura 2.7: Passos do pós-processamento: (a) quadro, (b) máscara do primeiro plano, (c) filtragem de manchas, (d) filtragem morfológica.

2.8 Proposta

Nesta seção são apresentadas duas propostas para tratar o problema de detecção de variações, sendo a ideia principal das abordagens propostas determinar a máscara do primeiro plano $\mathbf{M}_l(t)$ de acordo com a regra de decisão

$$\mathbf{M}_l(t) = \begin{cases} 1, & \text{se } g(d_l(t)) > \tau \\ 0, & \text{caso contrário} \end{cases}, \tag{2.43}$$

onde:

- $d_l(t)$ é a distância Euclidiana, definida como

$$d_l(t) = d(\mathbf{I}_l(t), \mathbf{B}_l(t)) = ||\mathbf{I}_l(t) - \mathbf{B}_l(t)||, \tag{2.44}$$

e $||\bullet||$ é a norma L_2[21];

- $g(\bullet)$ é uma função discriminante aplicada à distância Euclidiana;
- τ é o limiar de decisão que permite a detecção de variações.

Ao considerar a detecção de variações através da Equação (2.43), surgem dois problemas a tratar:

- Em relação a $d_l(t)$, a distância Euclidiana no espaço de cores RGB não quantifica a similaridade das cores. Entretanto, em geral pode-se dizer que a distância Euclidiana é sensível a variações de brilho, mas não é muito sensível a variações em matiz e saturação [74]. Uma métrica alternativa é a medida angular entre vetores, que é definida como

$$\lambda = 1 - \cos(\theta) \tag{2.45}$$

$$\cos(\theta) = \frac{\mathbf{I}_l(t)^T \mathbf{B}_l(t)}{||\mathbf{I}_l(t)||||\mathbf{B}_l(t)||}, \tag{2.46}$$

  onde θ é o ângulo entre $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ e λ é uma quantidade adimensional que varia entre $[0, 2]$.[22]Ao contrário da distância Euclidiana, a medida angular é insensível às diferenças de brilho, mas quantifica bem as diferenças de matiz[23]. Assim, considerando que a própria definição da distância Euclidiana contém a medida angular entre vetores, surge a possibilidade de obter uma representação da distância Euclidiana (indicada na Equação (2.43) através da aplicação da função discriminante $g(\bullet)$) que permita ponderar a informação da medida angular λ como das magnitudes de $||\mathbf{I}_l(t)||, ||\mathbf{B}_l(t)||$, tornando explícito o uso da informação de cromaticidade como de brilho no cálculo da diferença entre $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$.

- Em relação ao limiar τ, considerando que uma técnica de subtração de fundo trata diferentes desafios, supor um limiar fixo limita a capacidade de adaptação da técnica e a necessidade de uma sintonia contínua, ponto que é crítico em aplicações de vídeo que são de natureza *on-line*. Portanto, tem-se a necessidade de determinar o limiar τ de maneira auto-sintonizada, implicando que seu valor deve ser determinado a partir da distribuição de probabilidades vinculada à representação da distância Euclidiana $f(d_l(t))$.

[21]A norma L_2 para o vetor $\mathbf{x} = [x_1, \cdots, x_n]^T$ é calculada pela expressão: $||\mathbf{x}||_2 = ||\mathbf{x}|| = \sqrt{x_1^2 + \cdots + x_n^2}$.
[22]Considerando que $\cos(\theta) \in [-1, 1]$, então, $\lambda \in [0, 2]$, já que, $\lambda = 1 - \cos(\theta)$.
[23]No Anexo D é feita uma demostração desta afirmação usando as coordenadas do modelo de cores HSB.

Levando em conta estas considerações, foram elaboradas duas técnicas de detecção de variações, onde: (a) é definida uma medida de similaridade das cores obtida a partir da distância Euclidiana; (b) é estudada uma representação da distância Euclidiana como uma função bivariada de duas métricas. Note-se indicar que a detecção de variações é um passo importante, já que uma determinação ótima da máscara do primeiro plano a partir do quadro de entrada assegura um bom desempenho para uma técnica de subtração de fundo.

2.8.1 Técnica baseada na Distância Euclidiana Simplificada

A partir da Equação (2.44), a distância Euclidiana é representada como uma função das magnitudes de suas componentes. Assim, tem-se que

$$d_l^2(t) = ||\mathbf{I}_l(t)||^2 + ||\mathbf{B}_l(t)||^2 - 2||\mathbf{I}_l(t)\mathbf{B}_l(t)|| \cos(\theta), \tag{2.47}$$

e substituindo a Equação (2.45) na Equação (2.47) obtém-se

$$d_l^2(t) = ||\mathbf{I}_l(t)||^2 + ||\mathbf{B}_l(t)||^2 - 2||\mathbf{I}_l(t)\mathbf{B}_l(t)|| + 2||\mathbf{I}_l(t)\mathbf{B}_l(t)||\lambda. \tag{2.48}$$

Na Equação (2.48), são considerados os seguintes casos, em relação aos valores do ângulo θ:

- se $\theta = 0^o$, então, a distância Euclidiana é definida como

$$d_{0^o}^2 = ||\mathbf{I}_l(t)||^2 + ||\mathbf{B}_l(t)||^2 - 2||\mathbf{I}_l(t)\mathbf{B}_l(t)||; \tag{2.49}$$

- se $\theta = 90^o$, então, a distância Euclidiana é definida como

$$d_{90^o}^2 = ||\mathbf{I}_l(t)||^2 + ||\mathbf{B}_l(t)||^2. \tag{2.50}$$

Substituindo as Equações (2.49) e (2.50) na Equação (2.48), a distância Euclidiana é representada como uma combinação linear convexa. Assim

$$d_l^2(t) = \lambda d_{90^o}^2 + (1 - \lambda) d_{0^o}^2. \tag{2.51}$$

Note-se que: (a) da Equação (2.51), que λ atua ponderando a importância das magnitudes $d^2_{90^o}$ e $d^2_{0^o}$ na conformação da distância Euclidiana $d_l(t)$; (b) das Equações (2.49) e (2.50), que $d^2_{0^o}$ e $d^2_{90^o}$ dependem unicamente das magnitudes $||\mathbf{I}_l(t)||$, $||\mathbf{B}_l(t)||$, contendo, portanto, a informação de brilho e a saturação dos píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$[24]. Assim, uma simplificação da distância Euclidiana, que contenha tanto a diferença angular (informação da matiz) como a diferença das magnitudes (informação de brilho) dos píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ e, portanto, tenha um maior poder de discriminação quando $\mathbf{I}_l(t)$ é parte do primeiro plano, é dada pelo termo $(1-\lambda)d^2_{0^o}$ da Equação (2.51). Assim é definida uma medida de similaridade das cores como:

$$\rho_l(t) = g(d_l(t)) = \sqrt{1-\lambda}d_{0^o}. \tag{2.52}$$

A Equação (2.52) pode ser simplificada (sem perder sua capacidade de discriminação) ao considerar unicamente o primeiro termo da série de Taylor da expressão $\sqrt{1-\lambda}$. Portanto, a Equação (2.52) é rescrita como

$$\rho_l(t) = \left(1 - \frac{\lambda}{2}\right) d_{0^o}. \tag{2.53}$$

Um problema comum em detecção de variações, fruto da inexatidão do modelo do fundo, é a detecção de falsos positivos, frequentemente referidos como *fantasmas*. Em [18] um *fantasma* é definido como o conjunto de pontos conectados detectados em movimento por meio da técnica de subtração de fundo, que não correspondem a nenhum objeto real que esteja se deslocando. Idealmente, o valor de θ na Equação (2.45), para os píxeis de um *fantasma*, são próximos a 0^o já que eles realmente pertencem ao fundo. Portanto, uma maneira de excluí-los é pela limiarização dos valores de θ, envolvendo uma nova definição para λ, denotada por $\tilde{\lambda}$ que é

$$\tilde{\lambda} = \begin{cases} 1 - \cos(\theta), & \text{se } \theta > \tau_\theta \\ 0, & \text{caso contrário} \end{cases}, \tag{2.54}$$

onde τ_θ é o limiar estabelecido a um valor menor ou igual a 10^o. Substituindo a Equação (2.54) na Equação (2.53) obtém-se

[24]No anexo D é feita uma demostração desta afirmação usando as coordenadas do modelo de cores HSB.

$$\rho_l(t) = \left(1 - \frac{\tilde{\lambda}}{2}\right) d_{0^o}. \tag{2.55}$$

A Figura 2.8 apresenta um determinado quadro e sua correspondente máscara do primeiro plano (as áreas da imagem sombreadas em vermelho) quando a métrica usada é definida pela Equação (2.53) (ver Figura 2.8.a) e quando a métrica usada é definida pela Equação (2.55) (ver Figura 2.8.b). Deve-se observar que quando θ é limiarizado os "fantasmas" são eliminados.

(a) (b)

Figura 2.8: Eliminação dos *fantasmas* quando a métrica é definida por: (a) $d_l(t)$ ou (b) $\rho_l(t)$.

Neste trabalho, o histograma dos valores da medida de similaridade das cores $\rho_l(t)$ é modelado como uma distribuição normal de média zero e variância σ_{ρ_t}. Esta distribuição, além de contar com expressões fechadas para a estimação de seus parâmetros em função do dado observado, permite modelar de maneira aproximada as caudas da suposta distribuição que governa a $\rho_l(t)$, sendo estas as que permitem estabelecer o limiar para a detecção de variações[25]. O parâmetro desconhecido σ_{ρ_t} é estimado utilizando os valores de $\rho_l(t)$ para todo píxel l do quadro t. Assim, se $\rho(t) = \{\rho_l(t)|l = 1, \cdots, N_{\text{fil}} \times N_{\text{col}}\}$ é considerado como o dado observado, então a maximização da verossimilhança $p(\rho(t)|\sigma_{\rho_t})$ gera os valores mais adequados para σ_{ρ_t}.

Um limiar τ é calculado para produzir um nível de significância α, ou seja $\mathbb{P}(\rho_l(t) < \tau) = 1 -$

[25]Quando $\rho_l(t)$ toma um valor próximo de 0 implica que o píxel l pertence a uma região estática. Este caso sucede com muita frequência, implicando que no histograma de $\rho(t)$ o maior número de ocorrências esteja concentrada na origem. O caso contrário, ou seja, quando $\rho_l(t)$ toma um valor elevado, implicará que l pertence a uma região do primeiro plano. Este caso não acontece constantemente, implicando que, no histograma de $\rho(t)$, as caudas (que representam o menor número de ocorrências) estarão vinculadas às regiões do primeiro plano.

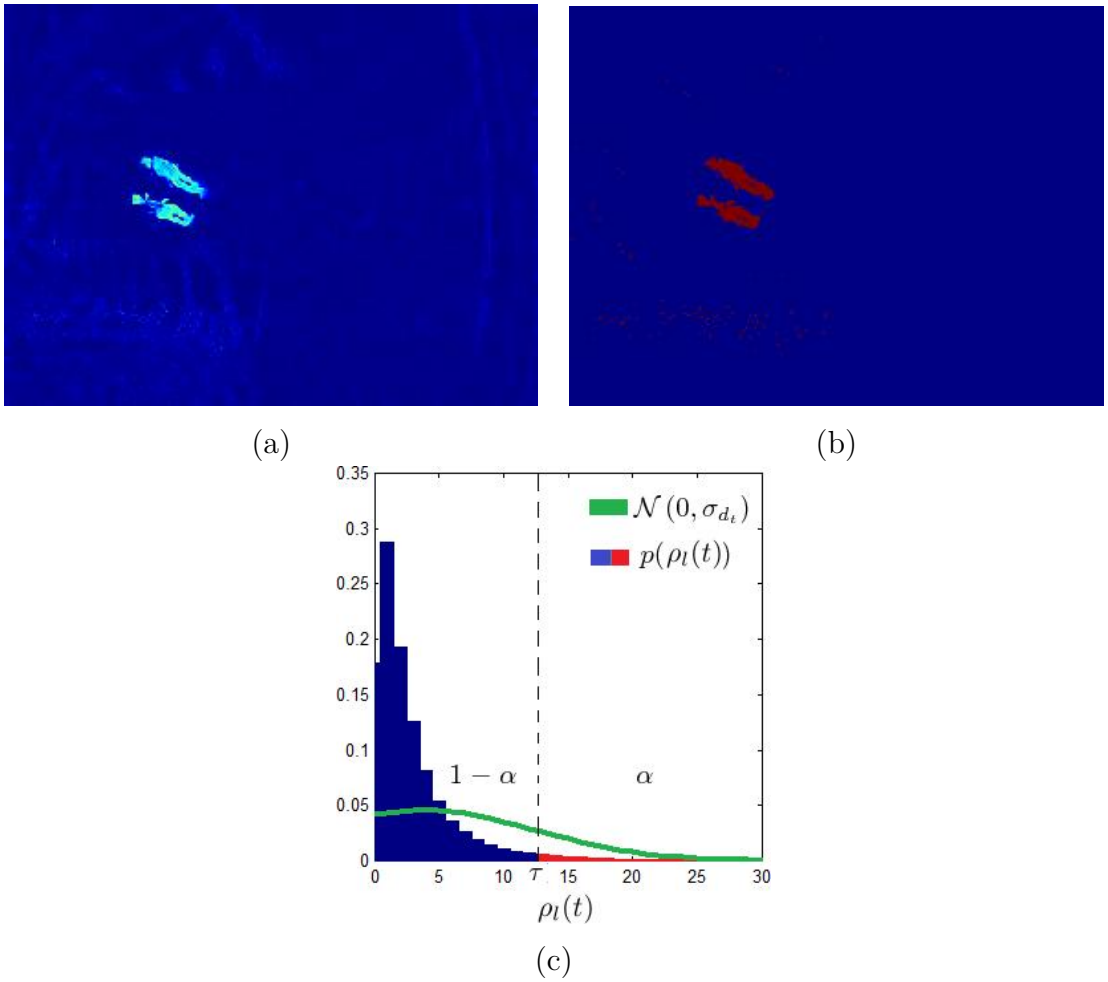


Figura 2.9: (a) $\rho(t)$, (b) $\mathbf{M}(t)$, (c) histograma de $\rho(t)$.

α [26] e, finalmente, a máscara do primeiro plano $\mathbf{M}_l(t)$ é calculada a partir da Equação (2.43). Um exemplo deste procedimento é apresentado na Figura 2.9, onde o histograma de $\rho(t)$ (ver Figura 2.9.c) é usado para estimar os parâmetros da distribuição normal, possibilitando assim calcular o limiar τ, o qual permite binarizar ρ (ver Figura 2.9.a) e obter a máscara $\mathbf{M}(t)$ (ver Figura 2.9.b).

2.8.2 Técnica baseada na Distância Euclidiana Bivariada

Uma generalização do procedimento anterior será apresentada a seguir, onde a distância Euclidiana é representada como uma variável bivariada, onda a primeira variável quantifica a diferença de magnitude (informação de brilho) e a segunda a diferença angular (informação da matiz) entre os píxeis do quadro atual e o modelo do fundo. Também é necessária a aplicação de operações de transformação lineares para chegar à sua definição.

As Equações (2.49) e (2.50) podem ser fatoradas como

$$\begin{aligned} d^2_{0^o} &= ||\mathbf{I}_l(t)||^2 + ||\mathbf{B}_l(t)||^2 - 2||\mathbf{I}_l(t)\mathbf{B}_l(t)|| \\ &= (||\mathbf{I}_l(t)|| - ||\mathbf{B}_l(t)||)^2, \qquad (2.56) \\ d^2_{90^o} &= ||\mathbf{I}_l(t)||^2 + ||\mathbf{B}_l(t)||^2 \\ &= \frac{1}{2}\left((||\mathbf{I}_l(t)|| + ||\mathbf{B}_l(t)||)^2 + (||\mathbf{I}_l(t)|| - ||\mathbf{B}_l(t)||)^2\right). \qquad (2.57) \end{aligned}$$

Substituindo as Equações (2.56) e (2.57) na Equação (2.51), obtém-se

$$\begin{aligned} d^2_l(t) &= \lambda d^2_{90^o} + (1-\lambda) d^2_{0^o} \\ &= d^2_{0^o} + \lambda(d^2_{90^o} - d^2_{0^o}) \\ &= \frac{\lambda}{2}(||\mathbf{I}_l(t)|| + ||\mathbf{B}_l(t)||)^2 + \left(1 - \frac{\lambda}{2}\right)(||\mathbf{I}_l(t)|| - ||\mathbf{B}_l(t)||)^2. \qquad (2.58) \end{aligned}$$

A magnitude do modelo do fundo $||\mathbf{B}_l(t)||$ atua como um valor de referência e, assim, para áreas onde existem objetos em movimento, $||\mathbf{I}_l(t)||$, apresenta um valor superior/inferior

[26] Neste caso, é considerado como hipótese nula, $\mathcal{H}_0$, que um píxel pertença ao fundo. Portanto, a probabilidade de falsos positivos é definida por $\mathbb{P}(\mathbf{M}_l(t) = 1|\mathcal{H}_0) = \alpha$ e, assim, $\mathbb{P}(\mathbf{M}_l(t) = 0|\mathcal{H}_0) = \mathbb{P}(\rho_l(t) < \tau) = 1 - \mathbb{P}(\mathbf{M}_l(t) = 1|\mathcal{H}_0) = 1 - \alpha$.

ao valor de referência definido por $||\mathbf{B}_l(t)||$. No caso contrário $||\mathbf{I}_l(t)||$ segue o nível de referência definido por $||\mathbf{B}_l(t)||$. Portanto, é válido supor que os valores de $||\mathbf{I}_l(t)||$ e $||\mathbf{B}_l(t)||$ estão relacionados pela equação

$$\kappa = |\ ||\mathbf{I}_l(t)|| - ||\mathbf{B}_l(t)||\ | = \begin{cases} ||\mathbf{I}_l(t)|| - ||\mathbf{B}_l(t)||, & ||\mathbf{I}_l(t)|| \geq ||\mathbf{B}_l(t)|| \\ ||\mathbf{B}_l(t)|| - ||\mathbf{I}_l(t)||, & ||\mathbf{I}_l(t)|| < ||\mathbf{B}_l(t)|| \end{cases}, \tag{2.59}$$

onde a variável κ quantifica a variação em relação à referência $||\mathbf{B}_l(t)||$. Ao substituir a Equação (2.59) na Equação (2.58), obtém-se

$$\begin{aligned} d_l^2(t) &= \begin{cases} \frac{\lambda}{2}(2||\mathbf{B}_l(t)|| + \kappa)^2 + \left(1 - \frac{\lambda}{2}\right)\kappa^2, & ||\mathbf{I}_l(t)|| \geq ||\mathbf{B}_l(t)|| \\ \frac{\lambda}{2}(2||\mathbf{I}_l(t)|| + \kappa)^2 + \left(1 - \frac{\lambda}{2}\right)\kappa^2, & ||\mathbf{I}_l(t)|| < ||\mathbf{B}_l(t)|| \end{cases} \\ &= \begin{cases} \kappa^2 + 2\lambda||\mathbf{B}_l(t)||\kappa + 2\lambda||\mathbf{B}_l(t)||^2, & ||\mathbf{I}_l(t)|| \geq ||\mathbf{B}_l(t)|| \\ \kappa^2 + 2\lambda||\mathbf{I}_l(t)||\kappa + 2\lambda||\mathbf{I}_l(t)||^2, & ||\mathbf{I}_l(t)|| < ||\mathbf{B}_l(t)|| \end{cases}. \end{aligned} \tag{2.60}$$

A partir da Equação (2.60), podem-se obter duas métricas que caracterizam a distância Euclidiana em função unicamente da diferença das magnitudes κ e a diferença angular λ correspondentes ao vetor $\mathbf{I}_l(t) - \mathbf{B}_l(t)$. Assim, tem-se que,

- supondo que os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ tenham a mesma magnitude, porém diferentes orientações ($\kappa = 0$, $\lambda \neq 0$),

$$d_l(t) = d_P(l,t) = \begin{cases} \sqrt{2\lambda}||\mathbf{B}_l(t)||, & ||\mathbf{I}_l(t)|| \geq ||\mathbf{B}_l(t)|| \\ \sqrt{2\lambda}||\mathbf{I}_l(t)||, & ||\mathbf{I}_l(t)|| < ||\mathbf{B}_l(t)|| \end{cases}; \tag{2.61}$$

- supondo que os vetores $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ tenham a mesma orientação, porém diferentes magnitudes ($\kappa \neq 0$, $\lambda = 0$),

$$d_l(t) = d_M(l,t) = \kappa = |\ ||\mathbf{I}_l(t)|| - ||\mathbf{B}_l(t)||\ |. \tag{2.62}$$

substituindo as Equações (2.61) e (2.62) na Equação (2.60), a distância Euclidiana é representada como

$$d_l^2(t) = d_M^2(l,t) + d_P^2(l,t) + \sqrt{2\lambda} d_M(l,t) d_P(l,t). \tag{2.63}$$

A partir de Equação (2.63) pode-se obter um limite inferior e superior para $d_l(t)$ em relação a $d_M(l,t) + d_P(l,t)$,

- o limite superior é obtido substituindo a cota superior de λ ($\lambda \in [0,2]$) na Equação (2.63), obtendo-se

$$d_l(t) \leq d_M(l,t) + d_P(l,t); \tag{2.64}$$

- o limite inferior é obtido levando em conta forma quadrática da Equação (2.63), sendo esta

$$d_l^2(t) = \frac{1}{2}\left(1 - \sqrt{\frac{\lambda}{2}}\right)(d_M(l,t) - d_P(l,t))^2 + \frac{1}{2}\left(1 + \sqrt{\frac{\lambda}{2}}\right)(d_M(l,t) + d_P(l,t))^2. \tag{2.65}$$

  Considerando que sempre $d_M(l,t), d_P(l,t) \geq 0$, e que o primeiro termo da soma definida na Equação (2.65) é sempre menor que o segundo termo, então

$$d_l(t) \geq \frac{1}{\sqrt{2}}\sqrt{1 + \sqrt{\frac{\lambda}{2}}}(d_M(l,t) + d_P(l,t)), \tag{2.66}$$

  e, levando-se em conta que $\lambda \in [0,2]$, a Equação (2.66) pode expressar-se como

$$d_l(t) \geq \frac{1}{\sqrt{2}}(d_M(l,t) + d_P(l,t)). \tag{2.67}$$

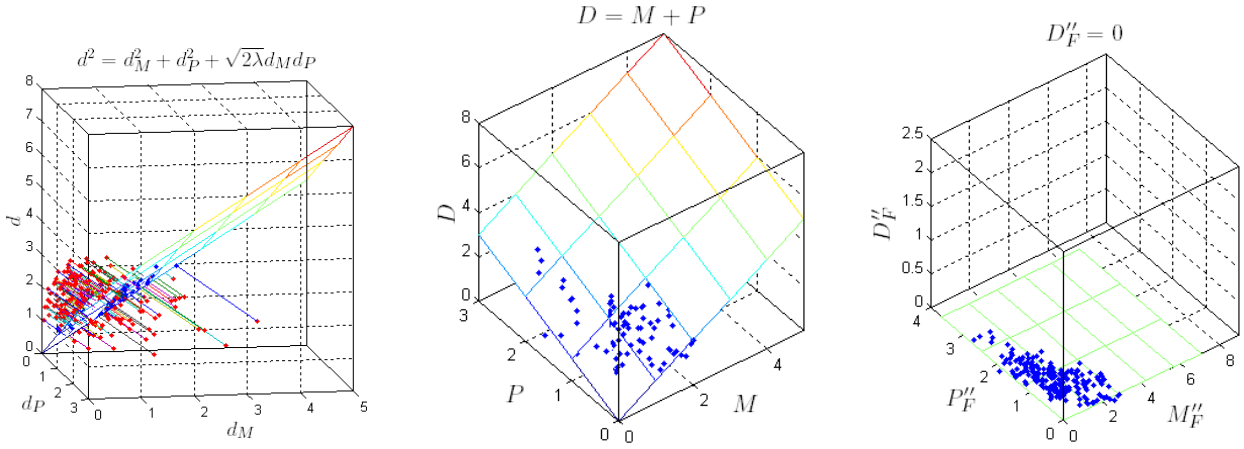
- Portanto, a partir das Equações 2.64 e 2.67, obtém-se:

$$\frac{1}{\sqrt{2}}(d_M(l,t) + d_P(l,t)) \leq d_l(t) \leq d_M(l,t) + d_P(l,t). \tag{2.68}$$

A Equação (2.68) mostra que os planos $d_l(t) = \frac{1}{\sqrt{2}}(d_M(l,t) + d_P(l,t))$ e $d_l(t) = d_M(l,t) + d_P(l,t)$ definem os limites inferior e superior para as distâncias $d_l(t)$. Se todo ponto $(d_l(t), d_M(l,t), d_P(l,t))$ é projetado no o plano $d_l(t) = d_M(l,t) + d_P(l,t)$ tal como é apresentado na Figura 2.10.a (gerando-se um erro não maior a $0,2929(d_M(l,t) + d_P(l,t))$[27]),

[27]Supondo que o ponto $(d_l(t), d_M(l,t), d_P(l,t))$ encontra-se no plano da cota mínima ($\frac{1}{\sqrt{2}}(d_M(l,t) + d_P(l,t))$) o erro de projetar tal ponto no plano $d_M(l,t) + d_P(l,t)$ é igual a $(d_M(l,t) + d_P(l,t)) - \frac{1}{\sqrt{2}}(d_M(l,t) + d_P(l,t)) = (1 - \frac{1}{\sqrt{2}})(d_M(l,t) + d_P(l,t)) = 0,2929(d_M(l,t) + d_P(l,t))$.

e logo são aplicadas operações de transformação lineares, o ponto $(d_l(t), d_M(l,t), d_P(l,t))$ (ver Figura 2.10.b) poderá representar-se num plano bidimensional equivalente (ver Figura 2.10.c). Esta transformação da forma quadrática definida pela Equação (2.63) num plano é obtida efetuando os seguintes passos:



(a) (b) (c)

Figura 2.10: Etapas da representação de um ponto $(d_l(t), d_M(l,t), d_P(l,t))$ num plano bidimensional equivalente (a) projeção dos pontos $(d_l(t), d_M(l,t), d_P(l,t))$ no plano $d_l(t) = d_M(l,t) + d_P(l,t)$, (b) aplicação de operações de transformação no plano $d_l(t) = d_M(l,t) + d_P(l,t)$, (c) representação bidimensional equivalente.

- Primeiro, todo ponto $(d_l(t), d_M(l,t), d_P(l,t))$ é projetado no plano $D(l,t) = M(l,t) + P(l,t)$, através da transformação linear

$$\begin{bmatrix} M(l,t) \\ P(l,t) \\ D(l,t) \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} d_M(l,t) \\ d_P(l,t) \\ d_l(t) \end{bmatrix}.$$

- Logo, a todo ponto do plano $D(l,t) = M(l,t) + P(l,t)$ é aplicada uma rotação em relação ao eixo P de 45^o (ver Figura 2.11.a), uma rotação em relação ao eixo M de -35.26^o (ver Figura 2.11.b) e uma transformação de cisalhamento (ver Figura 2.11.c) para eliminar a deformação que apresenta o último plano girado em relação ao eixo P''. Estas três transformações são compostas através da operação de multiplicação matricial

$$\begin{bmatrix} M''_F(l,t) \\ P''_F(l,t) \\ D''_F(l,t) \end{bmatrix} = \begin{bmatrix} 1 & -1/\sqrt{3} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{6}/3 & 1/\sqrt{3} \\ 0 & -1/\sqrt{3} & \sqrt{6}/3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} M(l,t) \\ P(l,t) \\ D(l,t) \end{bmatrix}$$
$$= \frac{\sqrt{6}}{3} \begin{bmatrix} \frac{1}{\sqrt{3}}(2M(l,t) - P(l,t) + D(l,t)) \\ \frac{1}{2}(2P(l,t) - M(l,t) + D(l,t)) \\ -\frac{1}{\sqrt{2}}(M(l,t) + P(l,t) - D(l,t)) \end{bmatrix}. \quad (2.69)$$

Já que $D(l,t) = M(l,t) + P(l,t)$, a expressão final é (ver Figura 2.11.d):

$$\begin{bmatrix} M''_F(l,t) \\ P''_F(l,t) \\ D''_F(l,t) \end{bmatrix} = \begin{bmatrix} \sqrt{2}M(l,t) \\ \frac{\sqrt{6}}{2}P(l,t) \\ 0 \end{bmatrix} \quad (2.70)$$

Assumindo que $M''_F(l,t)$ e $P''_F(l,t)$ são variáveis aleatórias com distribuições exponenciais de parâmetros $\mu_{M''_F}$ e $\mu_{P''_F}$, cujos valores são estimados a partir das distribuições marginais do gráfico de dispersão M''_F vs P''_F gerado para cada quadro, a saber,

$$p(M''_F) = \frac{1}{\mu_{M''_F}} \exp\left(-\frac{M''_F}{\mu_{M''_F}}\right), \quad (2.71)$$
$$p(P''_F) = \frac{1}{\mu_{P''_F}} \exp\left(-\frac{P''_F}{\mu_{P''_F}}\right), \quad (2.72)$$

é possível determinar a distribuição conjunta de $M''_F(l,t)$ e $P''_F(l,t)$ como

$$\begin{aligned} p(M''_F, P''_F) &= p(M''_F)p(P''_F) \\ &= \left(\frac{1}{\mu_{M''_F}} \exp\left(-\frac{M''_F}{\mu_{M''_F}}\right)\right)\left(\frac{1}{\mu_{P''_F}} \exp\left(-\frac{P''_F}{\mu_{P''_F}}\right)\right) \\ &= \frac{1}{\mu_{M''_F}\mu_{P''_F}} \exp\left(-\frac{1}{\mu_{M''_F}\mu_{P''_F}}\left(\mu_{P''_F}M''_F + \mu_{M''_F}P''_F\right)\right). \end{aligned} \quad (2.73)$$

A partir da Equação (2.73), observa-se que a distribuição conjunta depende de uma combinação linear das variáveis aleatórias M''_F e P''_F. Portanto, definindo-se

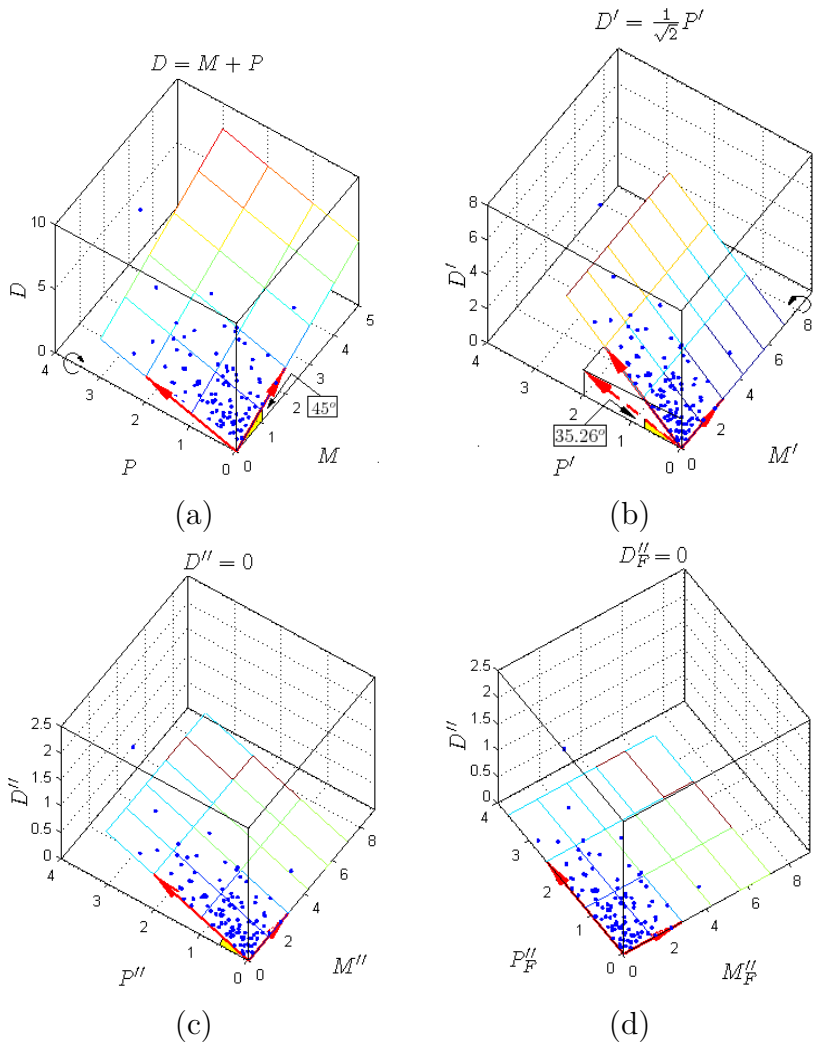


Figura 2.11: (a) Rotação do plano $D(l,t) = M(l,t) + P(l,t)$ em relação ao eixo P de 45^o, onde 45^o é a orientação do plano em relação aos eixos D e M . (b) Rotação em relação ao eixo M de $-35,26^o$, onde $35,26^o$ é a orientação do plano já rotacionado em relação aos eixos D' e P'. (c) aplicação de uma transformação de cisalhamento em relação ao eixo P'' por um fator de $1/\sqrt{3}$. (d) resultado da composição de transformações lineares afins aplicadas ao plano $D(l,t) = M(l,t) + P(l,t)$.

$$S = \mu_{P''_F} M''_F + \mu_{M''_F} P''_F, \tag{2.74}$$

$$\mu_S = \mu_{M''_F} \mu_{P''_F}, \tag{2.75}$$

a Equação (2.73) toma a forma

$$p(M''_F, P''_F) = p(S) = \frac{1}{\mu_S} \exp\left(-\frac{S}{\mu_S}\right). \tag{2.76}$$

Em forma equivalente ao caso inicial, o limiar τ é calculado para produzir um nível de significância α sobre a probabilidade $p(S)$, ou seja, $p(S < \tau) = 1 - \alpha$[28]. Este procedimento é equivalente a definir uma fronteira de separação linear no gráfico de dispersão M''_F versus P''_F. Finalmente, a máscara do primeiro plano $\mathbf{M}_l(t)$ é calculada empregando-se

$$\mathbf{M}_l(t) = \begin{cases} 1, & \text{se } S > \tau \\ 0, & \text{caso contrário} \end{cases}. \tag{2.77}$$

Um exemplo deste procedimento é apresentado na Figura 2.12, onde é plotado o gráfico de dispersão M''_F versus P''_F, as distribuições marginais $p(M''_F)$ e $p(P''_F)$ com suas respectivas distribuições exponenciais aproximadas, e é indicada a classificação dos pontos (M''_F, P''_F) em função do limiar τ, o qual permite obter a máscara $\mathbf{M}(t)$ através da Equação (2.77).

2.9 Resumo

Neste capítulo se expuseram os desafios a serem vencidos pelas técnicas de subtração de fundo, uma taxonomia das diferentes abordagens presentes na literatura, uma explicação detalhada de cada uma de suas partes constitutivas e das técnicas implementadas para cada uma delas (modelamento do fundo, detecção de variações e pós-processamento). Também, são propostas duas técnicas de detecção de variações, as quais estão baseadas (*a*) na distância Euclidiana calculada entre cada quadro do vídeo e o modelo do fundo, e (*b*) numa operação de limiarização, tal que o limiar é definido em relação a um nível de significância estatística.

[28]Neste caso, é considerado como hipótese nula, $\mathcal{H}_0$, que um píxel pertença ao fundo. Portanto a probabilidade de falsos positivos é definida por $\mathbb{P}(\mathbf{M}_l(t) = 1|\mathcal{H}_0) = \alpha$, e, assim, $\mathbb{P}(\mathbf{M}_l(t) = 0|\mathcal{H}_0) = \mathbb{P}(S < \tau) = 1 - \mathbb{P}(\mathbf{M}_l(t) = 1|\mathcal{H}_0) = 1 - \alpha$.

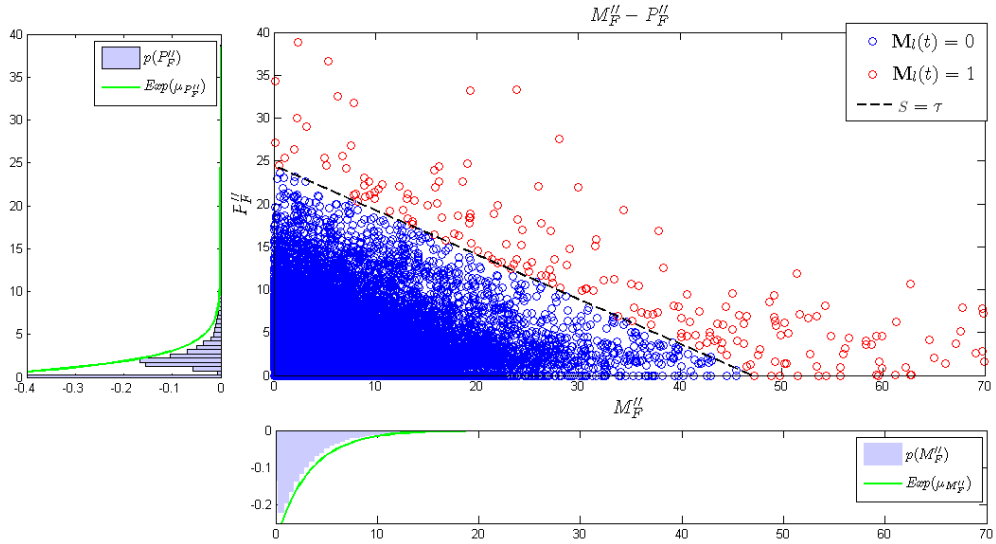


Figura 2.12: Gráfico de dispersão M''_F vs P''_F. Aqui, os pontos (M''_F, P''_F) de cor azul são classificados como fundo e os vermelhos são classificados como primeiro plano.

Capítulo 3

Avaliação das Técnicas de Subtração de Fundo

3.1 Introdução

Diversas abordagens têm sido propostas para avaliar o desempenho de um sistema de vigilância automatizado. Em alguns casos, a avaliação do desempenho é feita dando-se uma maior ênfase as técnicas de detecção de objetos [27][29][46][64][68], e em outros casos ela é feita de acordo com as técnicas de rastreamento de objetos [5][59][76]. Considerando-se que o resultado no passo de rastreamento de objetos depende fortemente da segmentação resultante do passo de detecção de objetos, a avaliação de uma técnica de subtração de fundo dentro de um sistema de vigilância automatizado desempenha um papel importante na análise do rendimento de todo o sistema.

A avaliação das técnicas de subtração de fundo não é uma tarefa trivial, e é, geralmente, realizada pela utilização de métodos de discrepância empíricos [80], também conhecidos como métodos de avaliação supervisionados [82]. Neste caso, a avaliação é realizada pela comparação entre as regiões vinculadas ao primeiro plano detectado e as regiões vinculadas ao primeiro plano definido na referência, também chamado *ground-truth*. Como resultado desta comparação, uma medida que indica o grau de sobreposição das duas regiões é obtido [81]. A definição desta medida implica considerar o critério de discrepância usado na avaliação. Por exemplo, o critério de discrepância pode ser definido como o número de píxeis mal classificados ou o número de objetos detectados corretamente contra o *ground-truth*, para cada quadro do vídeo.

Sendo assim, este capítulo se inicia definindo os valores e métricas de desempenho que caracterizam um problema de classificação binária, e na seção seguinte são apresentadas as diferentes propostas presentes na literatura para avaliar uma técnica de subtração de fundo. Na penúltima seção, é apresentada uma das contribuições deste trabalho, o uso da métrica da exatidão como uma métrica de desempenho de uma técnica de subtração de fundo. Finalmente, na última seção, é apresentado um resumo do capítulo. A principal contribuição exposta neste capítulo resume-se na determinação de um gráfico de desempenho que permite quantificar a taxa de acerto na detecção dos objetos presentes na cena, através de uma métrica que é calculada para todo quadro, para um vídeo ou para todo um banco de dados.

3.2 Definições

Considerando-se que para cada quadro de uma sequência de vídeo uma técnica de subtração de fundo gera uma máscara binária, a determinação do seu desempenho com base na comparação com a imagem do *ground-truth*, que especifica as verdadeiras áreas de mudança, pode ser representada como um problema de hipóteses nula ($\mathcal{H}_0$) e alternativa ($\mathcal{H}_1$). Neste caso, as métricas de desempenho de uma técnica de subtração de fundo são caracterizadas pelos valores definidos na matriz de confusão das hipótese nula e alternativa contra os resultados de um experimento (ver a Tabela 3.1). A matriz de confusão é útil quando se pretende determinar o desempenho de um problema de duas classes. No caso geral, estas classes são referidas como as classes positiva e negativa (para o problema de subtração de fundo as correspondentes classes são o primeiro plano e o fundo), e, dependendo das relações entre estas classes, os valores da matriz de confusão são definidos como:

- verdadeiro positivo (N_{VP}), que é o número de instâncias positivas que são classificadas como positivas;
- falso positivo (N_{FP}), que é o número de instâncias negativas que são classificadas como positivas;
- falso negativo (N_{FN}), que é o número de instâncias positivas que são classificadas como negativas;
- verdadeiro negativo (N_{VN}), que é o número de instâncias negativas que são classificadas como negativas.

Tabela 3.1: Matriz de confusão da hipótese nula e alternativa contra os resultados de um experimento.

		Realidade	
		$\mathcal{H}_0$ é falsa	$\mathcal{H}_0$ é verdadeira
Decisão	Rejeição da $\mathcal{H}_0$	Resultado correto N_{VP}	Erro do tipo I N_{FP}
	Não rejeição da $\mathcal{H}_0$	Erro do tipo II N_{FN}	Resultado correto N_{VN}

Segundo esta perspectiva, os dois tipos de erro presentes num problema de duas classes são: *a*) falso positivo, também conhecido como erro de tipo I, que ocorre quando as instâncias que devem ser classificadas como negativas (fundo) são classificadas como positivas (primeiro plano), *b*) falso negativo, também conhecido como erro de tipo II, que ocorre quando as instâncias que devem ser classificadas como positivas (primeiro plano) são classificadas como negativas (fundo). As métricas de desempenho que ponderam estes tipos de erro são definidas em termos das diferentes proporções entre as quatro quantidades N_{VP}, N_{FP}, N_{FN} e N_{VN}. As mais importantes são indicadas a seguir:

- taxa de verdadeiro positivo (m_{TVP}): essa métrica também é referida como a taxa de acerto ou sensibilidade, e indica a proporção de instâncias positivas que são corretamente classificadas como positivas, sendo calculada como

$$m_{TVP} = \frac{N_{VP}}{N_{VP} + N_{FN}}; \tag{3.1}$$

- taxa de falso positivo (m_{TFP}): essa métrica também é referida como a taxa de falso alarme, e indica a proporção de instâncias negativas que são erroneamente classificadas como positivas, sendo calculada como

$$m_{TFP} = \frac{N_{FP}}{N_{FP} + N_{VN}}; \tag{3.2}$$

- taxa de falso negativo (m_{TFN}): métrica que indica a proporção de instâncias positivas que são erroneamente classificadas como negativas ($= 1 - m_{TVP}$), sendo calculada como

$$m_{TFN} = \frac{N_{FN}}{N_{VP} + N_{FN}}; \tag{3.3}$$

- taxa de verdadeiro negativo (m_{TVN}): essa métrica também é referida como a especificidade, e indica a proporção de instâncias negativas que são corretamente classificadas como negativas, sendo calculada como

$$m_{TVN} = \frac{N_{VN}}{N_{FP} + N_{VN}}; \tag{3.4}$$

- valor preditivo positivo (m_{VPP}): essa métrica também é referida como precisão, e indica a proporção de instâncias positivas que são realmente positivas, sendo calculada como

$$m_{VPP} = \frac{N_{VP}}{N_{VP} + N_{FP}}; \tag{3.5}$$

- medida F (m_F): é a medida que combina as métricas m_{VPP} e m_{TVP} numa única medida, através do cálculo da média harmônica dos dois valores, a saber,

$$m_F = \frac{2 m_{VPP} m_{TVP}}{m_{VPP} + m_{TVP}}; \tag{3.6}$$

- exatidão (m_E): essa métrica também é referida como exatidão preditiva, e é a proporção de instâncias que são corretamente classificadas, sendo calculada como

$$m_E = \frac{N_{VP} + N_{VN}}{N_{VP} + N_{FN} + N_{FP} + N_{VN}}; \tag{3.7}$$

- taxa de erro (m_{ERR}): é a proporção de instâncias que são incorretamente classificadas, sendo calculada como

$$m_{ERR} = \frac{N_{FP} + N_{FN}}{N_{VP} + N_{FN} + N_{FP} + N_{VN}}; \tag{3.8}$$

- coeficiente de Jaccard (m_{JAC}): é a proporção de instâncias que são corretamente classificadas, desconsiderando as instâncias negativas, sendo calculada como

$$m_{JAC} = \frac{N_{VP}}{N_{VP} + N_{FN} + N_{FP}}. \tag{3.9}$$

3.3 Estado da Arte

No caso de uma técnica de subtração de fundo, o cálculo das métricas de desempenho definidas em função dos valores de N_{VP}, N_{FP}, N_{FN} e N_{VN} dependem do critério de discrepância

utilizado na avaliação. Assim, existem duas possibilidades de se efetuar o procedimento de avaliação: baseado em métricas orientadas a píxeis e baseado em métricas orientadas a objetos. Ambas propostas serão apresentadas a seguir.

3.3.1 Procedimento de avaliação baseado em métricas orientadas a píxeis

Aqui, é usado como critério de discrepância os píxeis classificados como primeiro plano. Neste caso, o problema da avaliação de uma técnica de subtração de fundo, em função de hipóteses nula e alternativa, é formulado da seguinte maneira:

- se é considerado que:
  - p_{fr} é rótulo do pixel l no quadro segmentado;
  - p_{gt} é rótulo do pixel l no quadro correspondente do *ground-truth*;
- então, as hipóteses são definidas como:
  - $\mathcal{H}_0 : p_{gt} = 0$: p_{gt} é um píxel do fundo;
  - $\mathcal{H}_1 : p_{gt} = 1$: p_{gt} é um píxel do primeiro plano,
- e as decisões associadas com cada uma das hipóteses definidas em função da resposta da técnica de subtração de fundo são:
  - $\mathcal{D}_0 : p_{fr} = 0$: p_{fr} é rotulado como fundo;
  - $\mathcal{D}_1 : p_{fr} = 1$: p_{fr} é rotulado como primeiro plano.

Assim, na Tabela 3.2 são indicados os valores dos elementos da matriz de confusão para um procedimento de avaliação baseado em métricas orientadas a píxeis considerando as hipóteses nula e alternativa contra os resultados gerados pela técnica de subtração de fundo.

Em [64][57][9] são avaliadas diferentes técnicas de subtração de fundo, onde os valores da matriz de confusão são definidos como:

- N_{VP}: o número de píxeis do primeiro plano corretamente detectados;

Tabela 3.2: Matriz de confusão para um problema de classificação binária, representando o procedimento de avaliação baseado em métricas orientadas a píxeis.

		Realidade	
		$\mathcal{H}_0$ é falsa $p_{gt} = 1$	$\mathcal{H}_0$ é verdade $p_{gt} = 0$
Decisão	Rejeição da $\mathcal{H}_0$ $\mathcal{D}_1 : p_{fr} = 1$	N_{VP}	N_{FP}
	Não rejeição da $\mathcal{H}_0$ $\mathcal{D}_0 : p_{fr} = 0$	N_{FN}	N_{VN}

- N_{FP}: o número de píxeis do fundo incorretamente rotulados como primeiro plano pela técnica em questão;
- N_{FN}: o número de píxeis do primeiro plano incorretamente rotulados como fundo pela técnica em questão;
- N_{VN}: o número de píxeis do fundo corretamente detectados;

e as métricas de desempenho utilizadas são definidas em termos das diferentes proporções entre as quatro quantidades N_{VP}, N_{FP}, N_{FN} e N_{VN}.

- Em [64], os autores propõem duas métricas chamadas erro absoluto e erro visual, definidas em função dos valores N_{FP}, N_{FN}, $N_{VP} + N_{FN}$ e $N_{FP} + N_{VN}$. Estas métricas são calculadas para alguns quadros de uma sequência de vídeo. A determinação do *ground-truth* de tais imagens é feita manualmente.
- Em [57], os autores utilizam a taxa de exatidão (m_E), o coeficiente de Jaccard (m_{JAC}) e o coeficiente de Yule[1] calculado sobre 4.000 imagens contendo uma única bola em movimento, utiliza-se um algoritmo simples de detecção de círculos para estabelecer o *ground-truth*.
- Em [9], a exatidão na detecção do primeiro plano é expressa por meio das taxas de sensibilidade (m_{TVP}) e precisão (m_{VPP}), e a medida F (m_F). Aqui, os autores propuseram um novo banco de dados gerado sinteticamente, próprio para a avaliação de técnicas de subtração de fundo. A vantagem de se usar um banco de dados sintético é que as anotações do *ground-truth* são exatas.

[1] É definido pela expressão, $|m_{VPP} + m_{TVN} - 1|$, e mede a correlação entre duas variáveis binarias.

3.3.2 Procedimento de avaliação baseado em métricas orientadas a objetos

Aqui, é usado como critério de discrepância o número de objetos detectados em movimento numa sequência de vídeo. Neste caso, se assume que todo objeto em movimento está vinculado a uma região do primeiro plano, onde a informação do *ground-truth* é representada em termos de retângulos (um retângulo vinculado a cada objeto em movimento na sequência de vídeo). O resultado da detecção geralmente é avaliado pela sobreposição do retângulo que contém o objeto detectado com o retângulo correspondente do *ground-truth*, tal que o grau de sobreposição entre os retângulos é um parâmetro do processo de avaliação que define o nível de associação de um objeto detectado com o *ground-truth*.

Tendo em conta o acima exposto, o problema da avaliação de uma técnica de subtração de fundo considerando como critério de discrepância o número de objetos detectados em movimento numa sequência de vídeo é formulado, em função das hipóteses nula e alternativa, da seguinte maneira:

- se é considerado que:
  - $\mathcal{P}$: é um grupo de píxeis conexos no quadro atual;
  - $\bar{\mathcal{R}}_{gt} = \{\mathcal{R}_{gt_i}\}_{i=1}^{N_{gt}}$: é o conjunto de retângulos do *ground-truth*;
  - $\bar{\mathcal{R}}_{dt} = \{\mathcal{R}_{dt_j}\}_{j=1}^{N_{dt}}$: é o conjunto de retângulos do quadro atual;
- então, as hipóteses definidas em função ao *ground-truth* são:
  - $\mathcal{H}_0 : \mathcal{P} \not\subset \bar{\mathcal{R}}_{gt}$ ($\mathcal{P}$ não está contido no conjunto de retângulos do *ground-truth*);
  - $\mathcal{H}_1 : \mathcal{P} \subset \bar{\mathcal{R}}_{gt}$ ($\mathcal{P}$ está contido no conjunto de retângulos do *ground-truth*);
- e as decisões associadas a cada uma das hipóteses definidas em função da resposta da técnica de subtração de fundo são:
  - $\mathcal{D}_0 : \mathcal{P} \not\subset \bar{\mathcal{R}}_{dt}$ ($\mathcal{P}$ não está contido no conjunto de retângulos do quadro atual);
  - $\mathcal{D}_1 : \mathcal{P} \subset \bar{\mathcal{R}}_{dt}$ ($\mathcal{P}$ está contido no conjunto de retângulos do quadro atual);

Assim, na Tabela 3.3 são indicados os valores dos elementos da matriz de confusão para um procedimento de avaliação baseado em métricas orientadas a objetos, considerando as hipóteses nula e alternativa, contra os resultados gerados pela técnica de subtração de fundo.

Em [46][29][39][47], os valores da matriz de confusão são definidos como:

Tabela 3.3: Matriz de confusão para um problema de classificação binária, representando o procedimento de avaliação baseado em métricas orientadas a objetos.

		Realidade	
		$\mathcal{H}_0$ é falsa $\mathcal{H}_1 : \mathcal{P} \subset \bar{\mathcal{R}}_{gt}$	$\mathcal{H}_0$ é verdade $\mathcal{H}_0 : \mathcal{P} \not\subset \bar{\mathcal{R}}_{gt}$
Decisão	Rejeição da $\mathcal{H}_0$ $\mathcal{D}_1 : \mathcal{P} \subset \bar{\mathcal{R}}_{dt}$	N_{VP}	N_{FP}
	Não rejeição da $\mathcal{H}_0$ $\mathcal{D}_0 : \mathcal{P} \not\subset \bar{\mathcal{R}}_{dt}$	N_{FN}	N_{VN}

- N_{VP} é o número de objetos detectados os quais são bem sobrepostos pelo *ground-truth*;
- N_{FP} é o número de objetos detectados que não foram sobrepostos pelo *ground-truth*;
- N_{FN} é o número de objetos do *ground-truth* que não estão suficientemente sobrepostos por um objeto detectado;
- N_{VN} para o caso das técnicas de avaliação baseadas em objetos é um valor que não é de utilidade, posto que não se conta com um conjunto de retângulos relativos ao fundo, ou seja, não se tem instâncias negativas no *ground-truth*, implicando que não é possível identificar verdadeiros negativos, o que não permite o cálculo do valor N_{VN}.

De forma equivalente ao caso anterior, as métricas de desempenho são também definidas em relação às quantidades N_{VP}, N_{FP}, N_{FN} e N_{VN}, tal como é indicado a seguir.

- Em [29], são determinadas as curvas das taxas de acerto (m_{TVP}) e falso alarme (m_{TFP}) contra vários valores do parâmetro de sobreposição, e é utilizada como uma medida do desempenho a área debaixo dessas curvas.
- Em [39], os autores propõem usar a metodologia de análise de eficácia (ou análise baseada na medida F) para evitar o problema da ausência de instâncias negativas (um valor não definido para N_{VN}). Aqui, os cálculos das taxas de precisão (m_{VPP}) e sensibilidade (m_{TVP}) são feitos para diferentes valores do parâmetro de sobreposição. Também é definido um parâmetro que controla a importância relativa das taxas de precisão e sensibilidade.
- Em [46] e [47], é analisado o caso onde existem vários retângulos detectados pela técnica de subtração de fundo e presentes no *ground-truth*. Neste caso, a avaliação

não é formulada como um problema de classificação binária, uma vez que é necessário ter em conta outros tipos de erros (além dos erros tipo I (N_{FP}) e tipo II (N_{FN})), que quantificam os erros devidos a todos os tipos possíveis de casamentos. Assim, os autores definiram:

- casamento um a zero: um retângulo do *ground-truth* não tem casamento. Este tipo de situação também é referido como Falha na Detecção (FD), e seu número de ocorrências é denotado pela variável N_{FD};
- casamento zero a um: um retângulo de um objeto detectado não tem casamento. Este tipo de casamento também é referido como Falso Alarme (FA), e seu número de ocorrências é denotado pela variável N_{FA};
- casamento um a um: um retângulo do *ground-truth* é casado com um retângulo de um objeto detectado, situação essa conhecida como Correta Detecção (CD), e seu número de ocorrências é denotado pela variável N_{CD};
- casamento um a muitos: um retângulo do *ground-truth* é casado com vários retângulos vinculados a diferentes objetos detectados. Este tipo de casamento também é referido como Separação (S), e seu número de ocorrências é denotado pela variável N_S;
- casamento muitos a um: vários retângulos do *ground-truth* são casados com um retângulo de um objeto detectado. Este tipo de casamento também é referido como União (U), e seu número de ocorrências é denotado pela variável N_U;
- casamento muitos a muitos: quando acontece simultaneamente uma separação e uma união. Este tipo de casamento também é referido como Separação - União (SU), e seu número de ocorrências é denotado pela variável N_{SU}.

O grau de sobreposição entre os retângulos presentes no *ground-truth* e os retângulos presentes nos quadros da sequência de vídeo foi calculado utilizando a medida de Pascal, definida como

$$\text{Pascal}(\mathcal{R}_{gt_i}, \mathcal{R}_{dt_j}) = \frac{\text{área}(\mathcal{R}_{gt_i} \cap \mathcal{R}_{dt_j})}{\text{área}(\mathcal{R}_{gt_i} \cup \mathcal{R}_{dt_j})}, \tag{3.10}$$

onde $\mathcal{R}_{gt_i}$ é o i-ésimo retângulo referente ao i-ésimo objeto presente no *ground-truth* e $\mathcal{R}_{dt_j}$ é o j-ésimo retângulo referente ao j-ésimo objeto detectado. Esta medida penaliza uma sub-segmentação[2] (ver Figura 3.1.a) e uma super-segmentação[3] (ver Figura 3.1.b). Finalmente, é definida uma matriz de decisão $\mathbf{D}_{\text{decs}}$, que indica os casamentos entre os retângulos presentes no quadro atual e no *ground-truth*, como

[2]É gerar uma segmentação incompleta, ou seja, que o fundo contenha regiões do primeiro plano.
[3]É gerar uma segmentação extra, ou seja, que o primeiro plano contenha regiões do fundo.

$$\mathbf{D}_{\text{decs}}(i,j) = \begin{cases} 1, & \text{Pascal}(\mathcal{R}_{gt_i}, \mathcal{R}_{dt_j}) > \tau_{\text{casamento}} \\ 0, & \text{Pascal}(\mathcal{R}_{gt_i}, \mathcal{R}_{dt_j}) < \tau_{\text{casamento}} \end{cases}, \tag{3.11}$$

onde $\tau_{\text{casamento}} \in [0,1]$ é o parâmetro de sobreposição que estabelece qual é o grau de sobreposição entre dois retângulos para que eles sejam casados. $\tau_{\text{casamento}}$ é um parâmetro no processo de avaliação, e será estudado com detalhe nas secções posteriores. Finalmente, o cálculo dos diferentes tipos de casamentos através da matriz de decisão $\mathbf{D}_{\text{decs}}$ é feito da seguinte maneira:

– primeiro é calculado o número de 1's em cada linha ou coluna de $\mathbf{D}_{\text{decs}}$, para o qual é necessário definir dois vetores auxiliares:

$$\mathbf{D}_{\text{decs}}{}^{L}(i) = \sum_{j=1}^{N_{dt}} \mathbf{D}_{\text{decs}}(i,j) \;\; i = 1, \cdots, N_{gt}, \tag{3.12}$$

$$\mathbf{D}_{\text{decs}}{}^{C}(j) = \sum_{i=1}^{N_{gt}} \mathbf{D}_{\text{decs}}(i,j) \;\; j = 1, \cdots, N_{dt}; \tag{3.13}$$

– logo, são calculadas as ocorrências de casamentos, segundo seu tipo, através das relações:

$$N_{FA} = |\mathcal{S}_{FA}| \;,\; \mathcal{S}_{FA} = \{\mathbf{D}_{\text{decs}}{}^{L}(i) = 0\}, \tag{3.14}$$

$$N_{FD} = |\mathcal{S}_{FD}| \;,\; \mathcal{S}_{FD} = \{\mathbf{D}_{\text{decs}}{}^{C}(i) = 0\}, \tag{3.15}$$

$$N_{CD} = |\mathcal{S}_{CD}| \;,\; \mathcal{S}_{CD} = \{\mathbf{D}_{\text{decs}}{}^{L}(i) = \mathbf{D}_{\text{decs}}{}^{C}(j) = 1 \wedge \mathbf{D}_{\text{decs}}(i,j) = 1\}, \tag{3.16}$$

$$N_{S} = |\mathcal{S}_{S}| \;,\; \mathcal{S}_{S} = \{\mathbf{D}_{\text{decs}}{}^{L}(i) > 1 \wedge \mathbf{D}_{\text{decs}}(i,j) = 1\}, \tag{3.17}$$

$$N_{U} = |\mathcal{S}_{U}| \;,\; \mathcal{S}_{U} = \{\mathbf{D}_{\text{decs}}{}^{C}(j) > 1 \wedge \mathbf{D}_{\text{decs}}(i,j) = 1\}, \tag{3.18}$$

$$N_{SU} = |\mathcal{S}_{SU}| \;,\; \mathcal{S}_{SU} = \{\mathbf{D}_{\text{decs}}{}^{L}(i) > 1 \wedge \mathbf{D}_{\text{decs}}{}^{C}(j) > 1 \wedge \mathbf{D}_{\text{decs}}(i,j) = 1\}, \tag{3.19}$$

onde $|\bullet|$ indica a cardinalidade do conjunto em questão.

Resumindo, num procedimento de avaliação baseado em métricas orientadas a píxeis, o objetivo é detectar todos os píxeis ativos para cada quadro de uma sequência de vídeo, implicando que a avaliação seja formulada como um problema de classificação binária. Por outro lado, num procedimento de avaliação baseado em métricas orientadas a objetos considera-se que cada objeto em movimento está ligado a uma região do primeiro plano, estabelecendo-se uma correspondência entre os objetos detectados e o *ground-truth*. Assim, no caso geral, devem ser considerados outros tipos de erros, o que implica que não se pode definir a avaliação como um problema de classificação binária. Mas, esta abordagem de validação tem dois pontos principais que devem ser destacados: (*a*) do ponto de vista do usuário, uma forma mais natural de avaliar uma técnica de subtração de fundo é determinar quantos objetos

foram detectados corretamente; (*b*) a geração manual do *ground-truth* é um processo demorado. Portanto, definir o *ground-truth* como um conjunto de retângulos associados com os objetos em movimento é mais simples do que determinar com precisão os contornos dos objetos em movimento numa sequência de vídeo[4].

(a) (b)

Figura 3.1: Aparição de: (a) uma sub-segmentação, (b) uma super-segmentação, onde, o retângulo de cor mais clara (verde) corresponde ao objeto detectado pela técnica e o retângulo do cor mais escura (azul) corresponde ao objeto marcado no *ground-truth*.

Tomando em conta o exposto, na próxima seção propõe-se um procedimento para quantificar o desempenho de uma técnica de subtração de fundo usando um procedimento de avaliação baseado em métricas orientadas a objetos.

3.4 Proposta

Do exposto, observa-se que um procedimento de avaliação baseado em métricas orientadas a objetos apresenta dois grandes problemas: (*a*) ao considerar-se o caso de múltiplos objetos detectados na cena e presentes no *ground-truth*, a avaliação é definida como um problema de classificação de múltiplas classes; (*b*) e como não se conta com um conjunto de retângulos relativos ao fundo, tem-se uma ausência de instâncias negativas (N_{VN}) no *ground-truth*. O primeiro problema foi contornado em [47], ao definir os tipos de casamento possíveis neste tipo de avaliação, ou seja, a determinação dos valores de: falha na detecção (N_{FD}), falso alarme (N_{FA}), correta detecção (N_{CD}), separação (N_S), união (N_U) e separação-união (N_{SU}). Porém, uma métrica que pondere o desempenho considerando a avaliação como um problema de classificação de múltiplas classes não é apresentada. O segundo problema exclui totalmente o uso da curva ROC como uma ferramenta para medir e especificar problemas

[4]A geração de uma máscara binária baseada no contorno dos objetos é um processo custoso, se ela é feita com precisão.

no desempenho de uma técnica de subtração de fundo[5] e para o caso da utilização da curva de precisão-sensibilidade experimentalmente foi observado nessa tese que a precisão e a sensibilidade não apresentam uma relação inversa entre elas[6], implicando que a correspondente medida F não oferecerá informação representativa do desempenho da técnica de detecção[7].

Considerando as restrições tanto da curva de precisão-sensibilidade quanto da curva ROC, nessa tese é proposto o uso da métrica de exatidão definida para um problema de classificação de múltiplas classes como uma métrica válida para a realização de uma avaliação baseada em objetos. Como gráfico de análise, propõe-se a curva de exatidão em relação ao parâmetro $\tau_{\text{casamento}}$, tal que, a partir deste gráfico, são determinados valores para $\tau_{\text{casamento}}$ que ponderam o fato de se considerar tanto uma pequena como uma grande sobreposição entre retângulos.

Métrica de Exatidão

Ao considerar-se o procedimento de avaliação baseado em métricas orientadas a objetos como um problema de classificação de múltiplas classes, onde as classes são definidas pelos diferentes tipos de casamentos, $\{CD, FD, FA, S, U, SU\}$, as métricas de desempenho da avaliação serão caracterizadas pelos valores definidos na matriz de confusão de múltiplas entradas representada na Tabela 3.4. Neste caso, o classificador sempre prediz instâncias positivas, já que o *ground-truth* contém unicamente objetos pertencentes ao primeiro plano. Portanto, uma métrica que oferece informação de importância é a métrica de exatidão, definida como

$$m_E = \frac{N_{CD}}{N_{CD} + N_{FD} + N_{FA} + N_S + N_U + N_{SU}}, \tag{3.20}$$

que define a proporção dos retângulos corretamente detectados em relação ao número total

[5]Na literatura tem-se utilizado a curva ROC para a avaliação de algoritmos de detecção de movimento [27][47][50]. No entanto, as implementações destas abordagens são discutíveis. Primeiro, elas são restritas para o nível de píxeis [27][50]. Segundo, elas nem sempre descrevem um método para selecionar o ponto de operação ótimo [47], que é uma das características principais da curva ROC, ao comparar diferentes técnicas.

[6]Essa é uma relação característica entre a precisão e a sensibilidade. Por exemplo, em sistemas de recuperação de informação pode-se incrementar a sensibilidade recuperando-se mais documentos, porém o número de documentos irrelevantes também aumenta (decaimento da precisão).

[7]A medida F (média harmônica) é utilizada quando se tem observações de grandezas inversamente proporcionais.

de retângulos detectados e não detectados.

Tabela 3.4: Matriz de confusão para um problema de classificação de múltiplas classes, representando o procedimento de avaliação baseado em métricas orientadas a objetos.

		Realidade					
		CD	FD	FA	S	U	SU
Decisão	CD	N_{CD}	0	0	0	0	0
	FD	N_{FD}	0	0	0	0	0
	FA	N_{FA}	0	0	0	0	0
	S	N_S	0	0	0	0	0
	U	N_U	0	0	0	0	0
	SU	N_{SU}	0	0	0	0	0

Uma característica da importância dos valores $N_{CD}, N_{FD}, N_{FA}, N_S, N_U$ e N_{SU} é que eles podem ser calculados para toda uma sequência de quadros de um vídeo, já que por serem valores de contagens (eles contam o número de casamentos válidos ou não) eles podem ser acumuláveis, não perdendo sua significância. Por exemplo, a soma dos N_{CD} para diferentes quadros de um vídeo sempre representa o número de casamentos válidos no vídeo. Como a ideia pode ser considerada para todo o banco de dados, é válido definir as medidas $N_{CD}, N_{FD}, N_{FA}, N_S, N_U$ e N_{SU} para cada vídeo e para todo o banco de dados, como

$$N_{CD_v} = \sum_{fr=1}^{N_{\text{quadros}}} N_{CD_{fr}} \qquad N_{CD_{db}} = \sum_{v=1}^{N_{\text{vídeos}}} N_{CD_v}, \tag{3.21}$$

$$N_{FD_v} = \sum_{fr=1}^{N_{\text{quadros}}} N_{FD_{fr}} \qquad N_{FD_{db}} = \sum_{v=1}^{N_{\text{vídeos}}} N_{FD_v}, \tag{3.22}$$

$$N_{FA_v} = \sum_{fr=1}^{N_{\text{quadros}}} N_{FA_{fr}} \qquad N_{FA_{db}} = \sum_{v=1}^{N_{\text{vídeos}}} N_{FA_v}, \tag{3.23}$$

$$N_{S_v} = \sum_{fr=1}^{N_{\text{quadros}}} N_{S_{fr}} \qquad N_{S_{db}} = \sum_{v=1}^{N_{\text{vídeos}}} N_{S_v}, \tag{3.24}$$

$$N_{U_v} = \sum_{fr=1}^{N_{\text{quadros}}} N_{U_{fr}} \qquad N_{U_{db}} = \sum_{v=1}^{N_{\text{vídeos}}} N_{U_v}, \tag{3.25}$$

$$N_{SU_v} = \sum_{fr=1}^{N_{\text{quadros}}} N_{SU_{fr}} \qquad N_{SU_{db}} = \sum_{v=1}^{N_{\text{vídeos}}} N_{SU_v}, \tag{3.26}$$

onde N_{quadros} é o número de quadros de um vídeo específico, e $N_{\text{vídeos}}$ é o número de vídeos

presentes no banco de dados em questão. Seguindo o mesmo raciocínio, também é válido calcular as correspondentes definições da exatidão para cada vídeo e para todo o banco de dados, a saber,

$$m_{E_v} = \frac{N_{CD_v}}{N_{CD_v} + N_{FD_v} + N_{FA_v} + N_{S_v} + N_{U_v} + N_{SU_v}}, \tag{3.27}$$

$$m_{E_{db}} = \frac{N_{CD_{db}}}{N_{CD_{db}} + N_{FD_{db}} + N_{FA_{db}} + N_{S_{db}} + N_{U_{db}} + N_{SU_{db}}}. \tag{3.28}$$

3.5 Análise Baseada na Curva de Exatidão

Qualquer técnica de subtração de fundo é governada por um conjunto de parâmetros (parâmetros da técnica), os quais são ajustados para que ela tenha um desempenho aceitável. Da mesma forma, um procedimento de avaliação baseado em métricas orientadas a objetos também requer um conjunto de parâmetros (parâmetros de avaliação), os quais determinam o grau de correspondência entre os objetos detectados e os objetos do *ground-truth* (um sistema típico para a avaliação do desempenho de uma técnica de subtração de fundo toma a sua saída e a compara com a informação do *ground-truth*, como é ilustrado na Figura 3.2). No caso estudado neste trabalho, o parâmetro de avaliação é o limiar $\tau_{\text{casamento}}$. Como já foi indicado, ele estabelece o grau de sobreposição entre dois retângulos. Portanto, seu valor afeta em grande medida à resposta da avaliação. Se for considerado como uma métrica válida do desempenho a exatidão da técnica, m_E, definida pela Equação (3.20), observa-se que ela é uma função dependente de $\tau_{\text{casamento}}$[8], e essa dependência é representada na curva m_E vs $\tau_{\text{casamento}}$. Um ponto de interesse é estabelecer um procedimento para determinar um valor adequado para o limiar $\tau_{\text{casamento}}$ que pondere o fato de considerar tanto uma pequena como uma grande sobreposição entre retângulos.

Experimentalmente, observa-se que a curva m_E vs $\tau_{\text{casamento}}$ tem uma forma de "S" (ver Figura 3.4), e em alguns casos é equivalente a uma função logística[9], implicando que o processo

[8] A métrica m_E depende dos valores de N_{CD}, N_{FD}, N_{FA}, N_S, N_U eN_{SU}, tal que, cada um destes valores são dependentes da matriz de decisão $\mathbf{D}_{\text{decs}}$ como é indicado nas Equações 3.14 - 3.19, porém todo elemento (i, j) em $\mathbf{D}_{\text{decs}}$ é dependente de $\tau_{\text{casamento}}$ como é indicado na Equação (3.11), por tanto, m_E será dependente de $\tau_{\text{casamento}}$.

[9] A expressão de uma função logística é definida pela fórmula matemática

$$f(x) = a\frac{1 + me^{-x/\tau}}{1 + ne^{-x/\tau}}$$

, para parâmetros reais a, m, n, e τ. Esta função tem um campo de aplicação muito amplo, desde a biologia

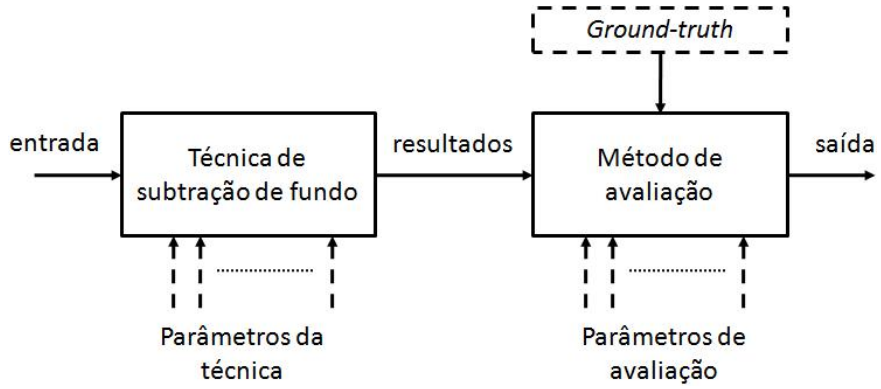


Figura 3.2: Um típico sistema de avaliação.

de casamentos entre a saída da técnica e o *ground-truth* tenha um comportamento equivalente a um sistema de competição entre duas alternativas, ou a um modelo de crescimento exponencial com recursos limitados [70]. Indícios que validam esta observação são dados a seguir: (a) o máximo número de casamentos nunca será superior ao número de retângulos detectados e presentes no *ground-truth*; (b) à medida em que $\tau_{\text{casamento}}$ aumenta de valor, a quantidade de casamentos vai diminuindo (se requer um maior grau de sobreposição entre os retângulos para que seja considerado um casamento válido) até alcançar o mínimo número de casamentos possíveis, ou seja zero.

Pelo exposto, aqui é considerado que a curva m_E vs $\tau_{\text{casamento}}$ é aproximada por uma função logística, e é proposto analisar o comportamento dela em função de seus pontos de inflexão (ver Figura 3.3.a), definidos como τ_L e τ_C. Considerando que $\tau_L < \tau_C$, tem-se que:

- o limiar τ_L localizado no lado esquerdo da curva m_E vs $\tau_{\text{casamento}}$ é nomeado de limiar liberal, já que ao se tomar valores próximos a $\tau_{\text{casamento}} = 0$, estabelece-se como válido um casamento (entre um objeto detectado e um retângulo do *ground-truth*) considerando um valor baixo de sobreposição, o que implica num relaxamento na exigência da avaliação;
- o limiar τ_C localizado no lado direito da curva m_E vs $\tau_{\text{casamento}}$ é nomeado de limiar conservador, já que ao se tomar valores próximos a $\tau_{\text{casamento}} = 1$, estabelece-se como válido um casamento(entre um objeto detectado e um retângulo do *ground-truth*) considerando um valor alto de sobreposição, o que implica num aumento da exigência da avaliação.

Levando-se em conta estes comportamentos, aqui é proposto usar a média entre ambos

à economia[36].

limiares, como um valor de limiar razoável, em relação à exigência na avaliação. Assim, é definido o limiar τ_R como

$$\tau_R = \frac{\tau_L + \tau_C}{2}. \tag{3.29}$$

Já que τ_L e τ_C estão vinculados aos pontos de inflexão da curva, eles são calculados a partir da derivada de m_E (ver Figura 3.3.b). Considerando que ela é aproximada por uma função logística, sua derivada é estabelecida pela expressão

$$\frac{dm_E}{d\tau_{\text{casamento}}} \approx m_E(1 - m_E), \tag{3.30}$$

e os valores de τ_L e τ_C serão obtidos ao resolver a Equação (3.31) numericamente.

$$\left.\frac{dm_E}{d\tau_{\text{casamento}}}\right|_{\tau_{\text{casamento}}=\{\tau_L,\tau_C\}} = 0, \tag{3.31}$$

Nas Figuras 3.4 e 3.5 são apresentadas as curvas m_{E_v} vs $\tau_{\text{casamento}}$ e $dm_{E_v}/d\tau_{\text{casamento}}$ vs $\tau_{\text{casamento}}$, respectivamente, para todos os vídeos do banco de dados PETS2004. Nelas, pode-se observar o comportamento já descrito e também o intervalo formado pelos limiares τ_L e τ_C, indicado pela área sombreada, além da linha de τ_R (linha vermelha).

Finalmente, no Algoritmo 1 são detalhados todos os passos necessários para a determinação do desempenho de uma técnica de subtração de fundo baseada na métrica de exatidão m_E, onde

- a função `AlgDetecObj` contém a implementação da técnica de subtração de fundo, onde os valores de seus parâmetros são representados por `ParamsAlg` e a saída esperada são os retângulos presentes no quadro atual, $\bar{\mathcal{R}}_{dt}$;
- a função `ExtracGroundTruth` extrai os retângulos vinculados ao *ground-truth* do quadro atual, $\bar{\mathcal{R}}_{gt}$;
- a função `CalNumCasamentos` calcula a ocorrência de casamentos determinando os valores para $N_{CD_f}, N_{FD_f}, N_{FA_f}, N_{S_f}, N_{U_f}$ e N_{SU_f}.

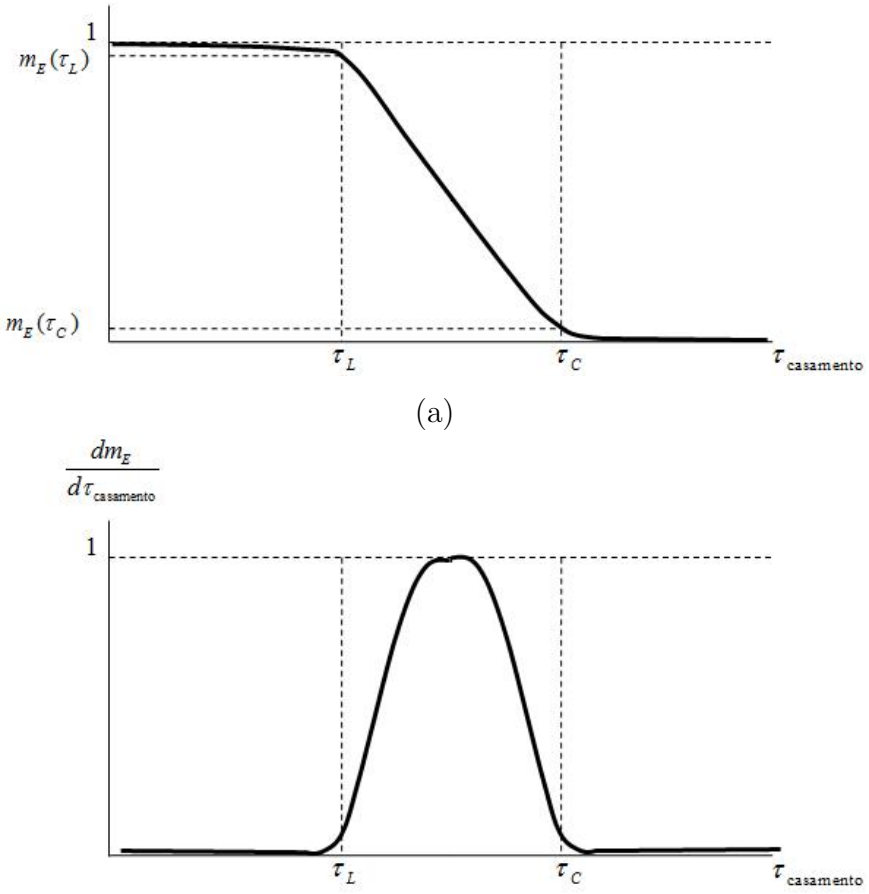


Figura 3.3: (a) Curva m_E ideal e (b) sua correspondente derivada.

Algoritmo 1: Cálculo da métrica de exatidão em nível de objetos.

Input: O conjunto de vídeos do banco de dados em questão $\{\mathbf{I}_{1,v}, \cdots, \mathbf{I}_{N_{\text{quadros}}(v),v}\}_{v=1}^{N_{\text{vídeos}}}$.
O conjunto de parâmetros da técnica de subtração de fundo, ParamsAlg.
Um conjunto de limiares $\{\tau_{\text{casamento}}\}$.

Output: As métricas de exatidão $\{m_{E_v}\}_{v=1}^{N_{\text{vídeos}}}$ e $m_{E_{db}}$ calculadas para cada valor do limiar $\tau_{\text{casamento}}$.

foreach $\tau_{casamento}$ **do** /* para cada limiar */
  $N_{CD_{db}} \leftarrow 0$; $N_{FA_{db}} \leftarrow 0$; $N_{FD_{db}} \leftarrow 0$;
  $N_{U_{db}} \leftarrow 0$; $N_{S_{db}} \leftarrow 0$; $N_{SU_{db}} \leftarrow 0$;
  for $v \leftarrow 1$ **to** $N_{vídeos}$ **do** /* para cada vídeo */
    $N_{CD_v} \leftarrow 0$; $N_{FA_v} \leftarrow 0$; $N_{FD_v} \leftarrow 0$;
    $N_{U_v} \leftarrow 0$; $N_{S_v} \leftarrow 0$; $N_{SU_v} \leftarrow 0$;
    for $f \leftarrow 1$ **to** $N_{quadros}(v)$ **do** /* para cada quadro do vídeo */
      // extração dos retângulos detectados pelo algoritmo
      $\bar{\mathcal{R}}_{dt} \leftarrow$ AlgDetecObj($\boldsymbol{I}_{f,v}$,ParamsAlg);
      // extração dos retângulos do *ground-truth*
      $\bar{\mathcal{R}}_{gt} \leftarrow$ ExtracGroundTruth($\boldsymbol{I}_{f,v}$);
      // determinação do número de casamentos segundo seu tipo
      $\{N_{CD_f}, N_{FD_f}, N_{FA_f}, N_{S_f}, N_{U_f}, N_{SU_f}\} \leftarrow$ CalNumCasamentos($\bar{\mathcal{R}}_{dt}$,$\bar{\mathcal{R}}_{gt}$,$\tau_{casamento}$);
      // medidas para todo um vídeo
      $N_{CD_v} \leftarrow N_{CD_v} + N_{CD_f}$; $N_{FA_v} \leftarrow N_{FA_v} + N_{FA_f}$; $N_{FD_v} \leftarrow N_{FD_v} + N_{FD_f}$;
      $N_{U_v} \leftarrow N_{U_v} + N_{U_f}$; $N_{S_v} \leftarrow N_{S_v} + N_{S_f}$; $N_{SU_v} \leftarrow N_{SU_v} + N_{SU_f}$;
    // medidas para todo um banco de dados
    $N_{CD_{db}} \leftarrow N_{CD_{db}} + N_{CD_v}$; $N_{FA_{db}} \leftarrow N_{FA_{db}} + N_{FA_v}$; $N_{FD_{db}} \leftarrow N_{FD_{db}} + N_{FD_v}$;
    $N_{U_{db}} \leftarrow N_{U_{db}} + N_{U_{db}}$; $N_{S_{db}} \leftarrow N_{S_{db}} + N_{S_{db}}$; $N_{SU_{db}} \leftarrow N_{SU_{db}} + N_{SU_{db}}$;
    // métrica de exatidão para cada vídeo
    $m_{E_v}(\tau_{\text{casamento}}) \leftarrow \frac{N_{CD_v}}{N_{CD_v}+N_{FD_v}+N_{FA_v}+N_{S_v}+N_{U_v}+N_{SU_v}}$;
  // métrica de exatidão para o banco de dados
  $m_{E_{db}}(\tau_{\text{casamento}}) \leftarrow \frac{N_{CD_{db}}}{N_{CD_{db}}+N_{FD_{db}}+N_{FA_{db}}+N_{S_{db}}+N_{U_{db}}+N_{SU_{db}}}$;

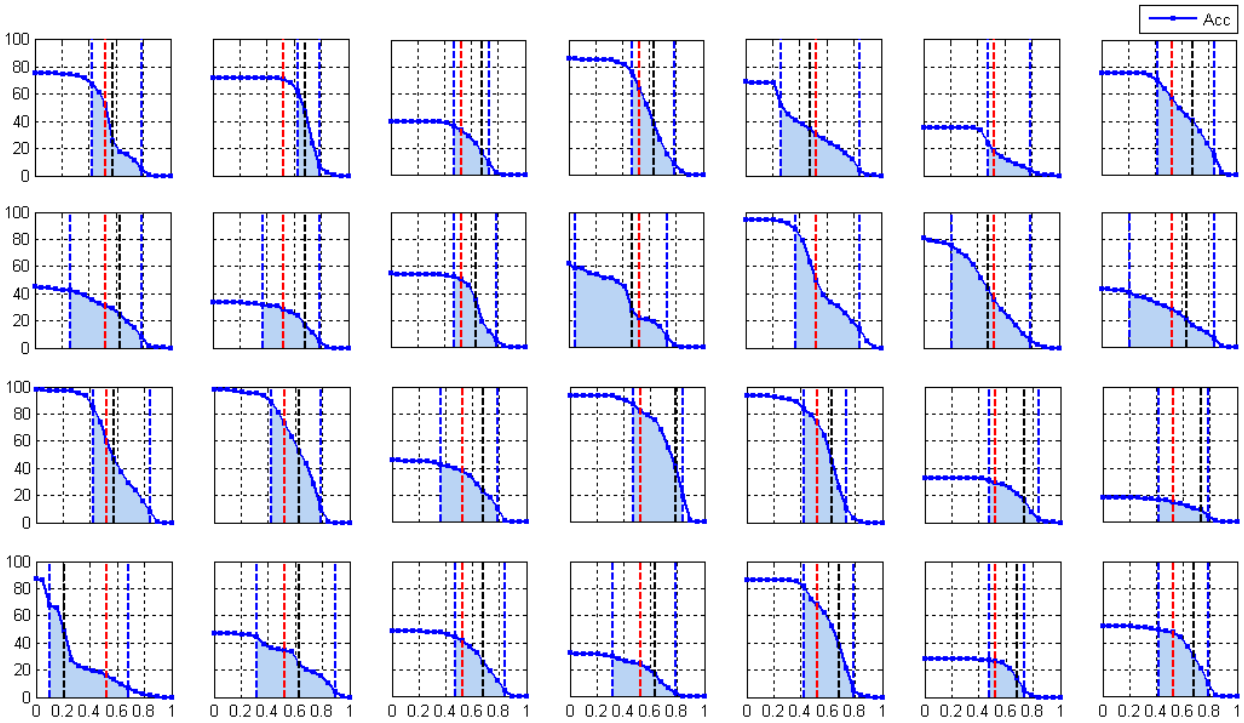


Figura 3.4: Curva m_{E_v} vs $\tau_{\text{casamento}}$ para todos os vídeos do banco de dados PETS2004.

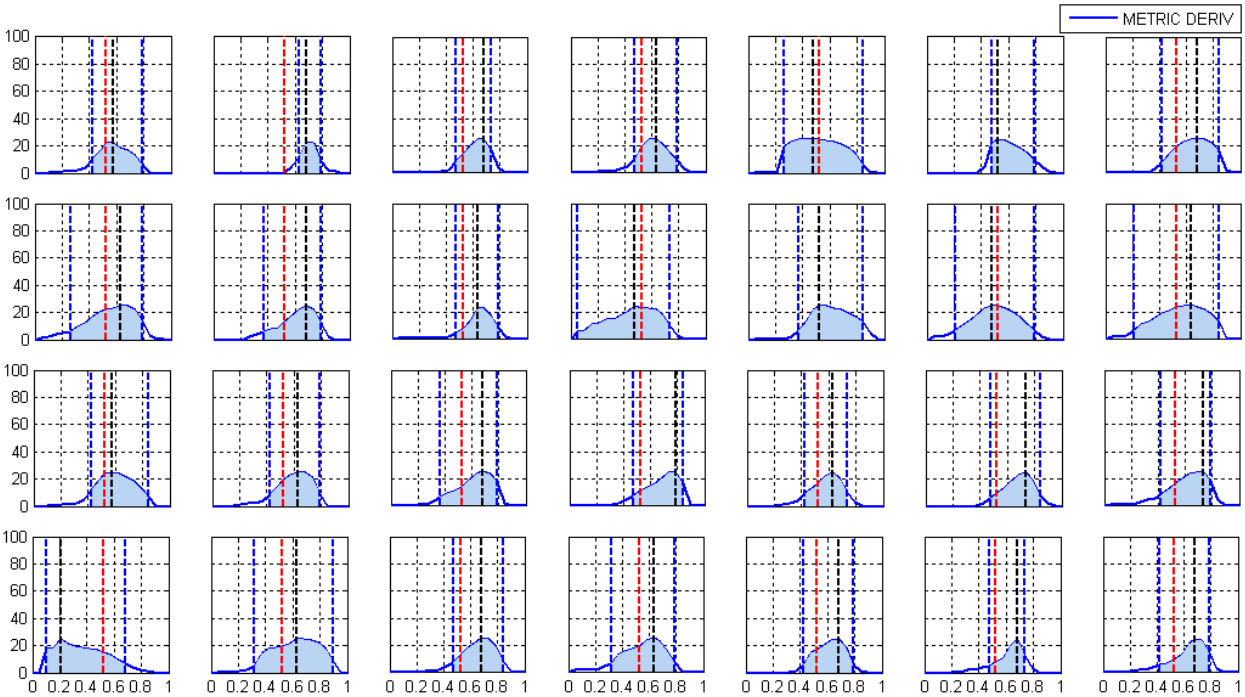


Figura 3.5: Curva $\frac{dm_{E_v}}{d\tau_{\text{casamento}}}$ vs $\tau_{\text{casamento}}$ para todos os vídeos do banco de dados PETS2004.

3.6 Resumo

Neste capítulo foram expostas as perspectivas na avaliação das técnicas de subtração de fundo. Considerando as diferentes vantagens que apresenta um procedimento de avaliação baseado em métricas orientadas a objetos, é proposto um procedimento que permite calcular a métrica de exatidão, seja para cada vídeo ou para todo o banco de dados. Esta métrica representa quão exata é uma técnica de subtração de fundo em encontrar os objetos em movimento ao longo dos quadros de um vídeo, considerando que tem-se o *ground-truth* do banco de dados a analisar. Nos próximos capítulos são avaliadas as diferentes técnicas de subtração de fundo estudadas neste trabalho, estabelecendo uma comparação do desempenho de cada uma delas.

Capítulo 4

Testes e Resultados

4.1 Introdução

Neste capítulo são apresentados os resultados experimentais obtidos para a avaliação das técnicas de subtração de fundo implementadas. Para testar tais técnicas, dois bancos de dados especializados são utilizados. O capítulo divide-se em seis seções. Nas cinco primeiras, realiza-se uma exposição do ambiente de desenvolvimento utilizado na implementação computacional das técnicas descritas nos capítulos anteriores, uma descrição dos bancos de dados utilizados na avaliação, e a apresentação e discussão dos resultados obtidos. Finalmente, na última seção é apresentado um resumo do capítulo.

4.2 O Ambiente Desenvolvido

Todas as técnicas foram desenvolvidas sobre o ambiente *MATrix LABoratory* (MATLAB), aproveitando-se as facilidades de implementação por ele proporcionadas. Este ambiente permite diminuir o tempo de *prototipagem* de sistemas complexos. Hoje em dia, com a adição de diversas ferramentas, organizadas em pacotes denominados *toolboxes*, é possível realizar, inclusive, aplicações profissionais.

No que tange à maneira de programar em MATLAB deve-se estar atento para evitar funções de execução lenta. Como tal linguagem é de propósito geral, é comum basear-se no estilo de programação de outras linguagens como JAVA ou C, o que acarreta um aumento no tempo

de processamento. Melhores resultados são obtidos programando-se de forma matricial, evitando-se o uso de laços, sobretudo, para tais operações. Outra de suas vantagens refere-se à visualização de dados em geral, possível através de um conjunto de recursos gráficos de fácil utilização. O MATLAB também possui funções de leitura e escrita especiais para arquivos de imagens, permitindo a utilização de um banco de imagens. Por todas estas facilidades, optou-se pelo desenvolvimento do sistema de classificação em tal ambiente.

4.3 Bancos de Dados

Os experimentos foram feitos utilizando os bancos de dados: (*a*) PETS2004 apresentado em [24][1]; (*b*) SABS apresentado em [9]. As características de ambos bancos de dados são detalhadas nas próximas subseções.

4.3.1 Banco de Dados PETS2004

O banco de dados PETS2004 é próprio para o desenvolvimento e teste de um sistema de vigilância em espaço público. Considerando que seu *ground-truth* contém todos os objetos rotulados através de retângulos, ele possibilita efetuar um procedimento de avaliação baseado em métricas orientadas a objetos.

O banco de dados PETS2004 consiste de 28 vídeos, compostos de quadros em escala de cores de 384×288 píxeis, capturados a uma taxa de 25 quadros por segundo, utilizando uma câmera com uma lente grande-angular na entrada de um edifício, gerando um total de 26500 quadros (rotulados), agrupados em 6 diferentes cenários de atividade. Na Tabela A.1 mostram-se a classificação e os nomes dos 28 vídeos que compõem o banco de dados. A enumeração a seguir apresenta os cenários de atividade indicados:

1. procura, que contém vídeos com uma pessoa que procura alguma coisa enquanto caminha da parte de trás para frente (*Browse1.mpg*), fica lendo (*Browse2.mpg*), fica sem movimentar-se por um longo período de tempo (*Browse_WhileWaiting1.mpg*), entre outros;

2. deixando objetos, que contém vídeos com uma pessoa deixando uma bolsa ou caixa na entrada, podendo deixar a bolsa na cadeira (*LeftBag_AtChair.mpg*) ou atrás dela

(*LeftBag_BehindChair.mpg*), também podendo deixar a bolsa e depois voltar para pegá-la (*LeftBag_PickedUp.mpg*);

3. encontros, que contém um grupo de duas ou mais pessoas, encontrando-se e caminhando juntas (*Meet_WalkTogether1.mpg*), caminhando juntas e logo se separando (*Meet_WalkSplit.mpg*), assim como quatro que pessoas se encontram, caminham juntas e se separam (*Meet_Crowd.mpg*), entre outros;

4. descansando, desmaiado ou lutando, que contém vídeos de uma pessoa descansando na cadeira (*Rest_InChair.mpg*), caída no chão (*Rest_SlumpOnFloor.mpg*), duas pessoas lutando (*Fight_RunAway1.mpg*) e uma perseguindo a outra (*Fight_Chase.mpg*);

5. caminhando, que contém vídeos de uma pessoa caminhando em linha reta (*Walk1.mpg*) e retornando (*Walk2.mpg*), entre outros.

A rotulação que define o *ground-truth* do banco de dados foi elaborada em [24], levando-se em conta os indivíduos presentes em cada vídeo. Assim, cada indivíduo é representado por um retângulo mais uma descrição de seus movimentos (inativo, ativo, caminhando ou correndo), e é rotulado somente quando inicia um tipo de movimento. Caso contrário, é considerado parte do fundo. Assim, cada quadro pode conter zero ou vários retângulos (ver Figura 4.1.a). Também, cada retângulo é vinculado a um identificador, o qual existirá enquanto o indivíduo for visível. Se ele desaparece e logo volta a aparecer, então o indivíduo obtém um novo identificador. Se o indivíduo é obstruído por alguns poucos quadros, então ele preserva o mesmo identificador. Além da rotulação dos indivíduos, o *ground-truth* também contém um rotulamento grupal, que representa a interação entre eles. Assim, esta interação é indicada por um retângulo de grupo mais uma descrição do movimento grupal (inativo, ativo ou em movimento) (ver Figura 4.1.b), e, de forma similar aos retângulos, cada retângulo de grupo é vinculado a um identificador.

(a) (b)

Figura 4.1: (a) quadro contendo três retângulos, (b) quadro contendo um retângulo de grupo e dois retângulos.

Finalmente, o *ground-truth* vinculado a cada vídeo é armazenado em um arquivo XML, segundo a estrutura proposta em [23]. Neste arquivo estão armazenados os retângulos (vinculados a cada objeto em movimento) e os retângulos de grupo (vinculados a múltiplas objetos em movimento), com seus respectivos descritores de movimento e identificadores presentes em cada quadro. Em geral, é possível dizer que esta estrutura de armazenamento é orientada a quadros, já que a informação é armazenada considerando a ordem sequencial dos quadros.

Na Figura 4.2 é apresentada a estrutura hierárquica dos campos que formam o arquivo XML, onde os retângulos são armazenados no campo *object* e os retângulos de grupo são armazenados no campo *group*, existindo tais campos para cada quadro do vídeo. Observe-se que também é armazenado o número de quadros e o número de objetos e grupos presentes no vídeo.

Uma limitação deste tipo de estruturação da informação dos indivíduos presentes no vídeo é que somente permite armazenar regiões retangulares, já que ela foi especializada para o caso de uma segmentação manual dos quadros. Assim, para o caso de análise do movimento humano não é possível utilizar tal estrutura, já que é preferível armazenar o contorno de um objeto a armazenar somente o retângulo que o contém.

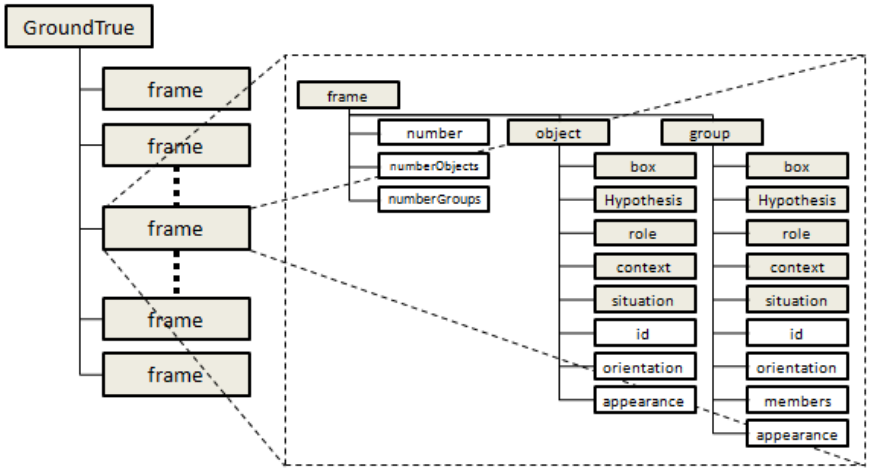


Figura 4.2: Estrutura hierárquica do arquivo XML que contém o *ground-truth* vinculado a cada vídeo do banco de dados PETS2004.

Para o caso das técnicas de subtração de fundo estudadas neste trabalho, e considerando as restrições da estrutura de armazenamento proposta para o PETS2004, decidiu-se utilizar a estrutura indicada pelo projeto *Labelme* [58]. No projeto *Labelme* é proposta uma estrutura de armazenamento orientada aos objetos presentes no vídeo, ou seja, são armazenados independentemente do quadro a que correspondem: (a) o contorno de um objeto, (b) o identificador, (c) um descritor e (d) o número do quadro ao qual pertence cada objeto. Desta

forma, pode-se analisar as características de cada objeto de forma independente.

Na Figura 4.3 é apresentada a estrutura hierárquica dos campos que formam o arquivo XML proposto no projeto *Labelme*, onde os objetos são armazenados no campo *Objects*. Assim, cada campo *Objects* contém um vetor de campos chamados *polygon* que armazenam os contornos do correspondente objeto em todo quadro em que ele aparece.

Vale ressaltar que uma parte importante na tarefa de programação foi a elaboração das funções necessárias para poder ler, armazenar nova informação e vincular os dois tipos de estruturas de armazenamento já explicados. Sem essas operações, a realização de testes seria muito custosa, em termos de tempo, devido ao volume de informação que é utilizado ao se trabalhar com um banco de dados de vídeos. Aqui, é necessário destacar que para poder armazenar e definir um rótulo específico para cada objeto detectado por uma técnica de subtração de fundo foi implementada uma etapa de rastreamento de objetos baseada em filtros de Kalman, explicada em detalhe no Apêndice C.

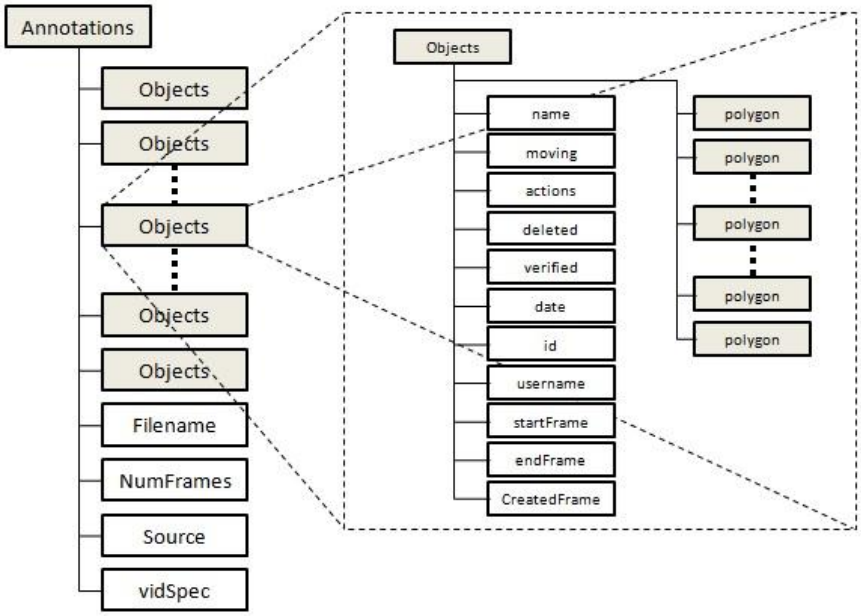


Figura 4.3: Estrutura hierárquica do arquivo XML proposto no projeto *Labelme*, que contém os contornos das regiões segmentadas de um vídeo.

4.3.2 Banco de Dados SABS

O banco de dados SABS é composto por imagens geradas artificialmente, o que possibilita efetuar um procedimento de avaliação baseado em métricas orientadas a píxeis e estudar o desempenho das técnicas de subtração de fundo considerando os desafios que se deve enfren-

tar. Para lidar com o problema que as imagens sintéticas provavelmente não representam fielmente a gama de dados reais [21], na elaboração do banco de dados foi usado um cenário típico de vigilância por vídeo, modelos $3D$ de alta qualidade e uma moderna tecnologia de *raytracing* para gerar uma síntese realista de imagens[1]. Os quadros têm uma resolução de 800×600 píxeis e são capturados a partir de um ponto de vista fixo. O ruído do sensor foi simulado usando ruído Gaussiano aditivo (de médio igual a 0 e desvio padrão igual a $0,0001$) para cada quadro. Em comparação a um *ground-truth* rotulado manualmente, o banco de dados SABS apresenta um *ground-truth* que não sofre de rotulamento imperfeito nem é composto por uma quantidade pequena de quadros rotulados. As características principais do banco de dados são:

- é composto por nove sequências de vídeo que representam diferentes cenários de vigilância, abrangendo cada um dos desafios típicos a serem tratados pela técnica de subtração de fundo decorrentes da natureza de um vídeo de vigilância (ver a Tabela A.2). A enumeração a seguir apresenta os cenários indicados:

  1. básico, é um cenário de vigilância básico que combina uma série de desafios, permitindo ter uma visão geral do desempenho;
  2. fundo dinâmico, em que são considerados deslocamentos de alguns objetos do fundo como as folhas da árvore em movimento e a mudança da iluminação dos semáforos;
  3. *bootstrapping*, este cenário não contém fase de treinamento, de forma que a determinação do modelo do fundo começa após o primeiro quadro;
  4. escurecimento, em que é simulada uma mudança gradual na cena ao diminuir a iluminação de forma constante. Assim, o fundo e o primeiro plano são escurecidos, e o contraste entre eles diminui;
  5. comutação da iluminação, em que mudanças pontuais são simuladas por comutar a luz da loja (quadro 901) e ligá-la novamente (quadro 1101);
  6. noite ruidosa, que é um cenário de vigilância básico de noite, com um incremento do ruído do sensor e um baixo contraste fundo/primeiro plano, resultando em uma maior camuflagem;
  7. sombras, em que são usados detalhes da região da rua para determinar o número de píxeis correspondentes às sombras que são classificados como primeiro plano;
  8. camuflagem, em que também são usados detalhes da região da rua, sendo comparado o desempenho entre uma sequência com pessoas vestindo roupas escuras

[1] A sequência foi renderizada utilizando o programa *Mental Ray*, um *raytracer* fornecido por *Autodesk Maya*, enquanto o *ground-truth* foi gerado pelo programa *Maya Vector*.

e carros cinzentos, e uma sequência contendo objetos do primeiro plano com uma cor significativamente diferente do fundo;

9. compressão de vídeo, em que são utilizadas sequências de vídeo comprimidas a diferentes taxas de bits, por um *codec* padrão de uso frequente em vídeo de vigilância[2].

- Cada vídeo do banco de dados é dividido em duas partes, uma parte correspondente à fase de treinamento (em geral, sem objetos do primeiro plano) e a outra parte correspondente à fase de teste (com objetos do primeiro plano);
- O *ground-truth* relacionado a cada quadro foi elaborado como uma imagem de múltiplos rótulos (um rótulo para cada objeto de interesse no cenário de vigilância). Desta maneira, vários objetos do primeiro plano podem ser destacados, podendo assim utilizar o *ground-truth* para a avaliação de técnicas de rastreamento de objetos.

4.4 Resultados Considerando o Banco de Dados PETS2004

Considerando que o banco de dados PETS2004 permite determinar o desempenho de uma técnica de subtração de fundo utilizando um procedimento de avaliação baseado em métricas orientadas a objetos, seu uso se deu para: (a) quantificar o desempenho através da métrica de exatidão, calculada para cada vídeo e para todo o banco de dados, através das Equações (3.27) e (3.28) definidas na seção 3.4; (b) determinar as métricas de exatidão relacionadas com cada um dos tipos de casamentos, calculadas para cada vídeo e para todo o banco de dados, através das equações

$$m_{E-S_v} = \frac{N_{S_v}}{N_{T_v}} \qquad m_{E-S_{db}} = \frac{N_{S_{db}}}{N_{T_{db}}}, \tag{4.1}$$

$$m_{E-U_v} = \frac{N_{U_v}}{N_{T_v}} \qquad m_{E-U_{db}} = \frac{N_{U_{db}}}{N_{T_{db}}}, \tag{4.2}$$

$$m_{E-SU_v} = \frac{N_{SU_v}}{N_{T_v}} \qquad m_{E-SU_{db}} = \frac{N_{SU_{db}}}{N_{T_{db}}}, \tag{4.3}$$

$$m_{E-FD_v} = \frac{N_{FN_v}}{N_{T_v}} \qquad m_{E-FD_{db}} = \frac{N_{FN_{db}}}{N_{T_{db}}}, \tag{4.4}$$

$$m_{E-FA_v} = \frac{N_{FP_v}}{N_{T_v}} \qquad m_{E-FA_{db}} = \frac{N_{FP_{db}}}{N_{T_{db}}}, \tag{4.5}$$

onde $N_{T_v} = N_{CD_v} + N_{FD_v} + N_{FA_v} + N_{S_v} + N_{U_v} + N_{SU_v}$ e $N_{T_{db}} = N_{CD_{db}} + N_{FD_{db}} + N_{FA_{db}} + N_{S_{db}} + N_{U_{db}} + N_{SU_{db}}$. Assim, as Equações (4.1) - (4.5) permitem determinar que tipo de erro tem uma maior presença na resposta de uma técnica de subtração de fundo.

[2]H.264, 40-640 kbits/s, 30 quadros por segundo

Estas métricas permitem quantificar: (a) a proporção dos objetos corretamente detectados (m_{E_v} ou $m_{E_{db}}$); (b) a proporção de objetos detectados, que podem ser oriundos de uma separação (m_{E-S_v} ou $m_{E-S_{db}}$), união (m_{E-U_v} ou $m_{E-U_{db}}$), ou a mistura de ambos (m_{E-SU_v} ou $m_{E-SU_{db}}$); (c) a proporção dos falsos alarmes detectados (m_{E-FA_v} ou $m_{E-FA_{db}}$) e as falsas detecções (m_{E-FD_v} ou $m_{E-FD_{db}}$).

Nas seguintes subseções são apresentados os resultados das provas feitas utilizando o banco de dados PETS2004, seguindo o procedimento descrito abaixo.

Procedimento 1 Passos para a realização dos testes efetuados utilizando o banco de dados PETS2004.

1. Cada vídeo do banco de dados foi dividido nos conjuntos de treinamento e de teste indicados na Tabela A.1 (ver página 116). Esta separação é necessária, já que as técnicas de subtração de fundo requerem uma etapa de inicialização;

2. foi realizado o cálculo das métricas de exatidão definidas pelas Equações (3.27), (3.28) e (4.1) - (4.5) segundo o procedimento descrito no Algoritmo 1, considerando o conjunto de teste estipulado para cada vídeo. Assim, os resultados são apresentados em função da curva m_E versus $\tau_{\text{casamento}}$.

Para descartar uma área da cena que contém indivíduos que não estão registrados no *ground-truth* (área onde se encontra uma recepcionista), tal como foi proposto em [29], é declarada uma região de não detecção através do mascaramento de cada quadro, tanto na etapa de treinamento como de teste. Na Figura 4.4.b é apresentado um quadro mascarado, onde a região de não detecção é definida pela área retangular preta.

(a) (b)

Figura 4.4: (a) quadro sem mascaramento; (b) quadro com mascaramento da região de não detecção.

Finalmente, na Tabela 4.1 são apresentados os valores numéricos dos parâmetros relacionados a cada técnica de subtração de fundo testada no banco de dados PETS2004. O critério para a sintonização dos parâmetros é explicado a seguir, inicialmente são utilizados como valores de referência os valores publicados na literatura, de cada uma das técnicas estudadas e, a partir destes valores é feita uma procura dos valores ótimos.

Tabela 4.1: Parâmetros das técnicas de: modelamento do fundo, detecção de variações e pós-processamento, quando é usado o banco de dados PETS2004.

Técnicas de Subtração de Fundo		Parâmetros
Média móvel Gaussiana	Distância Euclidiana Simplificada	$\alpha_{svt} = 0,3$; $\tau_\theta = 10$, $\alpha = 2 \times 10^{-3}$.
	Distância Euclidiana Bivariada	$\alpha_{svt} = 0,3$; $\alpha = 10^{-6}$.
	Teste de Significância	$\alpha_{svt} = 0,3$; $\sigma_{\text{invar}} = 27$, $N_{\mathcal{D}} = 25$, $B = 3$, $\alpha = 5 \times 10^{-4}$.
Histograma	Distância Euclidiana Simplificada	$\sigma_{ds} = 3$; $\tau_\theta = 10$, $\alpha = 2 \times 10^{-3}$.
	Distância Euclidiana Bivariada	$\sigma_{ds} = 3$; $\alpha = 10^{-6}$.
	Teste de Significância	$\sigma_{ds} = 3$; $\sigma_{\text{invar}} = 27$, $N_{\mathcal{D}} = 25$, $B = 3$, $\alpha = 5 \times 10^{-4}$.
Mistura de Gaussianas		$K_{mg} = 5$, $D_{vz} = 0,6$, $\alpha_{apr} = 25 \times 10^{-3}$, $\omega_0 = 10^{-3}$, $\sigma_0 = 45$, $\rho = 10^{-3}$,$T_{\text{fundo}} = 0,8$.
Pós-processamento		$N_{\text{MinArea}} = 60$

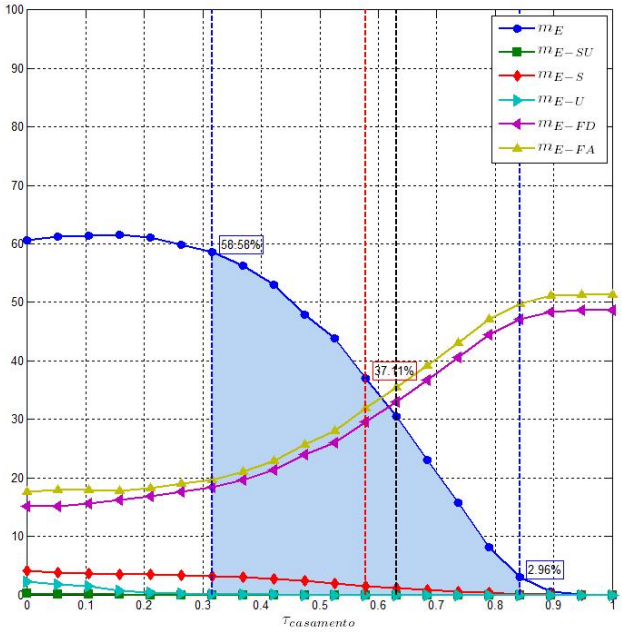
4.4.1 Técnica de Subtração de Fundo Baseada na Média Móvel Gaussiana

Seguindo o procedimento 1, são obtidas as Figuras 4.5, 4.6 e 4.7, onde observam-se as métricas de exatidão em relação ao limiar $\tau_{\text{casamento}}$. Considerando-se as três técnicas implementadas para o passo de detecção de variações, observa-se que a técnica baseada na distância Euclidiana simplificada apresenta a maior exatidão (ver Figura 4.5), onde, para o melhor caso ($\tau_L = 0,32$), esta técnica obtém uma exatidão de $m_{E_{db}} = 58,6\%$, e para o caso intermediário ($\tau_R = 0,58$) alcança uma exatidão igual a $m_{E_{db}} = 37,1\%$. Em todos os casos, a técnica baseada no teste de significância apresenta um desempenho menor, devido, em grande parte, ao erro oriundo dos falsos alarmes, que em geral toma valores superiores aos $17,3\%$ ($< m_{E-FA_{db}}$) (ver Figura 4.7), implicando que esta técnica apresenta uma determinada sensibilidade na detecção de regiões de troca. Analisando a Figura 4.6, vinculada às métricas de exatidão considerando a técnica de detecção de variações baseada na distância Euclidiana bivariada, observa-se que a presença do tipo de erro fruto das falhas na detecção é predominante (para o melhor caso tem-se $m_{E-FD_{db}} = 29,0\%$), implicando que esta técnica é conservadora no momento de detectar as áreas vinculadas aos objetos em movimento numa cena. Tal comportamento faz que esta abordagem, em comparação com as outras técnicas, seja menos susceptível a falsos alarmes (para o melhor caso tem-se $m_{E-FA_{db}} = 17,7\%$).

Nas Tabelas B.1, B.2 e B.3 (ver páginas 119, 120 e 121, respectivamente) são apresentados os números de ocorrências segundo o tipo de casamento, e as respectivas métricas de exatidão para o melhor caso, considerando as três técnicas implementadas para o passo de detecção de variações. As tabelas corroboram o já exposto, e indicam que: *a*) para o caso da técnica baseada na distância Euclidiana simplificada (ver Tabela B.1 na página 119), tem-se dez vídeos com uma métrica de exatidão maior que 80%, e seis vídeos com uma exatidão inferior a 40%. Nestes seis vídeos é possível ver as taxas de falso alarme, que variam de $22,6\%$ até $59,3\%$; *b*) para o caso da técnica baseada na distância Euclidiana bivariada (ver Tabela B.2 na página 120), tem-se doze vídeos com uma métrica de exatidão maior que 60%, e seis vídeos com uma exatidão inferior a 40%. Nestes seis vídeos é possível ver altas taxas de falso alarme (variando de $1,1\%$ até $60,4\%$) e falhas na detecção (variando de 15% até $71,1\%$); *c*) para o caso da técnica baseada no teste de significância (ver Tabela B.3 na página 121), tem-se dez vídeos com uma métrica de exatidão maior que 60%, e nove vídeos com uma exatidão inferior a 40%. Nestes nove vídeos é possível ver as taxas de falso alarme, que variam de $0,8\%$ até $63,2\%$.

Finalmente, na Figura 4.8 observa-se alguns exemplos onde os falsos alarmes, na maioria dos casos, correspondem aos indivíduos detectados na área iluminada (ver Figuras 4.8.d-f e

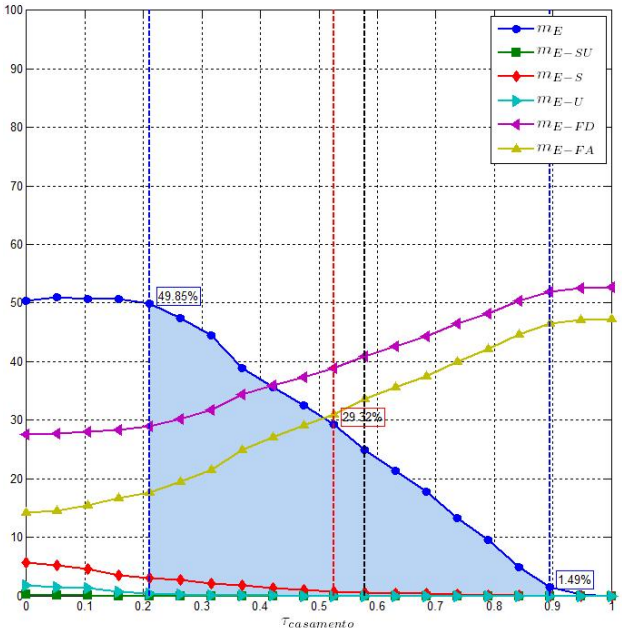
4.8.i) e as falhas na detecção correspondem a indivíduos que não foram detectados completamente devido ao problema de camuflagem (ver Figuras 4.8.b e 4.8.k) ou são inicialmente tão pequenos que são filtrados pela etapa de pós-processamento (ver Figuras 4.8.e e 4.8.h). Também, tem-se casos de falsos alarmes devido a: *a*) indivíduos que, por ter um mínimo movimento, não são rotulados no *ground-truth*, sendo considerados parte do fundo, representando o problema de fundo dinâmico (ver Figuras 4.8.g e 4.8.i); *b*) indivíduos que, na etapa de treinamento eram parte do fundo e na etapa de teste foram parte do primeiro plano (ver Figuras 4.8.a, 4.8.c, 4.8.j, 4.8.k e 4.8.l). Este comportamento evidencia que a técnica de subtração de fundo baseada na média móvel gaussiana tem uma forte tendência a preservar o modelo do fundo determinado na etapa de treino, implicando que à medida que a atualização vai acontecendo, o modelo do fundo não sofre variações de importância em relação aos valores já assumidos no treinamento.



exatidão	parâmetro de sobreposição[a]		
	τ_L (0,32)	τ_R (0,58)	τ_C (0,84)
m_E	**58,6%**	**37,1%**	**3,0%**
m_{E-SU}	0,0%	0,0%	0,0%
m_{E-S}	3,2%	1,5%	0,1%
m_{E-U}	0,1%	0,0%	0,0%
m_{E-FD}	18,4%	29,5%	47,1%
m_{E-FA}	19,7%	31,9%	49,8%

[a] τ_L é o limiar **l**iberal, τ_R é o limiar **r**azoável e τ_C é o limiar **c**onservador, definidos na seção 3.5.

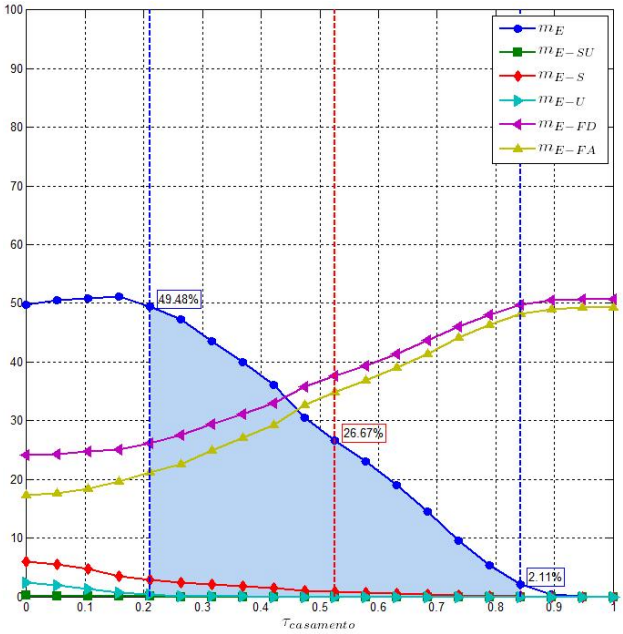
Figura 4.5: Gráfico das métricas de exatidão em relação ao parâmetro de sobreposição $\tau_{\text{casamento}}$, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana simplificada.



exatidão	parâmetro de sobreposição[a]		
	τ_L (0,21)	τ_R (0,53)	τ_C (0,89)
m_E	**49,9%**	**29,3%**	**1,5%**
m_{E-SU}	0,0%	0,0%	0,0%
m_{E-S}	3,1%	0,7%	0,0%
m_{E-U}	0,4%	0,0%	0,0%
m_{E-FD}	29,0%	38,9%	52,0%
m_{E-FA}	17,7%	31,1%	46,5%

[a] τ_L é o limiar **l**iberal, τ_R é o limiar **r**azoável e τ_C é o limiar **c**onservador, definidos na seção 3.5.

Figura 4.6: Gráfico das métricas de exatidão em relação ao parâmetro de sobreposição $\tau_{\text{casamento}}$, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana bivariada.



exatidão	parâmetro de sobreposição[a]		
	τ_L (0,21)	τ_R (0,53)	τ_C (0,84)
m_E	**49,5%**	**26,7%**	**2,1%**
m_{E-SU}	0,0%	0,0%	0,0%
m_{E-S}	2,8%	0,8%	0,0%
m_{E-U}	0,4%	0,0%	0,0%
m_{E-FD}	26,1%	37,7%	49,7%
m_{E-FA}	21,1%	34,9%	48,2%

[a] τ_L é o limiar **l**iberal, τ_R é o limiar **r**azoável e τ_C é o limiar **c**onservador, definidos na seção 3.5.

Figura 4.7: Gráfico das métricas de exatidão em relação ao parâmetro de sobreposição $\tau_{\text{casamento}}$, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada no teste de significância.

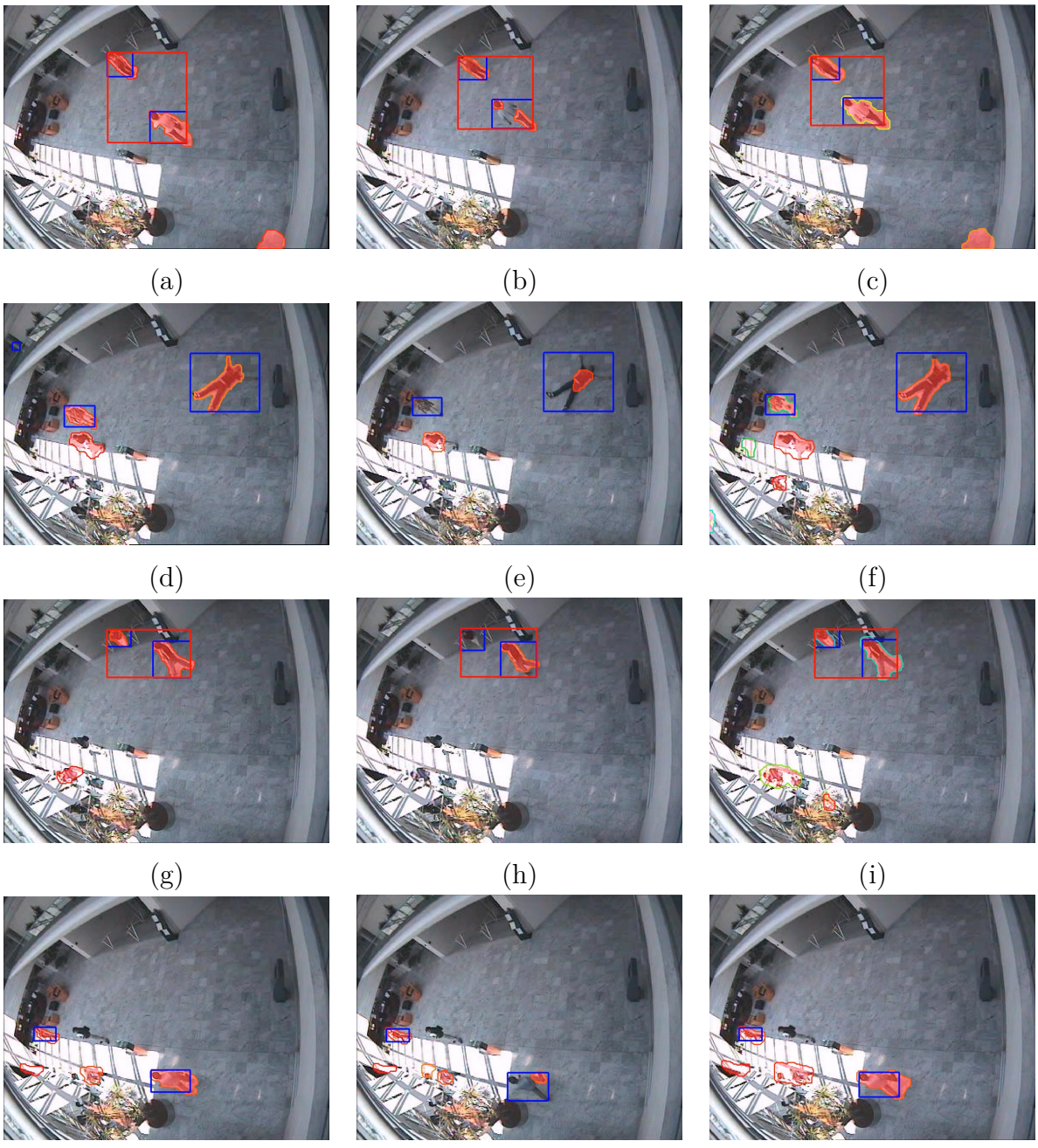


Figura 4.8: Erros nos resultados da detecção, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada (a,d,g,j) na distância Euclidiana simplificada; (b,e,h,k) na distância Euclidiana bivariada; (c,f,i,l) no teste de significância (os retângulos fazem referência ao *ground-truth*, e os contornos indicam os objetos detectados pela técnica de subtração de fundo).

4.4.2 Técnica de Subtração de Fundo Baseada no Histograma

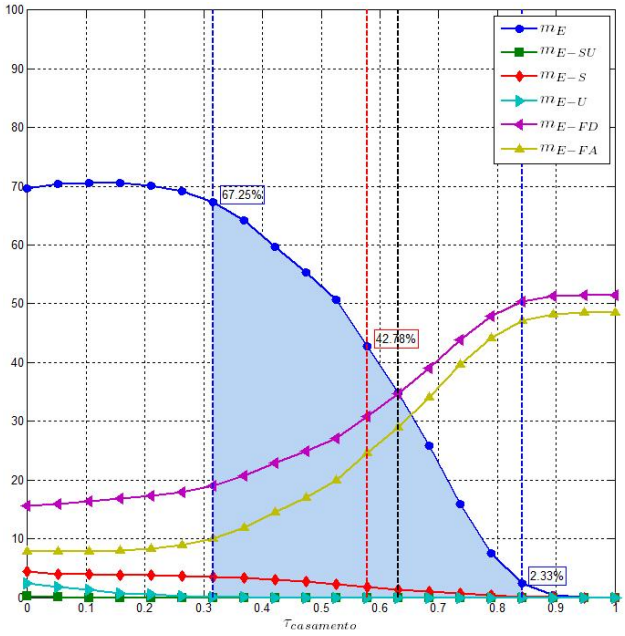
Para o caso da técnica de subtração de fundo baseada no histograma, as Figuras 4.9, 4.10 e 4.11 apresentam as métricas de exatidão em relação ao limiar $\tau_{\text{casamento}}$, considerando as três técnicas implementadas para o passo de detecção de variações. Aqui, também, a maior exatidão na resposta é gerada pela técnica baseada na distância Euclidiana simplificada (ver Figura 4.9), onde para o melhor caso ($\tau_L = 0,32$) tem-se uma métrica de exatidão de $m_{E_{db}} = 67,2\%$ e para o caso intermediário ($\tau_R = 0,58$) tem-se uma exatidão igual a $m_{E_{db}} = 42,8\%$. Em todos os casos, a técnica baseada no teste de significância apresenta um desempenho menor, devido também ao erro fruto dos falsos alarmes, que toma valores superiores a $11,4\%$ ($< m_{E-FA_{db}}$) (ver Figura 4.11). Analisando a Figura 4.10, vinculada às métricas de exatidão considerando a técnica de detecção de variações baseada na distância Euclidiana bivariada, observa-se uma maior exatidão em relação a seu equivalente no caso anterior (ver Figura 4.6), devido, principalmente, a uma diminuição dos falsos alarmes e das falhas na detecção (para o melhor caso, tem-se uma diminuição de $3,7\%$ nas falhas de detecção e de $7,2\%$ nos falsos alarmes).

Nas Tabelas B.4, B.5 e B.6 (ver páginas 122, 123 e 124, respectivamente) são apresentados os números de ocorrências segundo o tipo de casamento, e as respectivas métricas de exatidão para o melhor caso, considerando as três técnicas implementadas para o passo de detecção de variações. Aqui observa-se que: *a*) para o caso da técnica baseada na distância Euclidiana simplificada (ver Tabela B.4 na página 122), tem-se treze vídeos com uma métrica de exatidão maior que 80%, e seis vídeos com uma exatidão inferior a 43%. Nestes seis vídeos é possível ver as taxas de falso alarme, que variam de $3,9\%$ até $37,6\%$; *b*) para o caso da técnica baseada na distância Euclidiana bivariada (ver Tabela B.5 na página 123), tem-se dez vídeos com uma métrica de exatidão maior que 80%, e sete vídeos com uma exatidão inferior a 40%. Nestes sete vídeos é possível ver altas taxas de falso alarme (variando de $1,1\%$ até $43,7\%$) e falhas na detecção (variando de $14,4\%$ até $68,4\%$); *c*) para o caso da técnica baseada no teste de significância (ver Tabela B.6 na página 124), tem-se cinco vídeos com uma métrica de exatidão maior que 80%, e nove vídeos com uma exatidão inferior a 40%. Nestes nove vídeos é possível ver as taxas de falso alarme, que variam de $0,5\%$ até $50,6\%$.

Comparando a Figura 4.9 com as Figuras 4.5-4.7, observa-se que a técnica de subtração de fundo baseada no histograma, trabalhando em conjunto com a técnica baseada na distância Euclidiana simplificada, apresenta um melhor desempenho. Este resultado é importante, já que ambas técnicas têm como suposição principal que o modelo para cada processo de um píxel é definido por uma distribuição unimodal, e para o caso da técnica baseada na média

móvel gaussiana é particularizada por uma distribuição normal. Entretanto, as técnicas divergem em relação à estatística que define o modelo do fundo (no primeiro caso é definida como a moda do histograma, e no segundo caso como a média da distribuição normal) estando aqui a possível explicação deste melhor desempenho. Considerando a presença de pontos duvidosos ou ruidosos em cada processo de um píxel, a moda apresenta um comportamento mais estável em relação à média. Portanto, a técnica baseada no histograma será mais robusta ao ruído, explicando-se, assim, seu melhor desempenho.

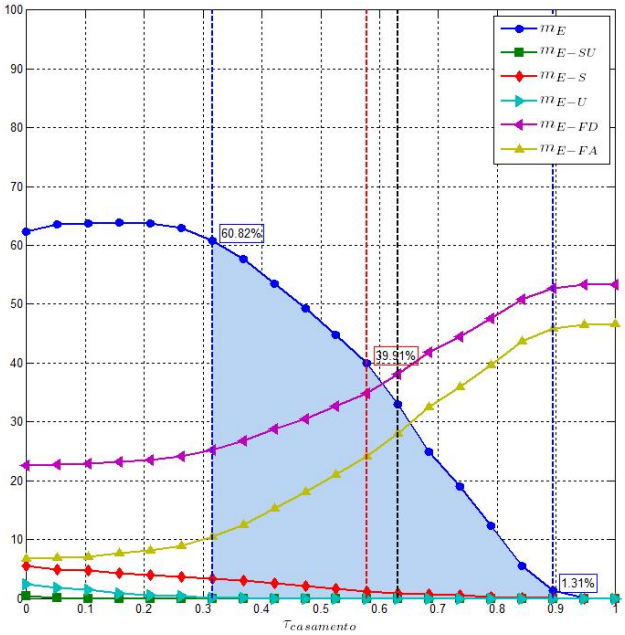
Finalmente, na Figura 4.12 observam-se alguns exemplos, onde os falsos alarmes, na maioria dos casos, correspondem ao problema de fundo dinâmico (ver Figuras 4.8.d-f e 4.8.g-l) e as falhas na detecção correspondem a indivíduos que estão nas áreas obscuras (ver Figuras 4.8.a-c) ou são filtrados pela etapa de pós-processamento (ver Figura 4.8.h).



exatidão	parâmetro de sobreposição[a]		
	τ_L (0,32)	τ_R (0,58)	τ_C (0,84)
m_E	**67,2%**	**42,8%**	**2,3%**
m_{E-SU}	0,0%	0,0%	0,0%
m_{E-S}	3,5%	1,7%	0,1%
m_{E-U}	0,1%	0,0%	0,0%
m_{E-FD}	19,1%	30,9%	50,4%
m_{E-FA}	10,0%	24,6%	47,2%

[a] τ_L é o limiar **l**iberal, τ_R é o limiar **r**azoável e τ_C é o limiar **c**onservador, definidos na seção 3.5.

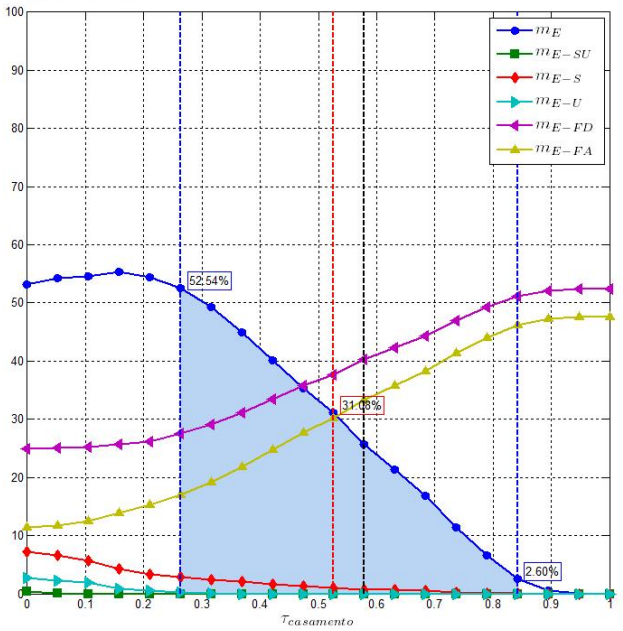
Figura 4.9: Gráfico das métricas de exatidão em relação ao parâmetro de sobreposição $\tau_{\text{casamento}}$, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana simplificada.



exatidão	parâmetro de sobreposição[a]		
	τ_L (0,32)	τ_R (0,58)	τ_C (0,89)
m_E	**60,8%**	**39,9%**	**1,3%**
m_{E-SU}	0,0%	0,0%	0,0%
m_{E-S}	3,3%	1,2%	0,0%
m_{E-U}	0,1%	0,0%	0,0%
m_{E-FD}	25,3%	34,8%	52,8%
m_{E-FA}	10,5%	24,1%	45,9%

[a] τ_L é o limiar **l**iberal, τ_R é o limiar **r**azoável e τ_C é o limiar **c**onservador, definidos na seção 3.5.

Figura 4.10: Gráfico das métricas de exatidão em relação ao parâmetro de sobreposição $\tau_{\text{casamento}}$, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana bivariada.



exatidão	parâmetro de sobreposição[a]		
	τ_L (0,26)	τ_R (0,53)	τ_C (0,84)
m_E	**52,5%**	**31,1%**	**2,6%**
m_{E-SU}	0,0%	0,0%	0,0%
m_{E-S}	2,9%	0,9%	0,0%
m_{E-U}	0,2%	0,0%	0,0%
m_{E-FD}	27,5%	37,7%	51,2%
m_{E-FA}	16,9%	30,3%	46,2%

[a] τ_L é o limiar **l**iberal, τ_R é o limiar **r**azoável e τ_C é o limiar **c**onservador, definidos na seção 3.5.

Figura 4.11: Gráfico das métricas de exatidão em relação ao parâmetro de sobreposição $\tau_{\text{casamento}}$, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseado no teste de significância.

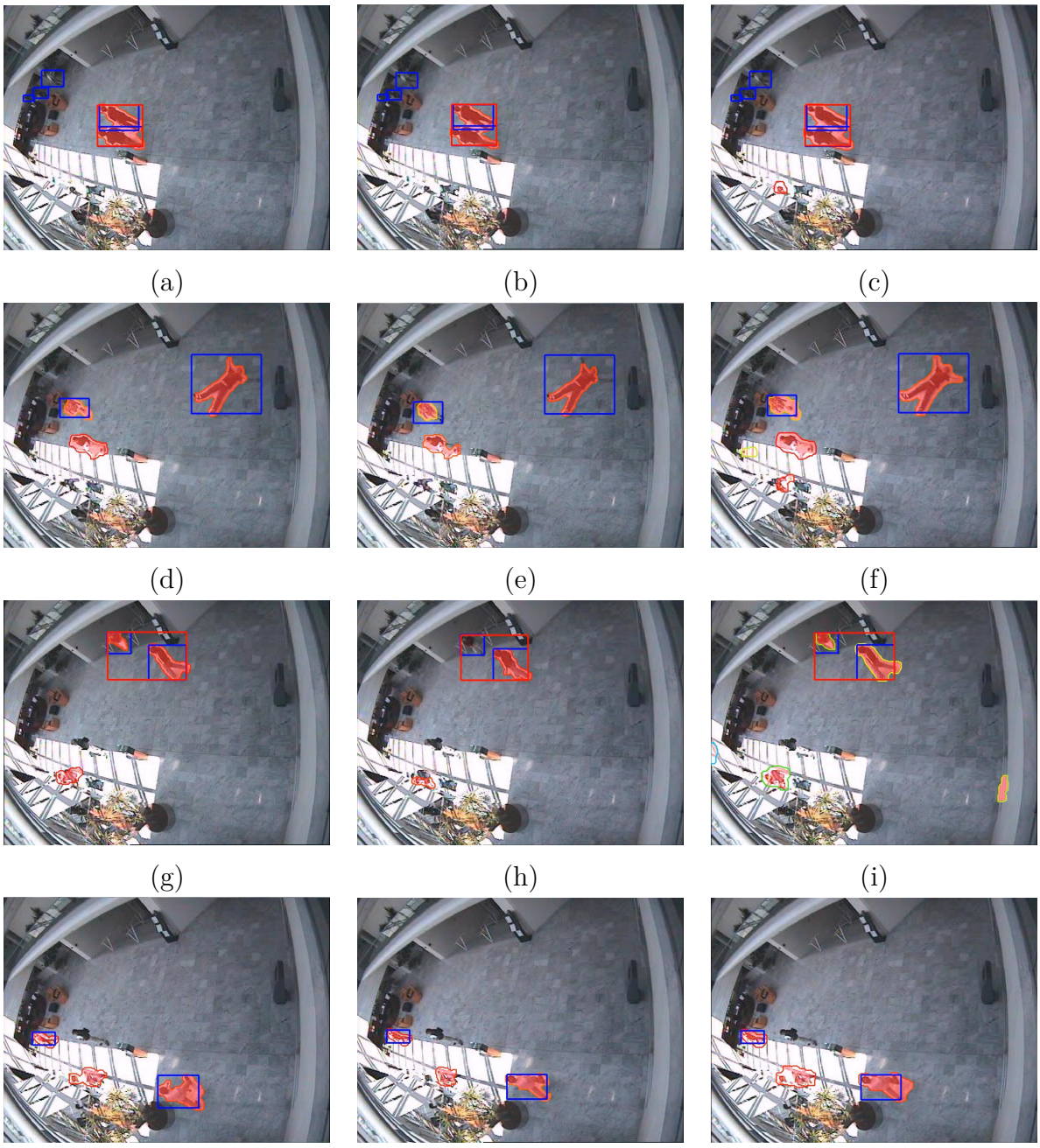


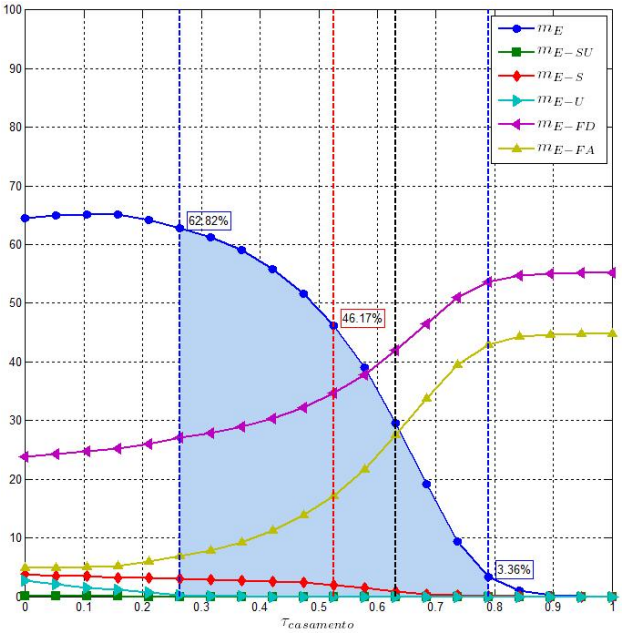
Figura 4.12: Erros nos resultados da detecção, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseada (a,d,g,j) na distância Euclidiana simplificada; (b,e,h,k) na distância Euclidiana bivariada; (c,f,i,l) no teste de significância (os retângulos fazem referência ao *ground-truth*, e os contornos indicam os objetos detectados pela técnica de subtração de fundo).

4.4.3 Técnica de Subtração de Fundo Baseada na Mistura de Gaussianas

A Figura 4.13 apresenta as métricas de exatidão em relação ao limiar $\tau_{\text{casamento}}$ da técnica de subtração de fundo baseada na mistura de Gaussianas, utilizando o procedimento para a determinação da máscara do primeiro plano própria desta abordagem. Observe-se que, para o melhor caso ($\tau_L = 0,26$), tem-se uma métrica de exatidão de $m_{E_{db}} = 62,8\%$ e para o caso intermediário ($\tau_R = 0,53$) tem-se uma taxa igual a $m_{E_{db}} = 46,2\%$. Esta técnica, também tem uma exatidão inferior à obtida pela técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica baseada na distância Euclidiana simplificada (ver Figura 4.9), devendo-se este menor valor na exatidão, fundamentalmente, à alta taxa de falhas na detecção, fruto do problema de primeiro plano adormecido. Na Tabela B.1 (ver página 119) são apresentados os números de ocorrências segundo seu tipo de casamento, e as respectivas métricas de exatidão para o melhor caso. A partir dela, é possível observar que tem-se sete vídeos com uma métrica de exatidão maior que 80%, e quatro vídeos com uma exatidão inferior a 50%, sendo que em tais vídeos também se observam as altas taxas de falha na detecção, que variam de $11,9\%$ até $54,7\%$.

Um ponto importante dessa abordagem é que ela tem um melhor desempenho ao lidar com o problema de um fundo dinâmico, devido principalmente à suposição de que todo processo de um píxel é explicado através de uma distribuição multimodal, permitindo representar os diferentes agrupamentos que definiriam um fundo dinâmico.

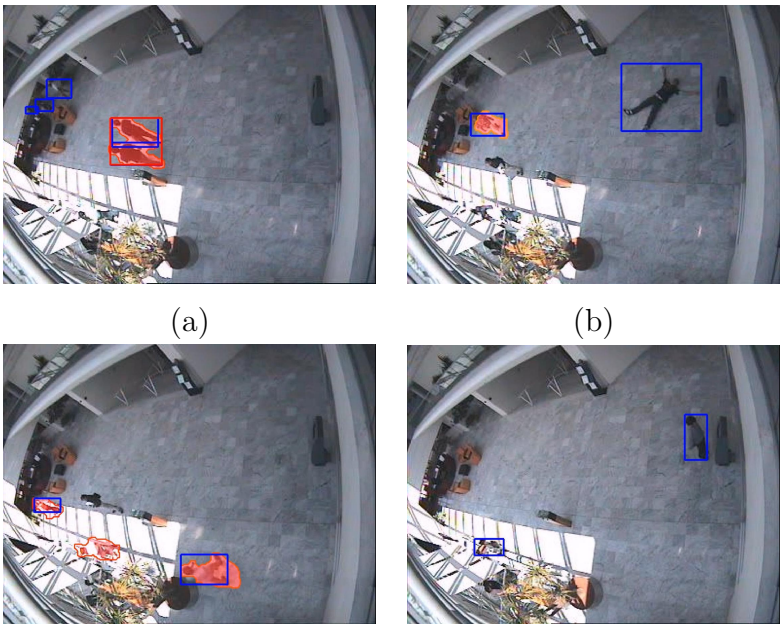
Finalmente, na Figura 4.14 observam-se alguns exemplos, de forma equivalente às abordagens anteriores, com a presença de falhas na detecção devido a indivíduos que estão nas áreas obscuras (ver Figura 4.14.a) ou devido ao problema de primeiro plano adormecido (ver Figuras 4.14.b e 4.14.d). Já para o caso dos falsos, alarmes esta técnica trata de uma forma muito melhor o problema de fundo dinâmico (ver Figura 4.8.c) sendo, entretanto, também susceptível a ele no tempo em que a mistura modela as oscilações do fundo.



exatidão	parâmetro de sobreposição[a]		
	τ_L (0,26)	τ_R (0,53)	τ_C (0,79)
m_E	**62,8%**	**46,2%**	**3,4%**
m_{E-SU}	0,0%	0,0%	0,0%
m_{E-S}	3,0%	2,0%	0,1%
m_{E-U}	0,3%	0,0%	0,0%
m_{E-FD}	27,0%	34,6%	53,6%
m_{E-FA}	6,9%	17,2%	42,9%

[a] τ_L é o limiar **l**iberal, τ_R é o limiar **r**azoável e τ_C é o limiar **c**onservador, definidos na seção 3.5.

Figura 4.13: Gráfico das métricas de exatidão em relação ao parâmetro de sobreposição $\tau_{\text{casamento}}$, quando é testada a técnica de subtração de fundo baseada na mistura de Gaussianas.



(a) (b)

(c) (d)

Figura 4.14: Erros nos resultados da detecção, quando é testada a técnica de subtração de fundo baseada na mistura de Gaussianas (os retângulos fazem referência ao *ground-truth*, e os contornos indicam os objetos detectados pela técnica de subtração de fundo).

4.5 Resultados Considerando o Banco de Dados SABS

Considerando que o banco de dados SABS foi elaborado para determinar o desempenho de uma técnica de subtração de fundo utilizando um procedimento de avaliação baseado em métricas orientadas a píxeis, foi estabelecido em [9] o uso da precisão, sensibilidade e a correspondente medida F como as métricas a empregar para quantificar o desempenho de uma técnica de subtração de fundo quando é usado este banco de dados. Os passos para determinar estas métricas são explicados a seguir:

1. calculam-se para cada quadro de um vídeo os valores de $N_{VP_{fr}}$, $N_{FP_{fr}}$, $N_{FN_{fr}}$ e $N_{VN_{fr}}$, definidos como

$$N_{VP_{fr}} = \sum_{l} \mathbb{I}_{1,1}\left(\mathbf{M}_{dt_l}, \mathbf{M}_{gt_l}\right) \qquad N_{FP_{fr}} = \sum_{l} \mathbb{I}_{1,0}\left(\mathbf{M}_{dt_l}, \mathbf{M}_{gt_l}\right), \tag{4.6}$$

$$N_{FN_{fr}} = \sum_{l} \mathbb{I}_{0,1}\left(\mathbf{M}_{dt_l}, \mathbf{M}_{gt_l}\right) \qquad N_{VN_{fr}} = \sum_{l} \mathbb{I}_{0,0}\left(\mathbf{M}_{dt_l}, \mathbf{M}_{gt_l}\right), \tag{4.7}$$

   onde $\mathbf{M}_{dt_l}$ e $\mathbf{M}_{gt_l}$ são os valores binários na localização (x_l, y_l) relacionadas às máscaras do primeiro plano, obtida pela técnica de subtração de fundo analisada, e armazenada no *ground-truth*, respectivamente, sendo $\mathbb{I}(\bullet)$ a função indicadora, definida como

$$\mathbb{I}_{a,b}\left(\mathbf{M}_{dt_l}, \mathbf{M}_{gt_l}\right) = \begin{cases} 1, & \mathbf{M}_l = a, \mathbf{M}_l = b \\ 0, & \text{caso contrário} \end{cases};$$

2. calculam-se para cada vídeo os valores médios $\overline{N}_{VP_v}$, $\overline{N}_{FP_v}$, $\overline{N}_{FN_v}$ e $\overline{N}_{VN_v}$, definidos como

$$\overline{N}_{VP_v} = \frac{1}{N_{\text{quadros}}} \sum_{fr=1}^{N_{\text{quadros}}} N_{VP_{fr}} \qquad \overline{N}_{FP_v} = \frac{1}{N_{\text{quadros}}} \sum_{fr=1}^{N_{\text{quadros}}} N_{FP_{fr}}, \tag{4.8}$$

$$\overline{N}_{FN_v} = \frac{1}{N_{\text{quadros}}} \sum_{fr=1}^{N_{\text{quadros}}} N_{FN_{fr}} \qquad \overline{N}_{VN_v} = \frac{1}{N_{\text{quadros}}} \sum_{fr=1}^{N_{\text{quadros}}} N_{VN_{fr}}; \tag{4.9}$$

3. calculam-se os valores médios das métricas de sensibilidade (m_{TVP_v}), precisão (m_{VPP_v}), e medida F (m_{F_v}), relacionadas a cada vídeo do banco de dados SABS, através das equações

$$\overline{m}_{TVP_v} = \frac{\overline{N}_{VP_v}}{\overline{N}_{VP_v} + \overline{N}_{FN_v}}, \tag{4.10}$$

$$\overline{m}_{VPP_v} = \frac{\overline{N}_{VP_v}}{\overline{N}_{VP_v} + \overline{N}_{FP_v}}, \tag{4.11}$$

$$\overline{m}_{F_v} = \frac{2\overline{m}_{VPP_v}\overline{m}_{TVP_v}}{\overline{m}_{VPP_v} + \overline{m}_{TVP_v}}. \tag{4.12}$$

Intuitivamente, os conceitos de precisão e sensibilidade são explicados da seguinte maneira:

- a precisão refere-se à fração do número de píxeis corretamente classificados pela técnica como primeiro plano, em relação ao número total de píxeis classificados também pela técnica como primeiro plano. Ou seja, o conceito de precisão implica em responder à pergunta, que porcentagem dos píxeis classificados como primeiro plano coincide com o *ground-truth*?. Ou, resumidamente, que porcentagem dos píxeis classificados como primeiro plano é valida?. Assim, uma precisão de 100% significa que cada píxel classificado como primeiro plano pela técnica de fato é parte do primeiro plano;
- a sensibilidade refere-se à fração do número de píxeis corretamente classificados pela técnica como primeiro plano em relação ao número total de píxeis rotulados como primeiro plano no *ground-truth*. Ou seja, o conceito de sensibilidade implica em responder à pergunta, que porcentagem do primeiro plano no *ground-truth* é detectada pela técnica?. Ou, resumidamente, que porcentagem dos píxeis validos é classificada como primeiro plano?. Assim, uma sensibilidade de 100% significa que todo píxel rotulado como primeiro plano no *ground-truth* foi classificado como primeiro plano pela técnica;
- finalmente, existe uma relação inversa entre a precisão e sensibilidade, onde é possível aumentar uma em detrimento da redução da outra, de modo a não analisá-las de forma isolada, sendo ambas combinadas através da respectiva medida F, onde um valor alto da medida F indica uma maior exatidão na rotulação.

Sumarizando, o desempenho de uma técnica de subtração de fundo é apresentado em função dos valores de precisão e sensibilidade para cada um dos vídeos do banco de dados, e a comparação entre diferentes técnicas é feita pela análise dos valores da medida F.

Nas seguintes subseções são apresentados os resultados dos testes efetuados utilizando o banco de dados SABS, adotando o procedimento descrito abaixo.

Procedimento 2 Passos para a realização dos testes utilizando o banco de dados SABS.

1. Cada vídeo do banco de dados foi dividido nos conjuntos de treinamento e de teste, indicados na Tabela A.2 (ver página 117);

2. levando-se em conta o cenário de vigilância *básico*, que combina uma série de desafios, é realizado o cálculo da curva de precisão-sensibilidade em função de um parâmetro da técnica, tal que este parâmetro tem um forte impacto no desempenho do resultado. Assim, nestes casos, os parâmetros de interesse são:

   - para as técnicas de subtração de fundo baseadas tanto na média móvel gaussiana como no histograma, trabalhando em conjunto com as técnicas baseadas na distância Euclidiana, é selecionado o nível de significância α como parâmetro a otimizar;
   - para a técnica de subtração de fundo baseada na mistura de Gaussianas, é selecionada a taxa de aprendizagem α_{apr} como parâmetro a otimizar.

   O valor selecionado para o parâmetro otimizado é aquele relacionado ao maior valor da medida F;

3. para cada técnica de subtração de fundo, considerando o valor do parâmetro otimizado, são calculados os valores de precisão e sensibilidade e a medida F, definidos pelas Equações (4.10), (4.11) e (4.12), vinculados a cada vídeo do banco de dados SABS.

Finalmente, na Tabela 4.2 são apresentados os valores numéricos dos parâmetros relacionados a cada técnica de subtração de fundo testada no banco de dados SABS. O critério para a sintonização dos parâmetros é similar ao utilizado na determinação dos parâmetros da Tabela 4.1. Porém, como é indicado no passo 2 do **procedimento 2**, alguns parâmetros são otimizados segundo o maior valor da medida F que apresenta a curva de precisão-sensibilidade em função do parâmetro de interesse (estes valores estão em negrito na Tabela 4.2).

4.5.1 Técnica de Subtração de Fundo Baseada na Média Móvel Gaussiana

Efetuando o passo 2 do procedimento 2, são obtidas as Figuras 4.15 e 4.16, onde observam-se as curvas de precisão-sensibilidade em relação ao nível de significância α. Considerando as técnicas implementadas para o passo de detecção de variações baseadas na distância Euclidiana, tem-se que: (a) a técnica baseada na distância Euclidiana simplificada apresenta

Tabela 4.2: Parâmetros das técnicas de: modelamento do fundo, detecção de variações e pós-processamento, quando é usado o banco de dados SABS.

Técnicas de Subtração de Fundo		Parâmetros
Média móvel Gaussiana	Distância Euclidiana Simplificada	$\alpha_{svt} = 0,3$; $\tau_\theta = 10$, $\boldsymbol{\alpha = 2 \times 10^{-3}}$.
	Distância Euclidiana Bivariada	$\alpha_{svt} = 0,3$; $\boldsymbol{\alpha = 10^{-4}}$.
Histograma	Distância Euclidiana Simplificada	$\sigma_{ds} = 3$; $\tau_\theta = 10$, $\boldsymbol{\alpha = 2 \times 10^{-3}}$.
	Distância Euclidiana Bivariada	$\sigma_{ds} = 3$; $\boldsymbol{\alpha = 3,4 \times 10^{-5}}$.
Mistura de Gaussianas		$K_{mg} = 5$, $D_{vz} = 0,6$, $\boldsymbol{\alpha_{apr} = 4 \times 10^{-3}}$, $\omega_0 = 10^{-3}$, $\sigma_0 = 45$, $\rho = 10^{-3}$,$T_{\text{fundo}} = 0,8$.
Pós-processamento		$N_{\text{MinArea}} = 100$

um valor otimizado para o nível de significância de $\alpha = 0,002$, obtendo uma medida F de $57,4\%$ (ver Figura 4.15); (b) a técnica baseada na distância Euclidiana bivariada apresenta um valor otimizado para o nível de significância de $\alpha = 0,0001$, obtendo uma medida F de $61,9\%$ (ver Figura 4.16).

Efetuando o passo 3 do procedimento 2, são obtidas as Tabelas 4.3 e 4.4, para todas as situações do banco de dados SABS, observando-se que para os cenários:

- básico, com fundo dinâmico e sombras, ambas técnicas de detecção de variações baseadas na distância Euclidiana têm um desempenho razoável em termos da precisão e sensibilidade. Porém, a técnica baseada na distância Euclidiana bivariada apresenta uma maior medida F, uma vez que a técnica baseada na distância Euclidiana simplificada é mais sensível à iluminação do semáforo, como é observado ao se comparar as Figuras 4.17.a e 4.17.b. Cabe indicar aqui que o cenário de fundo dinâmico é a área do quadro correspondente à árvore na sequência básica. Note-se que ambas técnicas têm problemas no início com uma adequada representação dos movimento das folhas;

- *bootstrapping*, aqui ambas técnicas de detecção de variações apresentam uma diminuição forte na sensibilidade em relação ao cenário básico. A técnica baseada na distância Euclidiana bivariada sofre uma diminuição de $61,9\%$ para $40,9\%$, enquanto que a técnica baseada na distância Euclidiana sofre uma diminuição de $58,2\%$ para $41,5\%$, implicando que poucas regiões do primeiro plano foram detectadas. Este problema deve-se principalmente ao fato da técnica baseada na média móvel gaussiana requerem um número de quadros livres de objetos para gerar um modelo do fundo válido;

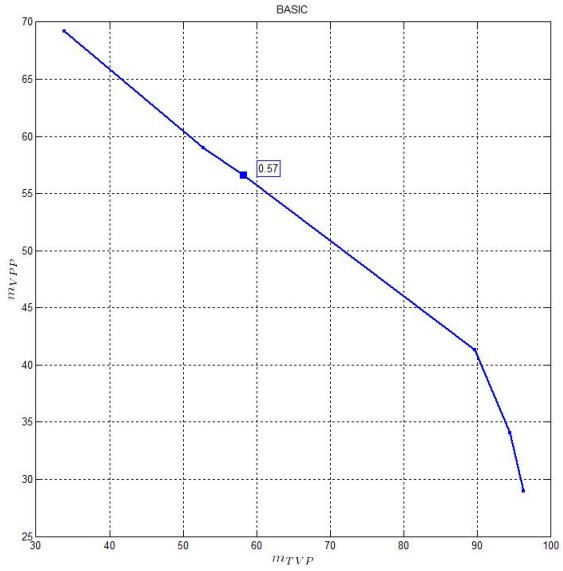
- escurecimento, ambas técnicas de detecção de variações têm um desempenho baixo,

principalmente porque a técnica baseada na média móvel gaussiana se adapta muito lentamente. Exemplos destas falhas na detecção devido à obscurecimento do cenário são apresentados nas Figuras 4.17.c e 4.17.d;

- comutação da iluminação e noite ruidosa, são os cenários que apresentam o maior desafio. Ambas técnicas de detecção de variações têm um desempenho baixo, uma vez que a técnica de modelamento do fundo baseada na média móvel gaussiana não foi capaz de lidar satisfatoriamente com súbitas mudanças na iluminação (cenário comutação da iluminação) e com o baixo contraste primeiro plano/fundo (cenário noite ruidosa), resultando nos problemas de camuflagem e abertura do primeiro plano. Porém, a técnica de detecção de variações mais sensível a este problema é a baseada na distância Euclidiana bivariada, apresentando uma medida F de $15,4\%$ para o cenário de comutação da iluminação e $3,1\%$ para o cenário de noite ruidosa. Exemplos de falhas na detecção (veículos não detectados) e falsos alarmes (janelas detectadas) devido à comutação da iluminação são apresentados nas Figuras 4.17.e e 4.17.f;

- camuflagem, em que ambas técnicas de detecção de variações têm um desempenho razoável em termos da precisão, porém a técnica baseada na distância Euclidiana simplificada apresenta uma maior medida F, devido, principalmente, a um maior valor de precisão ($61,7\%$ em comparação com $50,7\%$). Nas Figuras 4.17.g e 4.17.h é possível ver que a técnica baseada na distância Euclidiana simplificada detecta um dos pedestres, enquanto a técnica baseada na distância Euclidiana bivariada não é capaz de detectá-los. Cabe indicar aqui que o cenário de camuflagem é a área do quadro correspondente à esquina do semáforo, onde os pedestres se encontram;

- compressão de vídeo, em que os resultados mostram que ambas técnicas de detecção de variações não apresentaram uma diminuição de importância no desempenho. Por outro lado, a técnica baseada na distância Euclidiana bivariada se beneficia de um certo grau de compressão, provavelmente devido à eliminação de componentes de alta frequência do ruído (sensor) pelo *codec* (sua medida F aumenta de 38% para 57% à medida que é incrementada a taxa de bits na compressão).

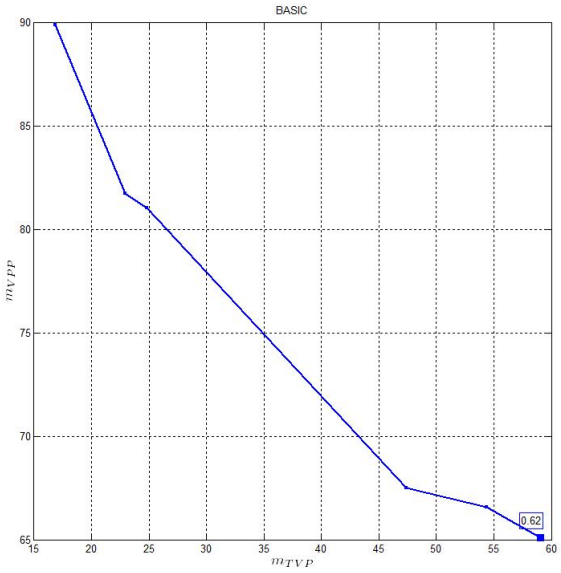


Parâmetro (α)	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
2e-005	33,8%	69,2%	45,4%
0,00101	52,8%	59,0%	55,7%
0,002	**58,2%**	**56,6%**	**57,4%**
0,068	89,7%	41,3%	56,6%
0,134	94,5%	34,1%	50,1%
0,2	96,4%	28,9%	44,5%

Figura 4.15: Gráfico da curva de precisão-sensibilidade e a medida F, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana simplificada, variando o nível de significância α.

Tabela 4.3: Valores para as métricas de sensibilidade, precisão e a medida F, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana simplificada.

Cenários	Vídeos	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
Básico	*Basic*	58,2%	56,6%	57,4%
Fundo dinâmico	*Basic*	61,4%	38,5%	47,3%
Bootstrapping	*Bootstrap*	41,5%	54,9%	47,3%
Escurecimento	*Darkening*	23,0%	64,1%	33,9%
Comutação da iluminação	*LightSwitch*	16,1%	22,2%	18,7%
Noite ruidosa	*NoisyNight*	14,5%	85,1%	24,8%
Sombras	*NoCamouflage*	45,4%	49,1%	47,1%
Camuflagem	*Camouflage*	36,6%	61,5%	45,9%
	NoCamouflage	44,2%	50,7%	47,2%
Compressão de vídeo	*MPEG4_40kbps*	47,7%	63,2%	54,3%
	MPEG4_80kbps	54,2%	60,1%	57,0%
	MPEG4_160kbps	56,2%	56,1%	56,1%
	MPEG4_320kbps	57,6%	56,0%	56,8%
	MPEG4_640kbps	57,4%	56,0%	56,7%



Parâmetro (α)	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
1e-008	16,8%	89,9%	28,3%
5,05e-007	22,9%	81,8%	35,8%
1e-006	24,8%	81,0%	38,0%
3,4e-005	47,4%	67,5%	55,7%
6,7e-005	54,4%	66,6%	59,9%
0,0001	**59,1%**	**65,1%**	**61,9%**

Figura 4.16: Gráfico da curva de precisão-sensibilidade e a medida F, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana bivariada, variando o nível de significância α.

Tabela 4.4: Valores para as métricas de sensibilidade, precisão e a medida F, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana bivariada.

Cenários	Vídeos	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
Básico	*Basic*	59,1%	65,1%	61,9%
Fundo dinâmico	*Basic*	66,2%	45,0%	53,6%
Bootstrapping	*Bootstrap*	30,1%	64,0%	40,9%
Escurecimento	*Darkening*	7,5%	59,2%	13,3%
Comutação da iluminação	*LightSwitch*	12,5%	19,9%	15,4%
Noite ruidosa	*NoisyNight*	1,6%	93,1%	3,1%
Sombras	*NoCamouflage*	43,6%	59,1%	50,1%
Camuflagem	*Camouflage*	22,0%	64,5%	32,8%
	NoCamouflage	39,6%	61,7%	48,3%
Compressão de vídeo	*MPEG4_40kbps*	26,3%	68,2%	38,0%
	MPEG4_80kbps	40,3%	71,8%	51,6%
	MPEG4_160kbps	49,3%	60,7%	54,4%
	MPEG4_320kbps	54,4%	55,5%	54,9%
	MPEG4_640kbps	54,8%	59,3%	57,0%

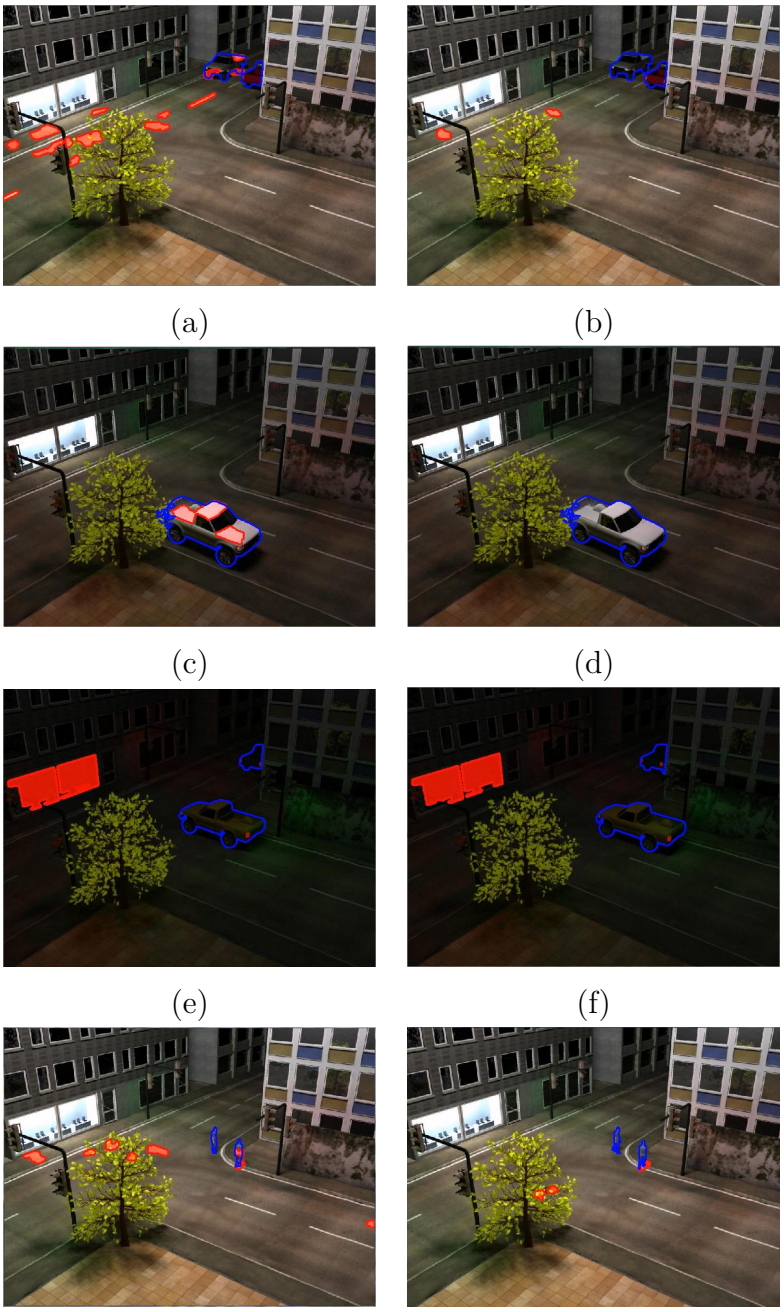


Figura 4.17: Erros nos resultados da detecção, quando é testada a técnica de subtração de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica de detecção de variações baseada (a,c,e,g) na distância Euclidiana simplificada; (b,d,f,h) na distância Euclidiana bivariada; (os contornos de cor azul fazem referência ao *ground-truth*, e as componentes conectadas de cor vermelho indicam os objetos detectados pela técnica de subtração de fundo).

4.5.2 Técnica de Subtração de Fundo Baseada no Histograma

Efetuando o passo 2 do procedimento 2, são obtidas as Figuras 4.18 e 4.19, onde observam-se a curva de precisão-sensibilidade em relação ao nível de significância α. Considerando as técnicas implementadas para o passo de detecção de variações baseadas na distância Euclidiana, tem-se que: (a) a técnica baseada na distância Euclidiana simplificadaapresenta um valor otimizado para o nível de significância de $\alpha = 0,002$, obtendo uma medida F de $64,3\%$ (ver Figura 4.18); (b) a técnica baseada na distância Euclidiana bivariada apresenta um valor otimizado para o nível de significância de $\alpha = 0,000034$, obtendo uma medida F de $58,7\%$ (ver Figura 4.19).

Efetuando o passo 3 do procedimento 2 são obtidas as Tabelas 4.5 e 4.6, para todos as situações do banco de dados SABS, observando-se que para os cenários:

- básico, com fundo dinâmico e sombras, ambas técnicas de detecção de variações baseadas na distância Euclidiana têm um desempenho razoável, em termos da precisão e sensibilidade. Entretanto, a técnica baseada na distância Euclidiana simplificada apresenta uma maior medida F, uma vez que esta técnica lida melhor com o problema da iluminação do semáforo, enquanto que a técnica baseada na distância Euclidiana bivariada é mais susceptível ao fundo dinâmico, como é observado ao se comparar as Figuras 4.20.a e 4.20.b;

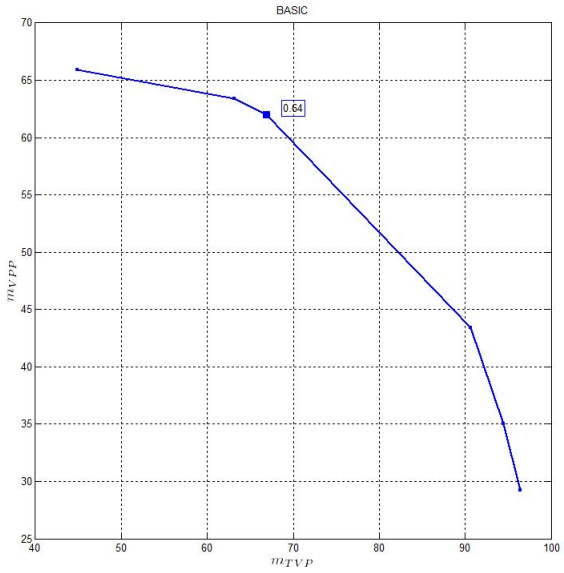
- *bootstrapping*, ambas técnicas de detecção de variações têm um desempenho razoável em termos da precisão e sensibilidade. Porém, a técnica baseada na distância Euclidiana bivariada apresenta uma diminuição de $16,1\%$ para a sensibilidade e de $7,2\%$ para a precisão, em relação ao cenário básico, devido fundamentalmente a que esta abordagem tem dificuldades no caso de um fundo dinâmico (ver Figuras 4.20.c e 4.20.d). Comparados com os resultados obtidos pela técnica de subtração de fundo baseada na média móvel gaussiana, é possível dizer que a técnica baseada no histograma trata melhor o problema de inicialização, já que o histograma, ao armazenar (como frequências) todo o histórico de um processo de um píxel, e definir como modelo do fundo a moda do histograma, é menos susceptível a considerar como fundo um píxel cujo valor de intensidade não se repete com frequência (característica que tem os píxeis pertencentes a um objeto do primeiro plano);

- escurecimento, comutação da iluminação e noite ruidosa, que são os cenários que apresentam o maior desafio. Ambas técnicas de detecção de variações têm um desempenho baixo, devido ao fato que a técnica de modelamento do fundo baseada no histograma não foi capaz de lidar satisfatoriamente com o baixo contraste primeiro plano/fundo

(cenário noite ruidosa), resultando nos problemas de camuflagem e abertura do primeiro plano, produzindo falhas na detecção. Exemplos destas falhas (veículos não detectados) são apresentados nas Figuras 4.20.e e 4.20.f. A técnica de detecção de variações baseada na distância Euclidiana bivariada é mais sensível a este problema, apresentando o menor valor para a medida F (de $2,3\%$ para o cenário noite ruidosa);

- camuflagem, ambas técnicas de detecção de variações têm um desempenho razoável em termos da precisão e sensibilidade. Entretanto, a técnica baseada na distância Euclidiana simplificada apresenta uma medida F levemente maior ($51,3\%$ em comparação com $50,4\%$ da técnica baseada na distância Euclidiana bivariada). Nas Figuras 4.20.g e 4.20.h é possível ver que ambas técnicas de detecção de variações detectam os pedestres, porém a técnica baseada na distância Euclidiana simplificada detecta a um deles quase completamente;

- compressão de vídeo, os resultados mostram que ambas técnicas de detecção de variações não apresentaram uma diminuição significativa no desempenho. A técnica baseada na distância Euclidiana simplificada mantém quase o mesmo valor de medida F à medida que é incrementada a taxa de bits na compressão, e a técnica baseada na distância Euclidiana bivariada (diferentemente do caso anterior, considerando a técnica de subtração de fundo baseada na média móvel gaussiana) não tem benefícios após da compressão do vídeo.

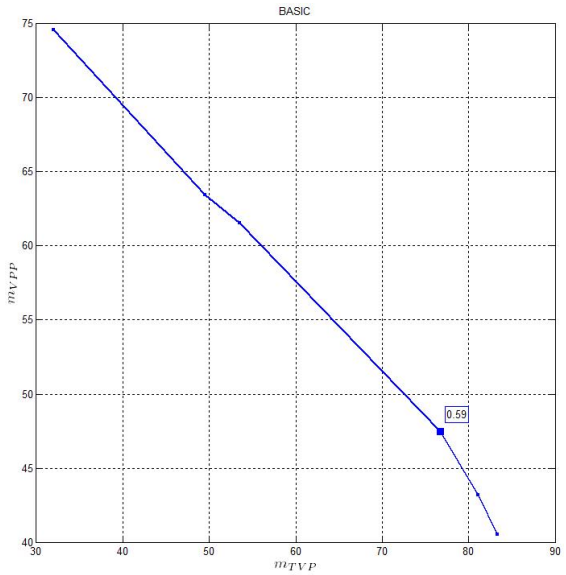


Parâmetro (α)	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
2e-005	44,9%	65,9%	53,4%
0,00101	63,1%	63,4%	63,2%
0,002	**66,9%**	**62,0%**	**64,3%**
0,068	90,6%	43,4%	58,7%
0,134	94,5%	35,0%	51,1%
0,2	96,4%	29,2%	44,9%

Figura 4.18: Gráfico da curva de precisão-sensibilidade e a medida F, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana simplificada, variando o nível de significância α.

Tabela 4.5: Valores para as métricas de sensibilidade, precisão e a medida F, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana simplificada.

Cenários	Vídeos	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
Básico	*Basic*	66,9%	62,0%	64,3%
Fundo dinâmico	*Basic*	73,5%	40,5%	52,2%
Bootstrapping	*Bootstrap*	55,2%	55,1%	55,2%
Escurecimento	*Darkening*	30,3%	46,6%	36,7%
Comutação da iluminação	*LightSwitch*	14,6%	16,8%	15,6%
Noite ruidosa	*NoisyNight*	17,7%	75,6%	28,7%
Sombras	*NoCamouflage*	53,0%	51,2%	52,1%
Camuflagem	*Camouflage*	44,2%	61,2%	51,3%
	NoCamouflage	50,4%	65,5%	57,0%
Compressão de vídeo	*MPEG4_40kbps*	66,0%	63,5%	64,7%
	MPEG4_80kbps	66,2%	62,7%	64,4%
	MPEG4_160kbps	66,2%	61,9%	63,9%
	MPEG4_320kbps	66,1%	62,4%	64,2%
	MPEG4_640kbps	67,3%	62,3%	64,7%



Parâmetro (α)	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
1e-008	32,0%	74,6%	44,8%
5,05e-007	49,5%	63,5%	55,6%
1e-006	53,5%	61,6%	57,3%
3,4e-005	**76,7%**	**47,5%**	**58,7%**
6,7e-005	81,1%	43,2%	56,4%
0,0001	83,4%	40,5%	54,5%

Figura 4.19: Gráfico da curva de precisão-sensibilidade e a medida F, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana bivariada, variando o nível de significância α.

Tabela 4.6: Valores para as métricas de sensibilidade, precisão e a medida F, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana bivariada.

Cenários	Vídeos	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
Básico	*Basic*	76,7%	47,5%	58,7%
Fundo dinâmico	*Basic*	86,8%	21,5%	34,4%
Bootstrapping	*Bootstrap*	60,6%	40,3%	48,5%
Escurecimento	*Darkening*	8,2%	46,2%	13,9%
Comutação da iluminação	*LightSwitch*	14,2%	15,2%	14,7%
Noite ruidosa	*NoisyNight*	1,2%	91,6%	2,3%
Sombras	*NoCamouflage*	59,7%	38,6%	46,9%
Camuflagem	*Camouflage*	42,0%	63,0%	50,4%
	NoCamouflage	53,6%	67,2%	59,6%
Compressão de vídeo	*MPEG4_40kbps*	69,1%	53,6%	60,4%
	MPEG4_80kbps	73,4%	45,8%	56,4%
	MPEG4_160kbps	76,6%	40,9%	53,4%
	MPEG4_320kbps	75,8%	42,5%	54,5%
	MPEG4_640kbps	83,2%	34,3%	48,6%

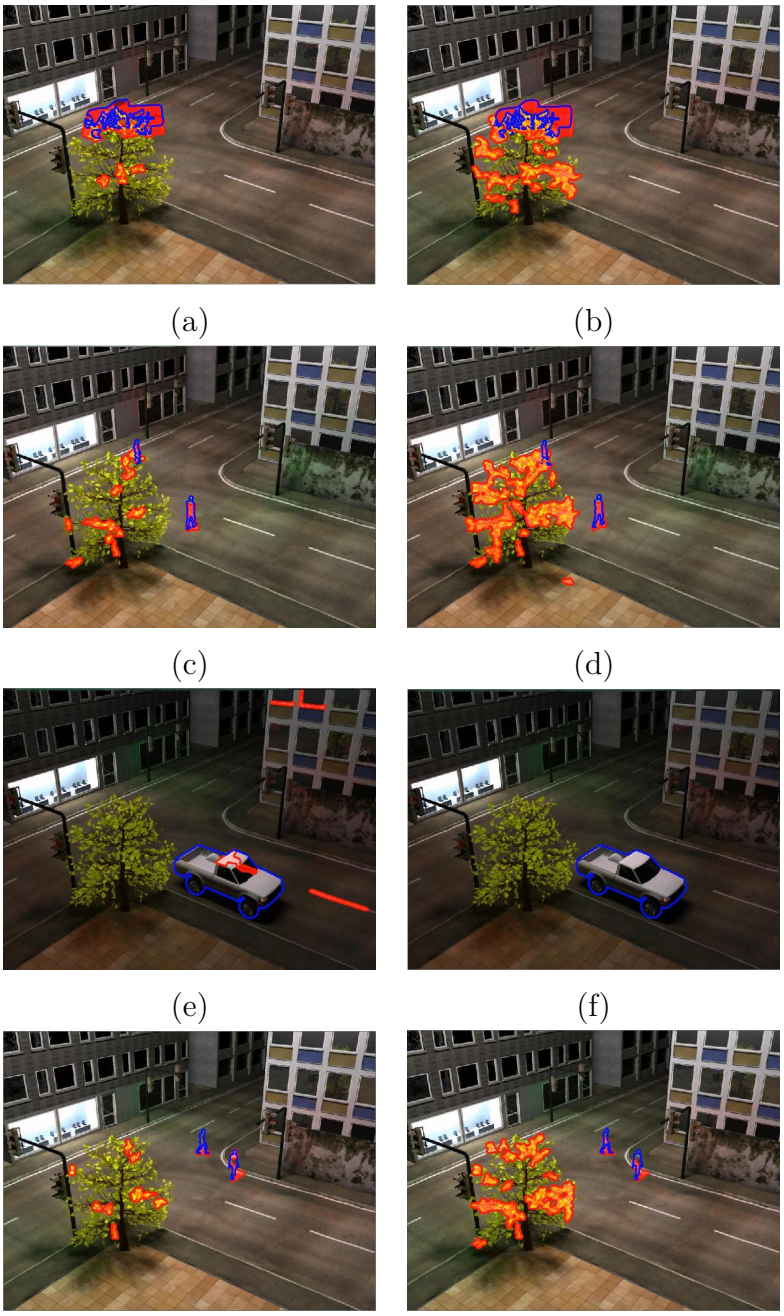


Figura 4.20: Erros nos resultados da detecção, quando é testada a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com a técnica de detecção de variações baseada (a,c,e,g) na distância Euclidiana simplificada; (b,d,f,h) na distância Euclidiana bivariada; (os contornos de cor azul fazem referência ao *ground-truth*, e as componentes conectadas de cor vermelho indicam os objetos detectados pela técnica de subtração de fundo).

4.5.3 Técnica de Subtração de Fundo Baseada na Mistura de Gaussianas

Efetuando o passo 2 do procedimento 2, é obtida a Figura 4.21, onde observa-se a curva de precisão-sensibilidade em relação à taxa de aprendizagem α_{apr}. Esta técnica apresenta um valor otimizado para a taxa de aprendizagem de $\alpha_{apr} = 0,004$, obtendo uma medida F de $72,8\%$.

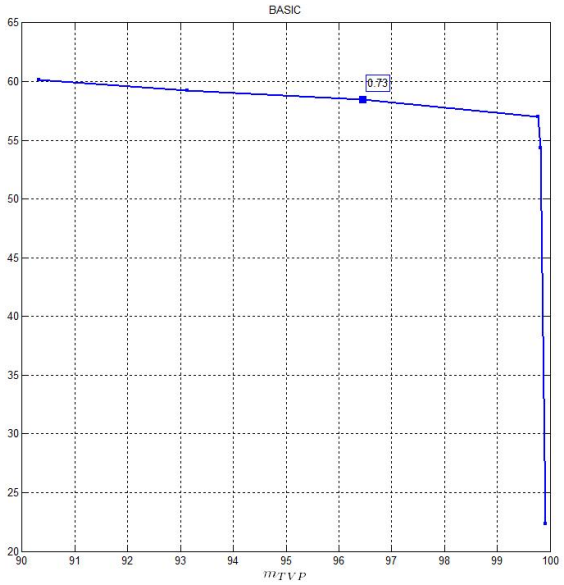
Efetuando o passo 3 do procedimento 2, é obtida a Tabela 4.7, para todas as situações do banco de dados SABS, observando-se que para os cenários:

- básico, com fundo dinâmico e sombras, esta técnica tem um desempenho superior às duas abordagens anteriores, devido a ter uma sensibilidade superior ao 96% nos três cenários, indicando que esta técnica detecta a maioria da área correspondente ao primeiro plano no *ground-truth*. Entretanto, a técnica também classifica partes da cena como primeiro plano quando elas não o são, apresentando assim uma precisão inferior a $58,5\%$ nos três casos. Esta técnica, ao modelar o processo de um píxel através de uma distribuição multimodal, trata bem o problema da representação do movimento das folhas no cenário correspondente ao fundo dinâmico, como é observado nas Figuras 4.22.a, 4.22.c e 4.22.d;

- *bootstrapping*, onde a técnica apresenta uma diminuição na sensibilidade em relação ao cenário básico de $4,9\%$. Porém, seus resultados ainda são superiores às abordagens anteriores, por se adaptar rapidamente às variações presentes em cada processo de um píxel;

- escurecimento, comutação da iluminação e noite ruidosa, que são os cenários que apresentam o maior desafio. Esta abordagem tem um menor desempenho no cenário comutação da iluminação, não sendo capaz de lidar satisfatoriamente com súbitas mudanças na iluminação, resultando numa sobre-detecção do primeiro plano (ver Figura 4.22.b) apresentando assim um baixo valor de precisão ($m_{VPP} = 36\%$). Entretanto, seu desempenho nos cenários escurecimento e noite ruidosa é razoável, devido principalmente a: *a*) um alto valor da sensibilidade ($m_{TVP} = 91,7\%$) para o cenário escurecimento; *b*) um alto valor de precisão ($m_{TVP} = 83\%$) para o cenário noite ruidosa. Um exemplo da detecção da área da cena que contém o veículo no cenário escurecimento é apresentado na Figura 4.22.c, podendo-se ver que detecta toda a área correspondente ao veículo (alta sensibilidade), detectando, entretanto, áreas que não correspondem a ele (baixa precisão);

- camuflagem, esta técnica tem um bom desempenho em termos da precisão e sensibilidade, apresentando, como na maioria dos cenários, um valor da sensibilidade superior a 90%. Na Figura 4.22.d é possível ver que esta abordagem detecta os dois pedestres;
- compressão de vídeo, em que os resultados mostram que esta técnica se beneficia de um certo grau de compressão, já que sua medida F é superior à obtida no cenário básico, considerando todas as taxas de bits na compressão.

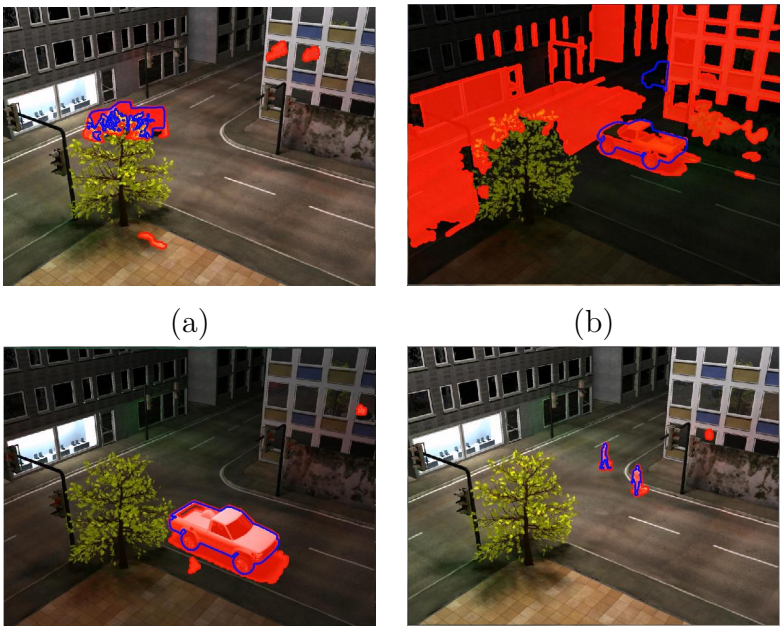


Parâmetro (α_{apr})	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
0,01	84,3%	61,1%	70,8%
0,0085	87,2%	60,5%	71,4%
0,007	90,3%	60,1%	72,2%
0,0055	93,1%	59,3%	72,4%
0,004	**96,5%**	**58,4%**	**72,8%**
0,0025	99,8%	57,0%	72,5%
0,001375	99,8%	54,4%	70,4%
0,00025	99,9%	22,3%	36,5%

Figura 4.21: Gráfico da curva de precisão-sensibilidade e a medida F, quando é testada a técnica de subtração de fundo baseada na mistura de Gaussianas, variando a taxa de aprendizagem α_{apr}.

Tabela 4.7: Valores para as métricas de sensibilidade, precisão e a medida F, quando é testada a técnica de subtração de fundo baseada na mistura de Gaussianas.

Cenários	Vídeos	sensibilidade (m_{TVP})	precisão (m_{VPP})	medida F (m_F)
Básico	*Basic*	96,4%	58,5%	72,8%
Fundo dinâmico	*Basic*	97,4%	47,9%	64,2%
Bootstrapping	*Bootstrap*	91,5%	50,7%	65,3%
Escurecimento	*Darkening*	91,7%	42,0%	57,6%
Comutação da iluminação	*LightSwitch*	54,6%	26,8%	36,0%
Noite ruidosa	*NoisyNight*	32,8%	83,0%	47,0%
Sombras	*NoCamouflage*	97,0%	53,0%	68,5%
Camuflagem	*Camouflage*	96,8%	61,9%	75,5%
	NoCamouflage	97,1%	61,9%	75,6%
Compressão de vídeo	*MPEG4_40kbps*	96,1%	60,5%	74,2%
	MPEG4_80kbps	96,3%	59,9%	73,8%
	MPEG4_160kbps	96,2%	59,6%	73,6%
	MPEG4_320kbps	96,4%	59,7%	73,8%
	MPEG4_640kbps	96,4%	59,9%	73,9%



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

Figura 4.22: Erros nos resultados da detecção, quando é testada a técnica de subtração de fundo baseada na mistura de Gaussianas (os contornos de cor azul fazem referência ao *ground-truth*, e as componentes conectadas de cor vermelho indicam os objetos detectados pela técnica de subtração de fundo).

4.6 Resumo

Neste capítulo foram realizadas os testes para a avaliação das técnicas de subtração de fundo implementadas utilizando os bancos de dados PETS2004 e SABS. Assim, tem-se que:

- em relação ao banco de dados PETS2004, nas Tabelas 4.8, 4.9 e 4.10 são apresentadas as métricas de exatidão global, vinculadas às falhas na detecção e aos falsos alarmes, respectivamente. Da Tabela 4.8, é possível concluir que a técnica de detecção de variações que apresenta o melhor desempenho é a técnica proposta baseada na distância Euclidiana simplificada, chegando a ter uma métrica de exatidão global de $67,2\%$, seguida pela técnica baseada na mistura de Gaussianas, com uma exatidão de $62,8\%$. Sua principal fonte de erro está nas falhas na detecção, se mantendo quase constante, variando entre $18,4\%$ e $19,1\%$ (ver Tabela 4.9). Para o caso dos falsos alarmes ela varia entre 10% e $19,7\%$ (ver Tabela 4.10). Do ponto de vista da técnica de subtração de fundo, a melhor resposta em todos os casos é para a técnica baseada no histograma, seguida pela técnica baseada na mistura de Gaussianas. Este resultado é explicado considerando que a técnica de subtração de fundo baseada no histograma trata melhor o problema de primeiro plano adormecido em relação à técnica de subtração de fundo baseada na mistura de Gaussianas, problema que é comum em situações de vigilância, portanto presente no banco de dados PETS2004;
- em relação ao banco de dados SABS, na Tabela 4.11 são apresentadas as medida F vinculadas a cada técnica de subtração de fundo implementada, observando-se que para os cenários:
  - básico, com fundo dinâmico e sombras, os modelos do fundo baseados na média móvel gaussiana e no histograma têm problemas com uma adequada representação do movimento das folhas, e também mostram pontos fracos com a iluminação do semáforo. Por outro lado, a técnica de subtração de fundo baseada na mistura de Gaussianas apresenta a melhor representação para um fundo dinâmico;
  - *bootstrapping*, cenário em que todas as técnicas de subtração de fundo apresentam uma diminuição na medida F quando operam sem fase de treinamento. No entanto, a técnica de subtração de fundo baseada na média móvel gaussiana, trabalhando em conjunto com a técnica baseada na distância Euclidiana bivariada, apresenta a maior perda de desempenho;
  - escurecimento, cenário em que todas as técnicas de subtração de fundo apresentam uma diminuição na medida F, sendo a mais afetada a técnica de subtração

de fundo baseada na média móvel gaussiana trabalhando em conjunto com a técnica baseada na distância Euclidiana bivariada. Isto ocorre porque a média móvel gaussiana se adapta muito lentamente à medida que o cenário vai escurecendo;

- comutação da iluminação e noite ruidosa, são cenários em que todas as técnicas de subtração de fundo apresentam uma diminuição forte na medida F, sendo estas as experiências que apresentam o maior desafio. Assim, obteve-se um desempenho baixo para todas as técnicas avaliadas, sendo a mais afetada a técnicas de detecção de variações baseada na distância Euclidiana bivariada;
- camuflagem, para o qual, além dos bons resultados da técnica de subtração de fundo baseada na mistura de Gaussianas, a técnica de subtração de fundo baseada no histograma trabalhando em conjunto com as técnicas baseadas na distância Euclidiana também apresenta um resultado razoável em termos de precisão e sensibilidade;
- compressão de vídeo, em que os resultados mostram que a maioria das técnicas não apresentaram uma diminuição considerável na medida F. Somente a técnica de subtração de fundo baseada na média móvel gaussiana tem uma diminuição na medida F para as baixas taxas de bits. Concomitantemente, a maioria das técnicas se beneficiam de um certo grau de compressão, provavelmente devido à eliminação de componentes de alta frequência do ruído pelo *codec*;

- em relação ao tempo de processamento das técnicas implementadas, na Tabela 4.12 são indicados os tempos médios de processamento por quadro em segundos, quando foi utilizado um processador *Intel Core i5* com 2.4GHz e 4GB de RAM e implementações em MATLAB. Se observa que,
  - considerando as técnicas de subtração de fundo, a técnica que apresenta o menor tempo de processamento é a baseada na média móvel Gaussiana, uma vez que ela basicamente realiza uma operação de atualização definida pela Equação (2.6), a técnica que apresenta o maior tempo de processamento é a baseada no histograma, uma vez que esta técnica tem que atualizar três histogramas (um para cada canal de cor) para cada píxel de um quadro, além de determinar de forma automática a maioria de seus parâmetros;
  - considerando as técnicas de detecção de variações, em geral as três técnicas apresentam tempos de processamento razonáveis, sendo a técnica que apresenta o maior tempo de processamento, a baseada no teste de significância, produto que ela requer de uma operação recursiva de limiarização;
  - considerando à etapa de pós-processamento, ela apresenta um tempo de processamento razonável, sendo, o filtragem das componentes conectadas menores a N_{MinArea} píxeis, a operação que maior tempo consume.

Tabela 4.8: Métrica de exatidão $m_{E_{db}}$ para o melhor caso, considerando todas as técnicas de modelamento do fundo e cada uma das técnicas de detecção de variações.

		Técnicas de Detecção de Variações			
	$m_{E_{db}}$	Teste de significância	Distância Euclidiana		Mistura de Gaussianas
			Simplificada	Bivariada	
Técnicas de Modelamento do Fundo	Média móvel Gaussiana	49,5%	58,6%	49,9%	
	Histograma	52,5%	**67,2%**	60,8%	
	Mistura de Gaussianas				62,8%

Tabela 4.9: Métrica de exatidão vinculada às falhas na detecção $m_{E-FD_{db}}$ para o melhor caso, considerando todas as técnicas de modelamento do fundo e cada uma das técnicas de detecção de variações.

		Técnicas de Detecção de Variações			
	$m_{E_{db}}$	Teste de significância	Distância Euclidiana		Mistura de Gaussianas
			Simplificada	Bivariada	
Técnicas de Modelamento do Fundo	Média móvel Gaussiana	26,1%	18,4%	29%	
	Histograma	27,5%	19,1%	25,3%	
	Mistura de Gaussianas				27%

Tabela 4.10: Métrica de exatidão vinculada às falsas alarmes $m_{E-FA_{db}}$ para o melhor caso, considerando todas as técnicas de modelamento do fundo e cada uma das técnicas de detecção de variações.

		Técnicas de Detecção de Variações			
	$m_{E_{db}}$	Teste de significância	Distância Euclidiana		Mistura de Gaussianas
			Simplificada	Bivariada	
Técnicas de Modelamento do Fundo	Média móvel Gaussiana	21,1%	19,7%	17,7%	
	Histograma	16,9%	10%	10,5%	
	Mistura de Gaussianas				27%

Tabela 4.11: Valores para a medida F, considerando todas as técnicas de modelamento do fundo e cada uma das técnicas de detecção de variações.

Cenários	Média móvel Gaussiana + Distância Euclidiana		Histograma + Distância Euclidiana		Mistura de Gaussianas
	Simplificada	Bivariada	Simplificada	Bivariada	
Básico	57,4%	61,9%	64,3%	58,7%	72,8%
Fundo dinâmico	47,3%	53,6%	52,2%	34,4%	64,2%
Bootstrapping	47,3%	40,9%	55,2%	48,5%	65,3%
Escurecimento	33,9%	13,3%	36,7%	13,9%	57,6%
Comutação da iluminação	18,7%	15,4%	15,6%	14,7%	36,0%
Noite ruidosa	24,8%	3,1%	28,7%	2,3%	47,0%
Sombras	47,1%	50,1%	52,1%	46,9%	68,5%
Camuflagem	45,9%	32,8%	51,3%	50,4%	75,5%
Compressão de vídeo	54,3%	38,0%	64,7%	60,4%	74,2%
	57,0%	51,6%	64,4%	56,4%	73,8%
	56,1%	54,4%	63,9%	53,4%	73,6%
	56,8%	54,9%	64,2%	54,5%	73,8%
	56,7%	57,0%	64,7%	48,6%	73,9%

Tabela 4.12: Tempos de processamento das técnicas de: modelamento do fundo, detecção de variações e pós-processamento.

		Tempo de Processamento por quadro (seg)
Técnicas de Modelamento do Fundo	Média móvel Gaussiana	0,035
	Histograma	6,58
	Mistura de Gaussianas	0,8
Técnicas de Detecção de Variações	Distância Euclidiana Simplificada	0,19
	Distância Euclidiana Bivariada	0,07
	Teste de Significância	0,77
Pós-processamento		0,18

Capítulo 5

Conclusões e Projetos Futuros

5.1 Conclusões

A partir dos objetivos definidos para a elaboração desta tese, podem-se destacar as seguintes conclusões:

Quanto às técnica de modelamento do fundo: foram implementadas três técnicas, a primeira baseada na média móvel gaussiana, a segunda baseada na mistura de Gaussianas e a terceira baseada no histograma. Assim, considerando a resposta destas técnicas utilizando o banco de dados SABS, observa-se que o modelo do fundo baseado na mistura de Gaussianas mostra os melhores resultados em termos da medida F, devido principalmente à capacidade deste modelo de tratar os problemas derivados de um fundo dinâmico. Entretanto, quando são considerados os resultados obtidos utilizando o banco de dados PETS2004, observa-se que o modelo do fundo baseado no histograma exibe a melhor resposta em termos da métrica de exatidão, devido principalmente à capacidade deste modelo de tratar o problema de primeiro plano adormecido, o qual é comum em aplicações de monitoramento. A divergência de resultados ao considerar ambos bancos de dados deve-se a que os cenários presentes no banco de dados SABS (gerados artificialmente) não apresentam o problema de primeiro plano adormecido já que a maioria dos objetos do primeiro plano (veículos e pedestres), a todo momento, estão se deslocando. Por outro lado, o banco de dados PETS2004, apesar de apresentar o problema de fundo dinâmico, sua ocorrência é pouco frequente. Uma conclusão de interesse, ratificada na literatura, é que para tratar o problema do fundo dinâmico um modelo multimodal é uma boa opção, porém esta solução padece de outros inconvenientes (no caso estudado é representado pelo problema do primeiro plano adormecido).

Quanto às técnicas de detecção de variações: foram implementadas três técnicas, a primeira baseada no teste de significância, a segunda baseada na distância Euclidiana simplificada e a terceira, uma generalização natural da técnica baseada na distância Euclidiana simplificada, denominada como distância Euclidiana bivariada. Aqui, destacam-se as abordagens baseadas na distância Euclidiana, já que são uma contribuição desta Tese, que tem como ideia principal a determinação da máscara do primeiro plano a partir da limiarização da resposta de uma função discriminante dependente da distância Euclidiana. Ambas propostas são capazes de operar no nível de cores, aproveitando não somente a informação de brilho (que é o caso ao trabalhar na escala de cinza, tal como é feito na técnica baseada no teste de significância) como também a informação de cromaticidade, além de fazer a detecção de variações utilizando um limiar automatizado, que dá uma maior capacidade de adaptação em relação aos desafios a tratar. Em termos dos resultados obtidos com os dois bancos de dados tem-se: (a) considerando-se a resposta destas técnicas utilizando o banco de dados SABS, observa-se que a técnica de subtração de fundo baseada no histograma, trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana simplificada, apresenta um desempenho razoável em termos da precisão e sensibilidade, sendo afetada pelos problemas de escurecimento, comutação da iluminação e noite ruidosa, situações que exibem um maior desafio para todas as técnicas testadas. Já no caso da técnica de detecção de variações baseada na distância Euclidiana bivariada, o desempenho menor deve-se ao fato da técnica ser fortemente afetada pelas três situações já mencionadas; (b) considerando-se a resposta destas técnicas utilizando o banco de dados PETS2004, observa-se que a técnica de subtração de fundo baseada no histograma, trabalhando em conjunto com a técnica de detecção de variações baseada na distância Euclidiana simplificada, apresenta um bom desempenho em termos da métrica de exatidão, obtendo uma taxa de exatidão global, sobre todo o banco de dados PETS2004, de $67,2\%$, o qual é o maior valor de exatidão alcançado nos testes. Um ponto de interesse a considerar é o porquê da técnica de detecção de variações, baseada na distância Euclidiana bivariada, ter um resultado inferior considerando os dois bancos de dados testados, tendo ela a capacidade de tratar separadamente a informação da diferença de magnitudes (informação do brilho) e da diferença angular (informação da matiz). Ao se examinar o modelo probabilístico usado para cada uma das variáveis da distância Euclidiana bivariada (distribuições de probabilidade exponenciais), já que a partir delas é obtida uma expressão para o limiar, observou-se (experimentalmente) que o limiar é bastante conservador, ou seja, no intuito de evitar a detecção de falsos positivos (representado por um valor baixo para o nível de significância α) o limiar evita rotular áreas correspondentes ao primeiro plano.

Quanto à avaliação das técnicas de subtração de fundo: levando-se em conta as diferentes vantagens que apresenta um procedimento de avaliação baseado em métricas orientadas a objetos, foi proposta uma abordagem que permite calcular a métrica de exatidão,

quer seja para cada vídeo ou para todo o banco de dados, sendo essa uma outra contribuição dessa Tese. Esta métrica representa quão exata é uma técnica de subtração de fundo em encontrar os objetos em movimento ao longo dos quadros de um vídeo, considerando que tem-se o *ground-truth* do banco de dados a analisar. Sendo assim, o desempenho de uma técnica de subtração de fundo fica representado na curva de exatidão em relação ao parâmetro $\tau_{\text{casamento}}$, e examinando seus pontos de inflexão, são determinados os valores para $\tau_{\text{casamento}}$ que ponderam o fato de se considerar tanto uma pequena como uma grande sobreposição entre retângulos. Uma característica importante observada é que a curva de exatidão não tem as restrições que apresentam tanto a curva de precisão-sensibilidade como a curva ROC, quando são consideradas como curvas de desempenho para uma avaliação das técnicas de subtração de fundo no nível de objetos. Este procedimento, aqui proposto, foi empregado para avaliar as técnicas de subtração de fundo implementadas, utilizando o banco de dados PETS2004, já que seu *ground-truth* contém todos os objetos rotulados através de retângulos, possibilitando efetuar um procedimento de avaliação baseado em métricas orientadas a objetos.

5.2 Temas a Serem Pesquisados

Considerando os resultados obtidos para as técnicas de subtração de fundo, tem-se: (a) em relação à técnica de subtração de fundo baseada no histograma, para poder superar o problema de fundo dinâmico, uma alternativa seria definir o modelo do fundo a partir de um conjunto de pontos representativos do histograma, e não unicamente através do ponto representativo definido pela moda. Também, é de interesse elaborar um procedimento algorítmico que permita diminuir o custo computacional dessa abordagem (em termos de uso de memória), uma vez que na maioria dos casos os histogramas são esparsos, sendo possível a utilização de uma estrutura de dados (por exemplo uma estrutura tipo árvore) que permita armazenar os histogramas de maneira eficiente; (b) em relação à técnica de subtração de fundo baseada na mistura de Gaussianas, ela é afetada pelo problema de primeiro plano adormecido, devendo-se basicamente ao fato dos pesos das misturas serem atualizados rapidamente. No entanto, tal como é indicado na Equação (2.9), a aprendizagem dos pesos é controlada pelo parâmetro α_{apr}. Porém, ao levar em conta que todos os pesos, em cada iteração, são normalizados para que sua soma seja 1, a atualização não depende mais diretamente de α_{apr}, sendo esse o motivo de α_{apr} assumir valores próximos a zero. Portanto, um ponto a pesquisar seria a elaboração de um procedimento de atualização dos pesos que gere um crescimento/decrescimento linear, mantendo a restrição que sua soma deva ser 1.

Considerando os resultados obtidos para as técnicas de detecção de variações, é de interesse continuar desenvolvendo a abordagem baseada na distância Euclidiana bivariada, a fim de determinar aqueles modelos probabilísticos que representem com maior precisão o comportamento estatístico das componentes da distância Euclidiana bivariada. Também é de interesse estudar a possibilidade de implementar uma técnica de modelamento do fundo baseada nesta proposta, já que ela, de maneira natural, contém a informação de proximidade (ou vizinhança) entre amostras consecutivas, permitindo definir o modelo do fundo como aquele que tenha um maior número de vizinhos, onde o critério de vizinhança pode ser definido no plano das variáveis que contêm a informação da diferença de magnitudes e da diferença angular.

Considerando o procedimento de avaliação baseado em métricas orientadas a objetos proposto, é de interesse avaliar as técnicas de subtração de fundo implementadas aqui, utilizando o banco de dados sugerido em [8], já que tal banco contém cenários: (a) de ambientes internos e externos, (b) que exibem diferentes condições climáticas (chuva, vento e variação de luminosidade do dia) e (c) vídeos de um cenário capturado de diferentes pontos de vista. Salienta-se que todos os objetos do primeiro plano que compõem o *ground-truth* deste banco de dados são rotulados através de retângulos, sendo essa uma condição necessária para efetuar uma avaliação orientada a objetos.

Apêndice A

Tabelas dos Bancos de Dados

Tabela A.1: Composição do banco de dados PETS2004[a].

Cenário	Vídeo	quadros para treinamento	quadros para teste
Procura	242 - 349 (10,4%)	350 - 750 (38,4%)	
	Browse2.mpg	249 - 296, 858 - 874 (7,4%)	300 - 560 (29,8%)
	Browse3.mpg	848 - 910 (6,9%)	222 - 850 (69,0%)
	Browse4.mpg	1098 - 1138 (3,6%)	104 - 600, 829 - 1092 (66,8%)
	Browse_WhileWaiting1.mpg	695 - 790 (12,1%)	0 - 692 (87,6%)
	Browse_WhileWaiting2.mpg	0 - 370 (19,6%)	826 - 1438 (32,3%)
Deixando objetos	*LeftBag.mpg*	610 - 735 (8,8%)	414 - 853 (30,6%)
	LeftBag_AtChair.mpg	648 - 741 (8,4%)	255 - 640, 742 - 1114 (68,1%)
	LeftBag_BehindChair.mpg	698 - 800 (9,7%)	159 - 697 (50,5%)
	LeftBag_PickedUp.mpg	203 - 266, 1152 - 1354 (19,7%)	267 - 874 (44,9%)
	LeftBox.mpg	0 - 70 (8,2%)	412 - 855 (51,4%)
Encontros	*Meet_Crowd.mpg*	0 - 20, 380 - 490 (26,9%)	26 - 375 (71,3%)
	Meet_Split_3rdGuy.mpg	0 - 42 (4,7%)	50 - 550 (54,3%)
	Meet_WalkSplit.mpg	550 - 620 (11,4%)	52 - 384 (53,5%)
	Meet_WalkTogether1.mpg	0 - 112, 522 - 610 (28,6%)	117 - 496 (53,7%)
	Meet_WalkTogether2.mpg	110 - 166 (6,9%)	167 - 438 (32,9%)
	Split.mpg	408 - 526 (21,6%)	0 - 377 (68,5%)
Descansando, desmaiado ou lutando	*Rest_FallOnFloor.mpg*	366 - 477 (11,1%)	478 - 998 (51,7%)
	Rest_InChair.mpg	227 - 246 (2,0%)	247 - 507 (25,9%)
	Rest_SlumpOnFloor.mpg	247 - 276 (3,3%)	314 - 647 (36,7%)
	Rest_WiggleOnFloor.mpg	297 - 478 (14,1%)	479 - 721 (18,8%)
	Fight_Chase.mpg	157 - 166 (2,3%)	191 - 425 (54,5%)
	Fight_OneManDown.mpg	806 - 958 (16,0%)	119 - 515 (41,4%)
	Fight_RunAway1.mpg	100 - 172 (13,2%)	173 - 480 (55,9%)
	Fight_RunAway2.mpg	150 - 190 (7,4%)	191 - 500 (56,3%)
Caminhando	*Walk1.mpg*	0 - 15 (2,6%)	234 - 514 (46,0%)
	Walk2.mpg	629 - 821 (18,3%)	294 - 571 (26,4%)
	Walk3.mpg	0 - 30 (2,2%)	210 - 384, 483 - 845, 1116 - 1378 (58,1%)

[a]As percentagens entre parênteses indicam a proporção dos quadros (quer seja para teste ou treino) em relação ao número total de quadros do vídeo em questão.

Tabela A.2: Composição do banco de dados SABS.

Cenários	Treino		Teste		área de ensaio
	Vídeos	quadros	Vídeos	quadros	
Básico	*NoForegroundDay*	0-800	*Basic*	0 - 599	
Fundo dinâmico	*NoForegroundDay*	0-800	*Basic*	0 - 599	i:200-560 j:100-380
Bootstrapping	*Bootstrap*	0 - 0	*Bootstrap*	1 - 1400	
Escurecimento	*NoForegroundDay*	0 - 800	*Darkening*	0 - 1399	
Comutação da iluminação	*NoForegroundNight*	0 - 800	*LightSwitch*	0 - 599	
Noite ruidosa	*NoForegroundNightNoisy*	0 - 800	*NoisyNight*	0 - 599	
Sombras	*NoForegroundDay*	0 - 800	*NoCamouflage*	0 - 599	
Camuflagem	*NoForegroundDay* *NoForegroundDay*	0 - 800 0 - 800	*Camouflage* *NoCamouflage*	0 - 599 0 - 599	i:150-500 j:400-600
Compressão de vídeo	*MPEG4_40kbps* *MPEG4_80kbps* *MPEG4_160kbps* *MPEG4_320kbps* *MPEG4_640kbps*	0 - 800	*MPEG4_40kbps* *MPEG4_80kbps* *MPEG4_160kbps* *MPEG4_320kbps* *MPEG4_640kbps*	0 - 599	

Apêndice B

Tabelas de Resultados

Tabela B.1: Valores da matriz de confusão e as métricas de exatidão considerando o parâmetro de sobreposição τ_L, quando é testada a técnica de subtração de fundo baseada numa média móvel gaussiana, utilizando a técnica de detecção de variações baseada na distância Euclidiana.

Vídeos	Matriz de confusão						Exatidão					
	N_{CD}	N_{SU}	N_S	N_U	N_{FD}	N_{FA}	m_{E-CD}	m_{E-SU}	m_{E-S}	m_{E-U}	m_{E-FD}	m_{E-FA}
Browse1.mpg	775	0	0	0	27	3	96,3%	0,0%	0,0%	0,0%	3,4%	0,4%
Browse2.mpg	252	0	0	0	9	5	94,7%	0,0%	0,0%	0,0%	3,4%	1,9%
Browse3.mpg	549	0	0	0	230	36	67,4%	0,0%	0,0%	0,0%	28,2%	4,4%
Browse4.mpg	741	0	0	0	26	37	92,2%	0,0%	0,0%	0,0%	3,2%	4,6%
Browse_WhileWaiting1.mpg	606	0	0	0	602	50	48,2%	0,0%	0,0%	0,0%	47,9%	4,0%
Browse_WhileWaiting2.mpg	610	0	0	0	3	622	49,4%	0,0%	0,0%	0,0%	0,2%	50,4%
LeftBag.mpg	157	0	101	0	11	28	52,9%	0,0%	34,0%	0,0%	3,7%	9,4%
LeftBag_AtChair.mpg	699	0	0	0	477	147	52,8%	0,0%	0,0%	0,0%	36,1%	11,1%
LeftBag_BehindChair.mpg	425	0	0	0	143	13	73,1%	0,0%	0,0%	0,0%	24,6%	2,2%
LeftBag_PickedUp.mpg	929	0	0	0	37	15	94,7%	0,0%	0,0%	0,0%	3,8%	1,5%
LeftBox.mpg	607	0	0	0	91	449	52,9%	0,0%	0,0%	0,0%	7,9%	39,1%
Meet_Crowd.mpg	196	0	91	0	67	160	38,1%	0,0%	17,7%	0,0%	13,0%	31,1%
Meet_Split_3rdGuy.mpg	746	0	6	21	108	258	65,5%	0,0%	0,5%	1,8%	9,5%	22,7%
Meet_WalkSplit.mpg	464	0	77	36	525	322	32,6%	0,0%	5,4%	2,5%	36,9%	22,6%
Meet_WalkTogether1.mpg	416	0	12	0	23	19	88,5%	0,0%	2,6%	0,0%	4,9%	4,0%
Meet_WalkTogether2.mpg	187	0	102	0	18	300	30,8%	0,0%	16,8%	0,0%	3,0%	49,4%
Split.mpg	437	0	0	2	45	6	89,2%	0,0%	0,0%	0,4%	9,2%	1,2%
Rest_FallOnFloor.mpg	495	0	0	0	38	11	91,0%	0,0%	0,0%	0,0%	7,0%	2,0%
Rest_InChair.mpg	241	0	0	0	18	3	92,0%	0,0%	0,0%	0,0%	6,9%	1,1%
Rest_SlumpOnFloor.mpg	325	0	0	0	9	358	47,0%	0,0%	0,0%	0,0%	1,3%	51,7%
Rest_WiggleOnFloor.mpg	193	0	0	0	173	27	49,1%	0,0%	0,0%	0,0%	44,0%	6,9%
Fight_Chase.mpg	550	0	22	18	192	72	64,4%	0,0%	2,6%	2,1%	22,5%	8,4%
Fight_OneManDown.mpg	516	0	84	0	313	366	40,3%	0,0%	6,6%	0,0%	24,5%	28,6%
Fight_RunAway1.mpg	251	0	23	0	163	274	35,3%	0,0%	3,2%	0,0%	22,9%	38,5%
Fight_RunAway2.mpg	264	0	75	0	95	632	24,8%	0,0%	7,0%	0,0%	8,9%	59,3%
Walk1.mpg	531	0	0	0	27	16	92,5%	0,0%	0,0%	0,0%	4,7%	2,8%
Walk2.mpg	273	0	0	0	298	5	47,4%	0,0%	0,0%	0,0%	51,7%	0,9%
Walk3.mpg	681	0	0	0	63	51	85,7%	0,0%	0,0%	0,0%	7,9%	6,4%
Total	12885	0	702	23	4056	4330	58,6%	0,0%	3,2%	0,1%	18,4%	19,7%

Tabela B.2: Valores da matriz de confusão e as métricas de exatidão considerando o parâmetro de sobreposição τ_L, quando é testada a técnica de subtração de fundo baseada numa média móvel gaussiana, utilizando a técnica de detecção de variações baseada na distância Euclidiana bivariada.

Vídeos	Matriz de confusão						Exatidão					
	N_{CD}	N_{SU}	N_S	N_U	N_{FD}	N_{FA}	m_{E-CD}	m_{E-SU}	m_{E-S}	m_{E-U}	m_{E-FD}	m_{E-FA}
Browse1.mpg	755	0	0	0	47	28	91,0%	0,0%	0,0%	0,0%	5,7%	3,4%
Browse2.mpg	218	0	0	0	43	4	82,3%	0,0%	0,0%	0,0%	16,2%	1,5%
Browse3.mpg	520	0	0	0	259	2	66,6%	0,0%	0,0%	0,0%	33,2%	0,3%
Browse4.mpg	562	0	16	0	189	77	66,6%	0,0%	1,9%	0,0%	22,4%	9,1%
Browse_WhileWaiting1.mpg	511	0	0	0	697	23	41,5%	0,0%	0,0%	0,0%	56,6%	1,9%
Browse_WhileWaiting2.mpg	562	0	0	0	51	624	45,4%	0,0%	0,0%	0,0%	4,1%	50,4%
LeftBag.mpg	147	0	62	0	60	64	44,1%	0,0%	18,6%	0,0%	18,0%	19,2%
LeftBag_AtChair.mpg	505	0	1	6	706	7	41,2%	0,0%	0,1%	0,5%	57,6%	0,6%
LeftBag_BehindChair.mpg	269	0	0	0	299	11	46,5%	0,0%	0,0%	0,0%	51,6%	1,9%
LeftBag_PickedUp.mpg	926	0	0	0	40	115	85,7%	0,0%	0,0%	0,0%	3,7%	10,6%
LeftBox.mpg	534	0	0	0	164	497	44,7%	0,0%	0,0%	0,0%	13,7%	41,6%
Meet_Crowd.mpg	154	0	49	0	197	305	21,8%	0,0%	7,0%	0,0%	27,9%	43,3%
Meet_Split_3rdGuy.mpg	568	0	4	21	288	54	60,7%	0,0%	0,4%	2,2%	30,8%	5,8%
Meet_WalkSplit.mpg	337	0	84	33	652	319	23,6%	0,0%	5,9%	2,3%	45,8%	22,4%
Meet_WalkTogether1.mpg	315	0	42	0	129	12	63,3%	0,0%	8,4%	0,0%	25,9%	2,4%
Meet_WalkTogether2.mpg	218	0	22	0	67	234	40,3%	0,0%	4,1%	0,0%	12,4%	43,3%
Split.mpg	386	0	0	0	100	38	73,7%	0,0%	0,0%	0,0%	19,1%	7,3%
Rest_FallOnFloor.mpg	402	0	0	0	131	10	74,0%	0,0%	0,0%	0,0%	24,1%	1,8%
Rest_InChair.mpg	196	0	1	0	62	24	69,3%	0,0%	0,4%	0,0%	21,9%	8,5%
Rest_SlumpOnFloor.mpg	328	0	0	0	6	378	46,1%	0,0%	0,0%	0,0%	0,8%	53,1%
Rest_WiggleOnFloor.mpg	103	0	0	0	263	4	27,8%	0,0%	0,0%	0,0%	71,1%	1,1%
Fight_Chase.mpg	498	0	19	30	221	143	54,7%	0,0%	2,1%	3,3%	24,3%	15,7%
Fight_OneManDown.mpg	471	0	50	0	392	350	37,3%	0,0%	4,0%	0,0%	31,0%	27,7%
Fight_RunAway1.mpg	240	0	16	0	184	40	50,0%	0,0%	3,3%	0,0%	38,3%	8,3%
Fight_RunAway2.mpg	219	0	50	0	165	663	20,0%	0,0%	4,6%	0,0%	15,0%	60,4%
Walk1.mpg	404	0	3	0	151	20	69,9%	0,0%	0,5%	0,0%	26,1%	3,5%
Walk2.mpg	176	0	6	0	468	52	25,1%	0,0%	0,9%	0,0%	66,7%	7,4%
Walk3.mpg	485	0	0	0	259	11	64,2%	0,0%	0,0%	0,0%	34,3%	1,5%
Total	10789	0	667	82	6284	3820	49,9%	0,0%	3,1%	0,4%	29,0%	17,7%

Tabela B.3: Valores da matriz de confusão e as métricas de exatidão considerando o parâmetro de sobreposição τ_L, quando é testada a técnica de subtração de fundo baseada numa média móvel gaussiana, utilizando a técnica de detecção de variações baseado no teste de significância.

Vídeos	Matriz de confusão						Exatidão					
	N_{CD}	N_{SU}	N_S	N_U	N_{FD}	N_{FA}	m_{E-CD}	m_{E-SU}	m_{E-S}	m_{E-U}	m_{E-FD}	m_{E-FA}
Browse1.mpg	774	0	0	0	28	35	92,5%	0,0%	0,0%	0,0%	3,3%	4,2%
Browse2.mpg	205	0	0	0	56	6	76,8%	0,0%	0,0%	0,0%	21,0%	2,2%
Browse3.mpg	500	0	0	0	279	1	64,1%	0,0%	0,0%	0,0%	35,8%	0,1%
Browse4.mpg	615	0	29	0	123	83	72,4%	0,0%	3,4%	0,0%	14,5%	9,8%
Browse_WhileWaiting1.mpg	540	0	0	0	668	39	43,3%	0,0%	0,0%	0,0%	53,6%	3,1%
Browse_WhileWaiting2.mpg	597	0	0	0	16	652	47,2%	0,0%	0,0%	0,0%	1,3%	51,5%
LeftBag.mpg	107	0	84	0	78	100	29,0%	0,0%	22,8%	0,0%	21,1%	27,1%
LeftBag_AtChair.mpg	500	0	0	4	712	13	40,7%	0,0%	0,0%	0,3%	57,9%	1,1%
LeftBag_BehindChair.mpg	269	0	0	0	299	9	46,6%	0,0%	0,0%	0,0%	51,8%	1,6%
LeftBag_PickedUp.mpg	941	0	5	0	20	97	88,5%	0,0%	0,5%	0,0%	1,9%	9,1%
LeftBox.mpg	568	0	0	0	130	461	49,0%	0,0%	0,0%	0,0%	11,2%	39,8%
Meet_Crowd.mpg	153	0	77	0	194	211	24,1%	0,0%	12,1%	0,0%	30,6%	33,2%
Meet_Split_3rdGuy.mpg	600	0	5	63	171	460	46,2%	0,0%	0,4%	4,8%	13,2%	35,4%
Meet_WalkSplit.mpg	482	2	65	53	490	310	34,4%	0,1%	4,6%	3,8%	35,0%	22,1%
Meet_WalkTogether1.mpg	342	0	25	0	121	5	69,4%	0,0%	5,1%	0,0%	24,5%	1,0%
Meet_WalkTogether2.mpg	206	0	19	0	82	243	37,5%	0,0%	3,5%	0,0%	14,9%	44,2%
Split.mpg	391	0	0	0	95	47	73,4%	0,0%	0,0%	0,0%	17,8%	8,8%
Rest_FallOnFloor.mpg	399	0	0	0	134	9	73,6%	0,0%	0,0%	0,0%	24,7%	1,7%
Rest_InChair.mpg	175	0	0	0	84	60	54,9%	0,0%	0,0%	0,0%	26,3%	18,8%
Rest_SlumpOnFloor.mpg	329	0	0	0	5	376	46,3%	0,0%	0,0%	0,0%	0,7%	53,0%
Rest_WiggleOnFloor.mpg	100	0	0	0	266	3	27,1%	0,0%	0,0%	0,0%	72,1%	0,8%
Fight_Chase.mpg	441	0	13	17	310	96	50,3%	0,0%	1,5%	1,9%	35,3%	10,9%
Fight_OneManDown.mpg	540	0	8	0	365	509	38,0%	0,0%	0,6%	0,0%	25,7%	35,8%
Fight_RunAway1.mpg	231	0	17	0	189	362	28,9%	0,0%	2,1%	0,0%	23,7%	45,3%
Fight_RunAway2.mpg	240	0	40	0	154	744	20,4%	0,0%	3,4%	0,0%	13,1%	63,2%
Walk1.mpg	506	0	0	0	52	11	88,9%	0,0%	0,0%	0,0%	9,1%	1,9%
Walk2.mpg	170	0	11	0	466	84	23,3%	0,0%	1,5%	0,0%	63,7%	11,5%
Walk3.mpg	487	0	0	0	257	16	64,1%	0,0%	0,0%	0,0%	33,8%	2,1%
Total	11188	2	638	92	5912	4781	49,5%	0,0%	2,8%	0,4%	26,1%	21,1%

Tabela B.4: Valores da matriz de confusão e as métricas de exatidão considerando o parâmetro de sobreposição τ_L, quando é testada a técnica de subtração de fundo baseada no histograma, utilizando a técnica de detecção de variações baseada na distância Euclidiana.

Vídeos	Matriz de confusão						Exatidão					
	N_{CD}	N_{SU}	N_S	N_U	N_{FD}	N_{FA}	m_{E-CD}	m_{E-SU}	m_{E-S}	m_{E-U}	m_{E-FD}	m_{E-FA}
Browse1.mpg	763	0	0	0	39	13	93,6%	0,0%	0,0%	0,0%	4,8%	1,6%
Browse2.mpg	255	0	0	0	6	4	96,2%	0,0%	0,0%	0,0%	2,3%	1,5%
Browse3.mpg	571	0	0	0	208	15	71,9%	0,0%	0,0%	0,0%	26,2%	1,9%
Browse4.mpg	746	0	0	0	21	47	91,6%	0,0%	0,0%	0,0%	2,6%	5,8%
Browse_WhileWaiting1.mpg	621	0	0	0	587	57	49,1%	0,0%	0,0%	0,0%	46,4%	4,5%
Browse_WhileWaiting2.mpg	598	0	0	0	15	215	72,2%	0,0%	0,0%	0,0%	1,8%	26,0%
LeftBag.mpg	164	0	98	0	7	38	53,4%	0,0%	31,9%	0,0%	2,3%	12,4%
LeftBag_AtChair.mpg	728	0	6	0	442	122	56,1%	0,0%	0,5%	0,0%	34,1%	9,4%
LeftBag_BehindChair.mpg	485	0	0	0	83	12	83,6%	0,0%	0,0%	0,0%	14,3%	2,1%
LeftBag_PickedUp.mpg	928	0	0	0	38	17	94,4%	0,0%	0,0%	0,0%	3,9%	1,7%
LeftBox.mpg	606	0	0	0	92	9	85,7%	0,0%	0,0%	0,0%	13,0%	1,3%
Meet_Crowd.mpg	216	0	82	0	47	155	43,2%	0,0%	16,4%	0,0%	9,4%	31,0%
Meet_Split_3rdGuy.mpg	765	0	4	27	77	177	72,9%	0,0%	0,4%	2,6%	7,3%	16,9%
Meet_WalkSplit.mpg	495	0	74	21	528	45	42,6%	0,0%	6,4%	1,8%	45,4%	3,9%
Meet_WalkTogether1.mpg	427	0	12	0	11	12	92,4%	0,0%	2,6%	0,0%	2,4%	2,6%
Meet_WalkTogether2.mpg	193	0	108	0	6	7	61,5%	0,0%	34,4%	0,0%	1,9%	2,2%
Split.mpg	447	0	0	3	33	6	91,4%	0,0%	0,0%	0,6%	6,7%	1,2%
Rest_FallOnFloor.mpg	497	0	0	0	36	13	91,0%	0,0%	0,0%	0,0%	6,6%	2,4%
Rest_InChair.mpg	254	0	0	0	5	6	95,8%	0,0%	0,0%	0,0%	1,9%	2,3%
Rest_SlumpOnFloor.mpg	330	0	0	0	4	61	83,5%	0,0%	0,0%	0,0%	1,0%	15,4%
Rest_WiggleOnFloor.mpg	218	0	0	0	148	7	58,4%	0,0%	0,0%	0,0%	39,7%	1,9%
Fight_Chase.mpg	548	0	6	3	238	22	67,1%	0,0%	0,7%	0,4%	29,1%	2,7%
Fight_OneManDown.mpg	541	0	79	0	293	366	42,3%	0,0%	6,2%	0,0%	22,9%	28,6%
Fight_RunAway1.mpg	245	0	49	0	143	230	36,7%	0,0%	7,3%	0,0%	21,4%	34,5%
Fight_RunAway2.mpg	273	0	76	0	85	261	39,3%	0,0%	10,9%	0,0%	12,2%	37,6%
Walk1.mpg	487	0	0	0	71	3	86,8%	0,0%	0,0%	0,0%	12,7%	0,5%
Walk2.mpg	266	0	0	0	295	6	46,9%	0,0%	0,0%	0,0%	52,0%	1,1%
Walk3.mpg	702	0	0	0	42	48	88,6%	0,0%	0,0%	0,0%	5,3%	6,1%
Total	13192	0	693	21	3743	1968	67,2%	0,0%	3,5%	0,1%	19,1%	10,0%

Tabela B.5: Valores da matriz de confusão e as métricas de exatidão considerando o parâmetro de sobreposição τ_L, quando é testada a técnica de subtração de fundo baseada no histograma, utilizando a técnica de detecção de variações baseada na distância Euclidiana bivariada.

Vídeos	Matriz de confusão						Exatidão					
	N_{CD}	N_{SU}	N_S	N_U	N_{FD}	N_{FA}	m_{E-CD}	m_{E-SU}	m_{E-S}	m_{E-U}	m_{E-FD}	m_{E-FA}
Browse1.mpg	769	0	0	0	33	5	95,3%	0,0%	0,0%	0,0%	4,1%	0,6%
Browse2.mpg	223	0	0	0	38	6	83,5%	0,0%	0,0%	0,0%	14,2%	2,2%
Browse3.mpg	532	0	0	0	247	9	67,5%	0,0%	0,0%	0,0%	31,3%	1,1%
Browse4.mpg	734	0	1	0	32	46	90,3%	0,0%	0,1%	0,0%	3,9%	5,7%
Browse_WhileWaiting1.mpg	557	0	0	0	651	21	45,3%	0,0%	0,0%	0,0%	53,0%	1,7%
Browse_WhileWaiting2.mpg	605	0	0	0	8	197	74,7%	0,0%	0,0%	0,0%	1,0%	24,3%
LeftBag.mpg	155	0	101	0	13	19	53,8%	0,0%	35,1%	0,0%	4,5%	6,6%
LeftBag_AtChair.mpg	593	0	3	2	594	53	47,6%	0,0%	0,2%	0,2%	47,7%	4,3%
LeftBag_BehindChair.mpg	341	0	0	0	227	12	58,8%	0,0%	0,0%	0,0%	39,1%	2,1%
LeftBag_PickedUp.mpg	932	0	2	0	32	47	92,0%	0,0%	0,2%	0,0%	3,2%	4,6%
LeftBox.mpg	589	0	0	0	109	10	83,2%	0,0%	0,0%	0,0%	15,4%	1,4%
Meet_Crowd.mpg	177	0	53	0	150	278	26,9%	0,0%	8,1%	0,0%	22,8%	42,2%
Meet_Split_3rdGuy.mpg	708	0	2	29	132	84	74,1%	0,0%	0,2%	3,0%	13,8%	8,8%
Meet_WalkSplit.mpg	437	0	80	48	527	37	38,7%	0,0%	7,1%	4,3%	46,7%	3,3%
Meet_WalkTogether1.mpg	392	0	20	0	49	9	83,4%	0,0%	4,3%	0,0%	10,4%	1,9%
Meet_WalkTogether2.mpg	154	0	133	0	20	12	48,3%	0,0%	41,7%	0,0%	6,3%	3,8%
Split.mpg	393	0	1	0	92	17	78,1%	0,0%	0,2%	0,0%	18,3%	3,4%
Rest_FallOnFloor.mpg	434	0	0	0	99	10	79,9%	0,0%	0,0%	0,0%	18,2%	1,8%
Rest_InChair.mpg	217	0	1	0	41	4	82,5%	0,0%	0,4%	0,0%	15,6%	1,5%
Rest_SlumpOnFloor.mpg	325	0	0	0	9	46	85,5%	0,0%	0,0%	0,0%	2,4%	12,1%
Rest_WiggleOnFloor.mpg	113	0	0	0	253	4	30,5%	0,0%	0,0%	0,0%	68,4%	1,1%
Fight_Chase.mpg	540	0	5	0	253	52	63,5%	0,0%	0,6%	0,0%	29,8%	6,1%
Fight_OneManDown.mpg	493	0	45	0	375	488	35,2%	0,0%	3,2%	0,0%	26,8%	34,8%
Fight_RunAway1.mpg	242	0	29	0	166	188	38,7%	0,0%	4,6%	0,0%	26,6%	30,1%
Fight_RunAway2.mpg	245	0	78	0	111	337	31,8%	0,0%	10,1%	0,0%	14,4%	43,7%
Walk1.mpg	503	0	0	0	55	10	88,6%	0,0%	0,0%	0,0%	9,7%	1,8%
Walk2.mpg	194	0	1	0	436	52	28,4%	0,0%	0,1%	0,0%	63,8%	7,6%
Walk3.mpg	635	0	0	0	109	15	83,7%	0,0%	0,0%	0,0%	14,4%	2,0%
Total	12080	0	655	26	5017	2085	60,8%	0,0%	3,3%	0,1%	25,3%	10,5%

Tabela B.6: Valores da matriz de confusão e as métricas de exatidão considerando o parâmetro de sobreposição τ_L, quando é testada a técnica de subtração de fundo baseada no histograma, utilizando a técnica de detecção de variações baseado no teste de significância.

Vídeos	Matriz de confusão						Exatidão					
	N_{CD}	N_{SU}	N_S	N_U	N_{FD}	N_{FA}	m_{E-CD}	m_{E-SU}	m_{E-S}	m_{E-U}	m_{E-FD}	m_{E-FA}
Browse1.mpg	766	0	0	0	36	40	91,0%	0,0%	0,0%	0,0%	4,3%	4,8%
Browse2.mpg	205	0	0	0	56	6	76,8%	0,0%	0,0%	0,0%	21,0%	2,2%
Browse3.mpg	501	0	0	0	278	1	64,2%	0,0%	0,0%	0,0%	35,6%	0,1%
Browse4.mpg	654	0	30	0	83	86	76,7%	0,0%	3,5%	0,0%	9,7%	10,1%
Browse_WhileWaiting1.mpg	545	0	0	0	663	42	43,6%	0,0%	0,0%	0,0%	53,0%	3,4%
Browse_WhileWaiting2.mpg	611	0	0	0	2	415	59,4%	0,0%	0,0%	0,0%	0,2%	40,4%
LeftBag.mpg	120	0	82	0	67	98	32,7%	0,0%	22,3%	0,0%	18,3%	26,7%
LeftBag_AtChair.mpg	496	0	0	15	694	25	40,3%	0,0%	0,0%	1,2%	56,4%	2,0%
LeftBag_BehindChair.mpg	272	0	0	0	296	10	47,1%	0,0%	0,0%	0,0%	51,2%	1,7%
LeftBag_PickedUp.mpg	943	0	5	0	18	128	86,2%	0,0%	0,5%	0,0%	1,6%	11,7%
LeftBox.mpg	581	0	0	0	117	15	81,5%	0,0%	0,0%	0,0%	16,4%	2,1%
Meet_Crowd.mpg	155	0	72	0	199	202	24,7%	0,0%	11,5%	0,0%	31,7%	32,2%
Meet_Split_3rdGuy.mpg	597	0	4	51	197	410	47,4%	0,0%	0,3%	4,1%	15,6%	32,6%
Meet_WalkSplit.mpg	476	0	66	38	521	334	33,2%	0,0%	4,6%	2,6%	36,3%	23,3%
Meet_WalkTogether1.mpg	338	0	25	0	125	10	67,9%	0,0%	5,0%	0,0%	25,1%	2,0%
Meet_WalkTogether2.mpg	204	0	19	0	84	240	37,3%	0,0%	3,5%	0,0%	15,4%	43,9%
Split.mpg	389	0	1	0	96	61	71,1%	0,0%	0,2%	0,0%	17,6%	11,2%
Rest_FallOnFloor.mpg	398	0	0	0	135	9	73,4%	0,0%	0,0%	0,0%	24,9%	1,7%
Rest_InChair.mpg	184	0	0	0	75	36	62,4%	0,0%	0,0%	0,0%	25,4%	12,2%
Rest_SlumpOnFloor.mpg	330	0	0	0	4	71	81,5%	0,0%	0,0%	0,0%	1,0%	17,5%
Rest_WiggleOnFloor.mpg	101	0	0	0	265	2	27,4%	0,0%	0,0%	0,0%	72,0%	0,5%
Fight_Chase.mpg	518	0	27	22	211	42	63,2%	0,0%	3,3%	2,7%	25,7%	5,1%
Fight_OneManDown.mpg	515	0	27	0	371	553	35,1%	0,0%	1,8%	0,0%	25,3%	37,7%
Fight_RunAway1.mpg	234	0	15	0	188	344	30,0%	0,0%	1,9%	0,0%	24,1%	44,0%
Fight_RunAway2.mpg	245	0	39	0	150	445	27,9%	0,0%	4,4%	0,0%	17,1%	50,6%
Walk1.mpg	502	0	0	0	56	15	87,6%	0,0%	0,0%	0,0%	9,8%	2,6%
Walk2.mpg	171	0	11	0	466	70	23,8%	0,0%	1,5%	0,0%	64,9%	9,7%
Walk3.mpg	512	0	0	0	232	15	67,5%	0,0%	0,0%	0,0%	30,6%	2,0%
Total	11312	0	618	35	5919	3648	52,5%	0,0%	2,9%	0,2%	27,5%	16,9%

Tabela B.7: Valores da matriz de confusão e as métricas de exatidão considerando o parâmetro de sobreposição τ_L, quando é testada a técnica de subtração de fundo baseada numa Mistura de Gaussianas.

Vídeos	Matriz de confusão						Exatidão					
	N_{CD}	N_{SU}	N_S	N_U	N_{FD}	N_{FA}	m_{E-CD}	m_{E-SU}	m_{E-S}	m_{E-U}	m_{E-FD}	m_{E-FA}
Browse1.mpg	450	0	0	0	352	22	54,6%	0,0%	0,0%	0,0%	42,7%	2,7%
Browse2.mpg	244	0	0	0	17	3	92,4%	0,0%	0,0%	0,0%	6,4%	1,1%
Browse3.mpg	459	0	0	0	320	9	58,2%	0,0%	0,0%	0,0%	40,6%	1,1%
Browse4.mpg	737	0	0	0	30	36	91,8%	0,0%	0,0%	0,0%	3,7%	4,5%
Browse_WhileWaiting1.mpg	610	0	0	0	598	30	49,3%	0,0%	0,0%	0,0%	48,3%	2,4%
Browse_WhileWaiting2.mpg	466	0	0	0	147	117	63,8%	0,0%	0,0%	0,0%	20,1%	16,0%
LeftBag.mpg	168	0	87	0	14	27	56,8%	0,0%	29,4%	0,0%	4,7%	9,1%
LeftBag_AtChair.mpg	702	0	1	0	473	69	56,4%	0,0%	0,1%	0,0%	38,0%	5,5%
LeftBag_BehindChair.mpg	417	0	0	0	151	11	72,0%	0,0%	0,0%	0,0%	26,1%	1,9%
LeftBag_PickedUp.mpg	539	0	0	0	427	18	54,8%	0,0%	0,0%	0,0%	43,4%	1,8%
LeftBox.mpg	431	0	0	0	267	13	60,6%	0,0%	0,0%	0,0%	37,6%	1,8%
Meet_Crowd.mpg	228	0	85	0	48	22	59,5%	0,0%	22,2%	0,0%	12,5%	5,7%
Meet_Split_3rdGuy.mpg	726	0	1	37	98	116	74,2%	0,0%	0,1%	3,8%	10,0%	11,9%
Meet_WalkSplit.mpg	589	0	61	44	405	73	50,3%	0,0%	5,2%	3,8%	34,6%	6,2%
Meet_WalkTogether1.mpg	421	0	10	0	19	17	90,1%	0,0%	2,1%	0,0%	4,1%	3,6%
Meet_WalkTogether2.mpg	198	0	92	0	17	32	58,4%	0,0%	27,1%	0,0%	5,0%	9,4%
Split.mpg	445	0	0	4	33	3	91,8%	0,0%	0,0%	0,8%	6,8%	0,6%
Rest_FallOnFloor.mpg	327	0	0	0	206	11	60,1%	0,0%	0,0%	0,0%	37,9%	2,0%
Rest_InChair.mpg	249	0	0	0	10	13	91,5%	0,0%	0,0%	0,0%	3,7%	4,8%
Rest_SlumpOnFloor.mpg	329	0	0	0	5	94	76,9%	0,0%	0,0%	0,0%	1,2%	22,0%
Rest_WiggleOnFloor.mpg	170	0	0	0	196	3	46,1%	0,0%	0,0%	0,0%	53,1%	0,8%
Fight_Chase.mpg	513	0	2	24	235	9	65,5%	0,0%	0,3%	3,1%	30,0%	1,1%
Fight_OneManDown.mpg	551	0	62	19	263	185	51,0%	0,0%	5,7%	1,8%	24,4%	17,1%
Fight_RunAway1.mpg	242	0	75	0	120	79	46,9%	0,0%	14,5%	0,0%	23,3%	15,3%
Fight_RunAway2.mpg	288	0	69	0	77	214	44,4%	0,0%	10,6%	0,0%	11,9%	33,0%
Walk1.mpg	480	0	0	0	78	15	83,8%	0,0%	0,0%	0,0%	13,6%	2,6%
Walk2.mpg	258	0	0	0	320	7	44,1%	0,0%	0,0%	0,0%	54,7%	1,2%
Walk3.mpg	685	0	0	0	59	20	89,7%	0,0%	0,0%	0,0%	7,7%	2,6%
Total	11907	0	567	53	5122	1306	62,8%	0,0%	3,0%	0,3%	27,0%	6,9%

Apêndice C

Técnica de Rastreamento de Objetos

C.1 Introdução

Em uma forma simples, o rastreamento de objetos pode ser definido como o problema de estimar a trajetória de um objeto no plano da imagem, à medida que este se desloca por uma cena. Em outras palavras, um rastreador atribui rótulos aos objetos rastreados em diferentes quadros de um vídeo, onde um objeto pode ser definido como qualquer coisa que seja de interesse para análise posterior.

Num sistema de vigilância visual os objetos em movimento (como veículos ou pedestres) são definidos pelos componentes conectados presentes na máscara do primeiro plano, gerada pela técnica de subtração de fundo. Devido à presença de ruído alguns componentes conectados não correspondem a objetos em movimento válidos. Porém, utilizando informação a priori na etapa de pós-processamento (como forma ou tamanho dos objetos) é possível eliminar a maioria de falsos objetos.

A maior dificuldade das técnicas de rastreamento de objetos é que, em geral, o número de objetos detectados de um quadro a outro quadro pode variar, mesmo no caso de uma perfeita detecção, já que alguns objetos são obstruídos e outros saem/ingressam na cena. Para lidar com os objetos desaparecendo e reaparecendo é necessário incluir o uso de filtros preditivos e técnicas de associação de dados.

Sendo assim, este Apêndice inicia com uma descrição do filtro de Kalman para o rastreamento das características vinculadas a um único objeto presente num vídeo. Aqui é apresen-

tado o modelo de movimento que governa as características do objeto, e também o modelo das observações vinculadas a tais características. Finalmente são descritos os passos principais na aplicação do filtro de Kalman. Na seção seguinte é descrita a problemática do rastreamento de múltiplos objetos, e a aplicação do filtro de Kalman para sua solução, implicando assim a inclusão das etapas de canalização, associação de dados e manutenção de trajetórias, terminando com a descrição da técnica implementada de rastreamento de múltiplos objetos.

Finalmente, é importante destacar que, a implementação da etapa de rastreamento de objetos baseada em filtros de Kalman, permite armazenar através de um rótulo específico, cada objeto detectado por uma técnica de subtração de fundo, possibilitando assim, fazer um procedimento de avaliação baseado em métricas orientadas a objetos.

C.2 Modelo para o Rastreamento de um Único Objeto

Matematicamente, o filtro de Kalman é um estimador que prevê e corrige os estados de uma grande variedade de processos lineares [54]. Não só é eficiente como também é teoricamente atrativo e preciso. Para o caso de técnicas de rastreamento de objetos, o vetor de estados, denotado por $\mathbf{x}(t)$, representará a dinâmica do comportamento do objeto a seguir no quadro t, sendo ele estimado a partir das observações, denotadas por $\mathbf{z}(t)$. Se é considerado que o objeto a seguir é representado por um ponto, seu centróide calculado em cada quadro t, então a observação $\mathbf{z}(t)$ será definida por

$$\mathbf{z}(t) = [c_x, c_y]^T, \tag{C.1}$$

onde c_x, c_y são a posição vertical e horizontal do centróide do objeto sendo rastreado. Em geral, a representação pontual é adequada para rastreamento de objetos que ocupam pequenas regiões numa imagem [79]. Assim, é possível definir o estado $\mathbf{x}(t)$ vinculado a cada observação $\mathbf{z}(t)$ pela posição, velocidade e aceleração do centróide nos eixos vertical e horizontal, já que estas variáveis capturam a dinâmica dos centróides à medida que sua posição varia com o tempo. Assim tem-se que

$$\mathbf{x}(t) = [c_x, c_y, v_x, v_y, a_x, a_y]^T, \tag{C.2}$$

onde v_x, v_y e a_x, a_y são a velocidade e aceleração para cada eixo, respectivamente. Para a estimação do estado $\mathbf{x}(t)$ a partir das observações $\mathbf{z}(t)$, via filtro de Kalman, um modelo de como os estados variam com o tempo e como eles são vinculados com as observações é essencial. Assim, são definidos:

- **Modelo do processo**:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1) + \mathbf{w}(t-1), \tag{C.3}$$

onde $\mathbf{A}$ representa a matriz de transição do estado $\mathbf{x}(t-1)$ para o estado $\mathbf{x}(t)$. O vetor $\mathbf{w}(t-1)$ é o ruído do processo, definido pela distribuição: $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, e as matrizes $\mathbf{A}$ e $\mathbf{Q}$ são definidas pelas expressões

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & T_\mathrm{a} & 0 & T_\mathrm{a}^2/2 & 0 \\ 0 & 1 & 0 & T_\mathrm{a} & 0 & T_\mathrm{a}^2/2 \\ 0 & 0 & 1 & 0 & T_\mathrm{a} & 0 \\ 0 & 0 & 0 & 1 & 0 & T_\mathrm{a} \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{Q} = \sigma_w^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

onde T_a é a taxa de amostragem do modelo cinemático discretizado de um ponto deslocando-se por um plano (para mais detalhes, ver [60]).

- **Modelo de observação**:

$$\mathbf{z}(t) = \mathbf{H}\mathbf{x}(t) + \mathbf{v}(t), \tag{C.4}$$

onde a matriz $\mathbf{H}$ representa a matriz de observações. O vetor $\mathbf{v}(t)$ é a medida de ruído definida pela distribuição $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ e as matrizes $\mathbf{H}$ e $\mathbf{R}$ são definidas pelas expressões

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{R} = \sigma_v^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Considerando que $\hat{\mathbf{x}}(t|t')$ é a estimação do estado $\mathbf{x}(t)$ baseada nas observações disponíveis até o instante t', então surgem dois casos de interesse:

- $\hat{\mathbf{x}}(t|t-1)$ é a estimação do estado $\mathbf{x}(t)$ no instante t baseada nas observações passadas, ou seja, é uma estimação a priori;
- $\hat{\mathbf{x}}(t|t)$ é a estimação do estado $\mathbf{x}(t)$ no instante t baseada na observação atual e nas observações passadas, ou seja, é uma estimação a posteriori.

Os erros de estimação correspondentes a estes dois casos são

$$\begin{aligned} \mathbf{e}(t|t) &= \mathbf{x}(t) - \hat{\mathbf{x}}(t|t) \\ \mathbf{e}(t|t-1) &= \mathbf{x}(t) - \hat{\mathbf{x}}(t|t-1), \end{aligned}$$

e $\mathbf{P}(t|t)$ e $\mathbf{P}(t|t-1)$ são as matrizes de covariância do erro, dadas por

$$\begin{aligned} \mathbf{P}(t|t) &= \mathbb{E}[\mathbf{e}(t|t)\mathbf{e}^T(t|t)] \\ \mathbf{P}(t|t-1) &= \mathbb{E}[\mathbf{e}(t|t-1)\mathbf{e}^T(t|t-1)]. \end{aligned}$$

O filtro de Kalman estima os estados $\hat{\mathbf{x}}(t|t-1)$ e $\hat{\mathbf{x}}(t|t)$ baseado em dois passos principais, a saber,

- **predição**: prevê o próximo estado do sistema e sua incerteza, a saber,

$$\hat{\mathbf{x}}(t|t-1) = \mathbf{A}\hat{\mathbf{x}}(t-1|t-1) \tag{C.5}$$
$$\hat{\mathbf{P}}(t|t-1) = \mathbf{A}\hat{\mathbf{P}}(t-1|t-1)\mathbf{A}^T + \mathbf{Q}. \tag{C.6}$$

  De forma equivalente, também é calculada a observação predita e sua incerteza, dadas por

$$\hat{\mathbf{z}}(t|t-1) = \mathbf{H}\hat{\mathbf{x}}(t|t-1) \tag{C.7}$$
$$\hat{\mathbf{O}}(t|t-1) = \mathbf{H}\hat{\mathbf{P}}(t|t-1)\mathbf{H}^T + \mathbf{Q}. \tag{C.8}$$

- **correção**: corrige, com base na observação do próximo estado, os valores

$$\mathbf{K}(t) = \hat{\mathbf{P}}(t|t-1)\mathbf{H}^T \left(\mathbf{H}\hat{\mathbf{P}}(t|t-1)\mathbf{H}^T + \mathbf{R}\right)^{-1} \quad \text{(C.9)}$$

$$\hat{\mathbf{x}}(t|t) = \hat{\mathbf{x}}(t|t-1) + \mathbf{K}(t)\left(\mathbf{z}(t) - \mathbf{H}\hat{\mathbf{x}}(t|t-1)\right) \quad \text{(C.10)}$$

$$\hat{\mathbf{P}}(t|t) = \left(\mathbf{I} - \mathbf{K}(t)\mathbf{H}\right)\hat{\mathbf{P}}(t|t-1). \quad \text{(C.11)}$$

  onde $\mathbf{K}(t)$ é o ganho de Kalman.

Os vetores $\hat{\mathbf{x}}(0|0)$ e $\hat{\mathbf{P}}(0|0)$ são iniciados com valores aleatórios.

Descrição da Técnica Implementada

A descrição do seguimento de um único objeto através do filtro de Kalman é apresentada no algoritmo 2, onde

- a função `CalculateBackSubTec` calcula a máscara do primeiro plano para cada quadro;
- a função `CalculateEstObs` calcula a observação atual, ou seja o centróide do objeto a seguir;
- a função `Prediction` executa o passo de predição do filtro de Kalman através das Equações (C.5)-(C.6);
- a função `Correction` executa o passo de correção do filtro de Kalman através das Equações (C.9)-(C.11).

C.3 Modelo para o Rastreamento de Múltiplos Objetos

Enquanto um filtro de Kalman é usado para seguir um único objeto, uma técnica de rastreamento de múltiplos objetos baseada no filtro de Kalman utiliza um conjunto de filtros para

Algoritmo 2: Rastreamento de um único objeto.

Input: O conjunto de quadros de um vídeo em questão $\{\mathbf{I}_t\}_{t=1}^{N_{\text{quadros}}}$.

Input: O conjunto do parâmetros do algoritmo de subtração de fundo, backparam.

Input: O conjunto do parâmetros do filtro de Kalman, kalparam $= \{\mathbf{A}, \mathbf{H}, \mathbf{Q}, \mathbf{R}\}$.

Output: O conjunto de estados seguintes, $\mathcal{X}$.

```
// inicialização
```

$\hat{\mathbf{x}}(0|0) \leftarrow \mathbf{x}_0$;$\hat{\mathbf{P}}(0|0) \leftarrow \mathbf{P}_0$;

for $t \leftarrow 1$ **to** $N_{quadros}$ **do** `/* para cada quadro do vídeo */`

```
    // cálculo da máscara do primeiro plano
```

$\mathbf{M}(t) \leftarrow$ `CalculateBackSubTec(`$\boldsymbol{I}_t$`,backparam);`

```
    // determinar a observação atual
```

$\mathbf{z}(t) \leftarrow$ `CalculateEstObs(`$\boldsymbol{M}(t)$`);`

```
    // prever o estado seguinte e observação seguinte
```

$\{\hat{\mathbf{x}}(t|t-1), \hat{\mathbf{P}}(t|t-1), \hat{\mathbf{z}}(t|t-1), \hat{\mathbf{O}}(t|t-1)\} \leftarrow$
`Prediction(`$\boldsymbol{x}(t-1|t-1), \boldsymbol{P}(t-1|t-1)$`,kalparam);`

```
    // corrigir o estado seguinte
```

$\{\hat{\mathbf{x}}(t|t), \hat{\mathbf{P}}(t|t), \mathbf{K}(t)\} \leftarrow$ `Correction(`$\boldsymbol{z}(t), \hat{\boldsymbol{x}}(t|t-1), \hat{\boldsymbol{P}}(t|t-1)$`,kalparam);`

```
    // armazenando o estado
```

$\mathcal{X} \leftarrow \mathcal{X} \cup \{\hat{\mathbf{x}}(t|t)\}$;

determinar a trajetória de um número não conhecido de objetos em movimento, onde a trajetória dos objetos num vídeo é modelada como uma sequência de estados. Assim, se $\mathbf{x}_i(t)$ é o estado do i-ésimo objeto no quadro t, então a sequência de estados $\mathbf{x}_i^{1:t} = \{\mathbf{x}_i(1), \cdots, \mathbf{x}_i(t)\}$ define a trajetória do i-ésimo objeto até o quadro t e $\mathcal{X}^t = \{\mathbf{x}_1(t), \cdots, \mathbf{x}_{n_T}(t)\}$ serão os estados vinculados a cada objeto presente no quadro t tal que $\mathcal{X}^{1:t} = \{\mathbf{x}_1^{1:t}, \cdots, \mathbf{x}_{n_T}^{1:t}\}$ são todas as trajetórias dos objetos até o quadro t.

De forma equivalente, a todo momento, observações são determinadas em cada quadro do vídeo. Assim, se $\mathbf{z}_i(t)$ é a observação vinculada ao i-ésimo objeto no quadro t, a sequência de observações $\{\mathbf{z}_i(1), \cdots, \mathbf{z}_i(t)\}$ serão todas as observações até o quadro t vinculadas ao i-ésimo objeto e $\mathcal{Z}^t = \{\mathbf{z}_1(t), \cdots, \mathbf{z}_{n_M}(t)\}$ serão todas as observações medidas no quadro t.

Na prática, a associação das observações $\mathcal{Z}^t$ com suas correspondentes trajetórias $\mathcal{X}^{1:t}$ é difícil por um número de razões. Primeiro, as observações se podem perder (ao serem obstruídas ou não detectadas); segundo, todas as falsas observações candidatas (por exemplo, manchas presentes na máscara do primeiro plano geradas pelas mudanças da iluminação) necessitam ser corretamente classificadas como falsas observações, e, por conseguinte, não serem associadas com alguma trajetória; finalmente, algumas trajetórias iniciam ou finalizam em algum quadro da sequência de vídeo.

A Figura C.1, tomada de [6], apresenta os elementos básicos de uma típica técnica de rastreamento de múltiplos objetos. Assume-se que para o quadro t tem-se as trajetórias já instanciadas de dados prévios $\mathcal{X}^{1:t}$, e que um novo conjunto de observações $\mathcal{Z}^t$ está disponível. Então, as observações são consideradas para serem parte de uma trajetória existente ou para inicializarem uma nova trajetória[1]. Assim, pode-se descrever o problema como: o processo de obtenção das observações $\mathcal{Z}^t$, associação das observações $\mathcal{Z}^t$ a suas correspondentes trajetórias $\mathcal{X}^{1:t}$, seguidas pela estimação dos estados futuros das trajetórias, onde o passo de manutenção das trajetórias é adicionado, depois da associação de dados, para controlar a inicialização, propagação e finalização delas.

Para o caso de técnica de rastreamento de múltiplos objetos, as etapas de canalização, associação de dados e manutenção de trajetórias são de grande importância, sendo cada uma delas explicadas detalhadamente na próxima seção.

[1]Ou seja, se uma observação está vinculada a um novo objeto, uma nova trajetória tem que ser criada. Esta é uma técnica de hipótese única, o que significa que toda observação é associada a somente uma trajetória existente, implicando que se for realizada uma associação errada (uma observação é associada a uma trajetória errada) a técnica será incapaz de corrigir este erro.

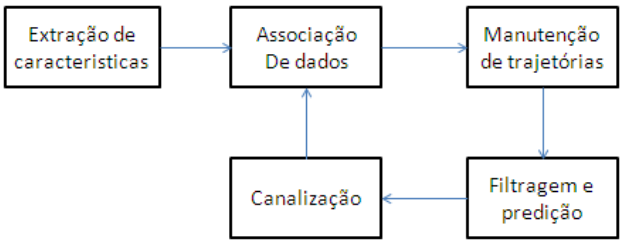


Figura C.1: Diagrama de fluxo de uma técnica de rastreamento de múltiplos objetos genérica [6].

C.3.1 Canalização

A canalização é um procedimento simples para a eliminação dos pares observação-trajetória menos prováveis. Uma vizinhança é definida ao redor de cada observação predita $\hat{\mathbf{z}}_i(t|t-1)$, e todas as observações que estão dentro da vizinhança são consideradas como candidatas para atualizar a i-ésima trajetória. A maneira pela qual as observações são selecionadas para atualizar uma trajetória específica depende do método de associação de dados; porém a maioria das técnicas de associação de dados utiliza a canalização para reduzir o número de combinações possíveis entre observações e trajetórias. A implementação da canalização é efetuada através da construção de uma matriz de custos retangular $\mathbf{R} = [r_{ij}]_{n_T \times n_M}$ que quantifica a distância de cada observação predita $\hat{\mathbf{z}}$ em relação às observações presentes $\mathbf{z}$ no quadro t, ou seja,

$$r_{ij} = \begin{cases} ||\hat{\mathbf{z}}_i(t|t-1) - \mathbf{z}_j(t)||, & \text{se } ||\hat{\mathbf{z}}_i(t|t-1) - \mathbf{z}_j(t)|| < c_{\max} \\ c_{\max}, & \text{caso contrário} \end{cases}, \tag{C.12}$$

onde o parâmetro $c_{\max}$ é o máximo custo aceitável.

C.3.2 Associação de Dados

O problema de associação de dados é visto como a associação entre as n_M observações e as n_T trajetórias, para cada quadro do vídeo. Uma enumeração exaustiva de todas as possíveis associações resulta em $2^{n_T \times n_M}$ associações, o que pode se tornar inviável mesmo para valores apenas moderadamente grandes de n_T e n_M.

Para rastreamento de objetos num vídeo, o problema da geração das hipóteses de associação, uma a uma, pode ser representado como um problema de acoplamentos mínimos num grafo

bipartido ponderado, também chamado problema de designação linear, sendo este um dos mais famosos problemas da programação linear e otimização combinatória.

Informalmente falando, seja uma matriz de custos $\mathbf{C} = [c_{ij}]_{n\times n}$, onde se deseja casar cada linha com uma única e particular coluna, de tal maneira que a soma dos correspondentes custos selecionados é minimizada. Em outras palavras, se deseja selecionar n elementos de $\mathbf{C}$ tal que tem-se exatamente um elemento em cada fila e coluna, e a soma dos custos correspondentes é mínima. Do ponto de vista da teoria de grafos, seja o grafo bipartido $\mathcal{G} = (\mathcal{U}, \mathcal{V}; \mathcal{E})$ tendo um vértice em $\mathcal{U}$ para cada linha e um vértice em $\mathcal{V}$ para cada coluna, e um custo c_{ij} associado com a aresta $[i, j](i, j = 1, \cdots, n)$. O problema é, então, determinar o custo de um casamento perfeito em $\mathcal{G}$, ou seja, achar um sub-conjunto de arestas tal que cada vértice pertença exatamente a uma aresta e a soma dos custos destas arestas é mínima. Pela introdução de uma matriz binária $\mathbf{U} = [u_{ij}]_{n\times n}$, tal que $u_{ij} = 1$ se a linha i é atribuída à coluna j, o problema de designação linear pode ser modelado como

$$\begin{aligned} \min \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}u_{ij} \\ \text{sujeito a} \quad & \sum_{j=1}^{n} u_{ij} = 1 \; i = 1, \cdots, n \\ & \sum_{i=1}^{n} u_{ij} = 1 \; j = 1, \cdots, n \\ & u_{ij} \in 0, 1 \end{aligned} \tag{C.13}$$

Existem algoritmos que resolvem o problema de designação linear, tal como o algoritmo de Jonker e Volgenant [35], que computa a casamento entre dois conjuntos do mesmo tamanho. Entretanto, um problema surge quando o número de observações disponíveis não é igual ao número de trajetórias, que é o caso mais comum em rastreamento de múltiplos objetos. Adicionalmente, posto que o casamento deve ser necessariamente um a um, uma trajetória poderia ocasionalmente ser forçada a aceitar uma não provável observação. Para solucionar este problema, duas transformações aplicadas à matriz de custo foram propostas em [73]:

- primeiro, para tratar as observações espúrias, ou a ausência delas, a matriz de custos retangular $\mathbf{R} \in \mathbb{R}^{n_T\times n_M}$ pode ser aumentada para uma matriz quadrada $\mathbf{C} \in \mathbb{R}^{n\times n}$ preenchendo-a com o valor $c_{\max}$, onde $n = \max\{n_T, n_M\}$. A matriz de permutação $\mathbf{U}$ que resolve o problema de designação linear em $\mathbf{C}$ diretamente induz um casamento de mínimo custo na matriz retangular $\mathbf{R}$, para qualquer valor da constante $c_{\max}$;

- segundo, para relaxar a restrição de um casamento um a um a matriz de custo $\mathbf{C}$ pode ser estendida para ser uma matriz $\mathbf{F} \in \mathbb{R}^{2n \times 2n}$, pela adição de 3 matrizes de tamanho $n \times n$ contendo o valor $c_{\max}$. Assim, a matriz de permutação $\mathbf{V}$ que resolve o problema de designação linear em $\mathbf{F}$ induz custo de casamento de pares que não excede o valor de $c_{\max}$. O parâmetro $c_{\max}$ é, portanto, o máximo custo de casamento aceitável. Por conseguinte, permite relaxar a restrição de casamento no problema original $\mathbf{R}$.

Pelas combinações destas duas operações, a solução do problema de associação de dados é resolvida com um algoritmo padrão para o problema de designação linear, enquanto lida adequadamente com a oclusão temporária de objetos e ruído adicional.

C.3.3 Manutenção de Trajetórias

Esta etapa permite definir uma nova trajetória para aquelas observações que não foram associadas na iteração anterior, em relação à iteração atual. Assim, seja $\mathcal{Z}^{\text{ant}}$ e $\mathcal{Z}^{\text{act}}$ o conjunto de observações nos quadros $t-1$ e t que não foram associadas a alguma trajetória, tal que a cardinalidade destes conjuntos é denotada por $n_{M_{\text{ant}}}$ e $n_{M_{\text{act}}}$, respectivamente. Os passos para o estabelecimento de uma nova trajetória, considerando os conjuntos mencionados, são:

- é determinado o conjunto de observações no instante t que não foram associadas a alguma trajetória

$$\mathcal{Z}^{\text{act}} = \{\mathbf{z}_j(t) \in \mathcal{Z}^t | \sum_{i=1}^{n_T} v_{ij} = 0\}; \tag{C.14}$$

- é calculada a matriz de custos $\mathbf{R}' = [r'_{ij}]_{n_{M_{\text{ant}}} \times n_{M_{\text{act}}}}$, a saber,

$$r'_{ij} = \begin{cases} ||\mathbf{z}_i^{\text{ant}} - \mathbf{z}_j^{\text{act}}||, & \text{se } ||\mathbf{z}_i^{\text{ant}} - \mathbf{z}_j^{\text{act}}|| < c_{\max} \\ c_{\max}, & \text{caso contrário} \end{cases}, \tag{C.15}$$

  onde $\mathbf{z}_i^{\text{ant}} \in \mathcal{Z}^{\text{ant}}$ e $\mathbf{z}_j^{\text{act}} \in \mathcal{Z}^{\text{act}}$;

- utilizando o mesmo procedimento que para o passo de associação de dados, é calculada a matriz de permutação $\mathbf{V}'$ vinculada à matriz de custos $\mathbf{R}'$, e a partir desta são determinadas as observações que geram uma nova trajetória, sendo elas pertencentes ao conjunto de observações $\mathcal{Z}^{\text{ant}}$ que tem um casamento com alguma observação do conjunto $\mathcal{Z}^{\text{act}}$, a saber,

$$\mathcal{Z}^{\text{novo}} = \{\mathbf{z}_i^{\text{ant}} \in \mathcal{Z}^{\text{ant}} | \sum_{j=1}^{n_{M_{\text{act}}}} v'_{ij} = 1\}; \tag{C.16}$$

- é determinado o conjunto de observações que ainda não estão associadas a alguma trajetória, ou seja,

$$\mathcal{Z}^{\text{ant}} = \{\mathbf{z}_j^{\text{act}} \in \mathcal{Z}^{\text{act}} | \sum_{i=1}^{n_{M_{\text{ant}}}} v'_{ij} = 0\}.$$

Assim, o conjunto de observações $\mathcal{Z}^{\text{novo}}$ permite inicializar um novo conjunto de trajetórias que serão adicionadas às trajetórias já existentes.

C.3.4 Descrição da Técnica Implementada

A descrição da técnica de rastreamento de múltiplos objetos através do filtro de Kalman é apresentada no Algoritmo 3, onde:

- A função `CalculateBackSubTec` calcula a máscara do primeiro plano para cada quadro.
- A função `CalculateEstObs` calcula o conjunto de observações atuais, ou seja os centróides de cada objeto presente na máscara do primeiro plano.
- A função `Prediction` executa o passo de predição do filtro de Kalman para todo o conjunto de estados do quadro atual.
- A função `Gating` calcula a matriz de custos $\mathbf{R}$ através da Equação (C.12).
- A função `DataAssociation` calcula a matriz de permutação $\mathbf{V}$ associada à $\mathbf{R}$ através do algoritmo de Jonker e Volgenant, o qual soluciona o problema de designação linear.
- A função `trackMaintenance` trata da inicialização de novas trajetórias considerando os valores diferentes de zero da matriz de permutação $\mathbf{V}$.
- A função `Correction` executa o passo de correção do filtro de Kalman para todo o conjunto de estados do quadro atual.

Algoritmo 3: Rastreamento de múltiplos objetos.

Input: O conjunto de quadros de um vídeo em questão $\{\mathbf{I}_t\}_{t=1}^{N_{\text{quadros}}}$.

Input: O conjunto do parâmetros do algoritmo de subtração de fundo, backparam.

Input: O conjunto do parâmetros do filtro de Kalman, kalparam $= \{\mathbf{A}, \mathbf{H}, \mathbf{Q}, \mathbf{R}\}$.

Output: O conjunto de estados seguintes,$\mathcal{X}$.

```
// inicialização
```

$\mathcal{X}^{0|0} \leftarrow \mathcal{X}_0$;

for $t \leftarrow 1$ **to** $N_{quadros}$ **do** `/* para cada quadro do vídeo */`

```
// cálculo da máscara do primeiro plano
```

$\mathbf{M}(t) \leftarrow$ `CalculateBackSubTec(`$\boldsymbol{I}_t$`,backparam);`

```
// determina o conjunto de observações atuais
```

$\mathcal{Z}^t \leftarrow$ `CalculateEstObs(`$\boldsymbol{M}(t)$`);`

if $n_T > 0$ **then**

```
// prevê o conjunto de estados e observações seguintes
```

$[\hat{\mathcal{X}}^{t|t-1}, \hat{\mathcal{Z}}^{t|t-1}, n_T] \leftarrow$ `Prediction(`$\mathcal{X}^{t-1|t-1}$`, kalparam);`

$\mathbf{V} \leftarrow \mathbf{0}_{n_T \times n_M}$;

if $n_M > 0$ **then**

```
// Canalização
```

$\mathbf{R} \leftarrow$ `Gating(`$\hat{\mathcal{Z}}^{t|t-1}, \mathcal{Z}^t, c_{max}$`);`

```
// Associação de dados
```

$\mathbf{V} \leftarrow$ `DataAssociation(` $\boldsymbol{R}, c_{max}$`);`

```
// Manutenção de trajetórias
```

$[\mathbf{V}, \hat{\mathcal{X}}^{t|t-1}, n_T] \leftarrow$ `trackMaintenance(` $\boldsymbol{V}, \hat{\mathcal{X}}^{t|t-1}, \mathcal{Z}^t$`);`

```
// corrigir o estado seguinte
```

$[\hat{\mathcal{X}}^{t|t}, n_T] \leftarrow$ `Correction(`$\hat{\mathcal{X}}^{t|t-1}, \mathcal{Z}^t$, $\boldsymbol{V}$`);`

```
// armazenando o conjunto de estados
```

$\mathcal{X}^{1:t} \leftarrow \mathcal{X}^{1:t-1} \cup \hat{\mathcal{X}}^{t|t}$;

Apêndice D

Modelo de cor HSB e a Diferença de Imagens

D.1 Introdução

Cor é a principal propriedade da luz visível pela qual um observador humano pode distinguir diferentes tipos de luz. Ela é uma propriedade subjetiva, e em geral, a cor de uma fonte de luz não é estabelecida pela medida de alguma propriedade física fundamental da luz. Foi verificado experimentalmente que para descrever uma determinada cor da luz é necessário estabelecer 3 valores[1]. Assim, a cor é uma propriedade tridimensional no sentido matemático. No caso de uma outra propriedade tridimensional, a localização de um ponto no espaço, pode ser descrito através de diversas definições de três parâmetros (coordenadas)[2]. Similarmente, no caso da cor, diferentes sistemas de três parâmetros podem ser utilizadas. Um sistema particular é chamado de modelo de cor, e seus parâmetros são referidos como sendo as coordenadas de seu espaço de cor tridimensional.

Um modelo de cor que está bem relacionado com a percepção humana da cor usa as seguintes propriedades como suas coordenadas: brilho, matiz e saturação. As propriedades de matiz e saturação descrevem em forma conjunta a cromaticidade da cor de interesse e uma vez que a cromaticidade contém duas das coordenadas da cor, ela é uma propriedade bidimensional.

[1]Baseado na visão tricromática, que é a capacidade de possuir três canais para transmitir informação de cor. Assim, a retina contém três tipos de receptores para cor (nos vertebrados, são as células cones) com diferentes espectros de absorção.

[2]Por exemplo: coordenadas cartesianas, coordenadas cilíndricas, coordenadas esféricas, coordenadas geodésicas.

Também existem outros pares de propriedades (diferentes de matiz e saturação) que podem ser usados para descrever a cromaticidade, permitindo assim diferentes modelos de cores dentro da família de brilho-cromaticidade.

Este Apêndice tem por finalidade determinar aquelas características das cores que se relacionam diretamente com as componentes da distância Euclidiana, para o qual é necessário expressar os píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ através de um modelo de cor pertencente à família de brilho-cromaticidade. Como nessa Tese é usado o modelo de cor HSB, o Apêndice inicia com uma descrição dos modelos de cores RGB e HSB, seguido por uma descrição da conversão de RGB a HSB e vice-versa. Na última seção é analisada a diferença de cores no modelo de cor HSB.

D.2 Modelo de Cores RGB

O modelo de cores RGB é um modelo aditivo, no qual o vermelho, o verde e o azul (usados em modelos de cor aditivos) são combinados de várias maneiras para reproduzir outras cores. O nome do modelo e a abreviação RGB vêm das três cores primárias: vermelho (*Red*), verde (*Green*) e azul (*Blue*). A quantidade de cada cor primária contida numa cor em *hardware* é representada por um número no intervalo de $[0, 255]$. Sendo assim, cada cor é definida como um vetor de três elementos, formado pelos valores assumidos por cada cor primária. Portanto, cada cor existe num cubo de tamanho $[0, 255]^3$, onde cada uma das cores primárias constitui um dos eixos principais do cubo. O cubo de cor RGB é ilustrado na Figura D.1. A origem do cubo, $(0, 0, 0)$, representa a total ausência de cor, que corresponde à cor preta. O canto mais distante a partir da origem é a soma das maiores intensidades de vermelho, verde e azul, ou seja $(255, 255, 255)$. Isso produz a cor branca. Os outros cantos do cubo representam as várias cores primárias (R, G e B) e secundárias[3] (C, M e Y). A diagonal principal do cubo RGB é ilustrada na Figura D.1 como uma linha desenhada entre as cores preto e branco. Esta linha, denominada eixo neutro, representa as cores, no cubo RGB, que contem quantidades iguais de vermelho, verde e azul. Todos os pontos nesta linha são cinza[4]. Quanto mais próximo a cor está da origem, mais escura ela é, e quanto mais perto a cor está do branco, mais clara ela é.

[3] As três cores secundárias são: ciano (*Cyan*), magenta (*Magenta*) e amarelo (*Yellow*).

[4] Se uma coordenada (R, G, B) é projetada sobre o eixo neutro (segundo uma media ponderada que toma em conta a percepção humana), a coordenada é convertida em tons de cinza.

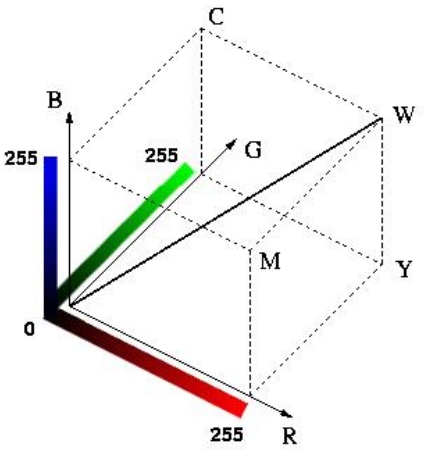


Figura D.1: O cubo de cor RGB. Cada eixo do cubo representa os valores de vermelho, verde ou azul no intervalo $[0, 255]$. Onde R é vermelho, G é verde, B é azul, C é Ciano, M é magenta, Y é Amarelo e W é branco.

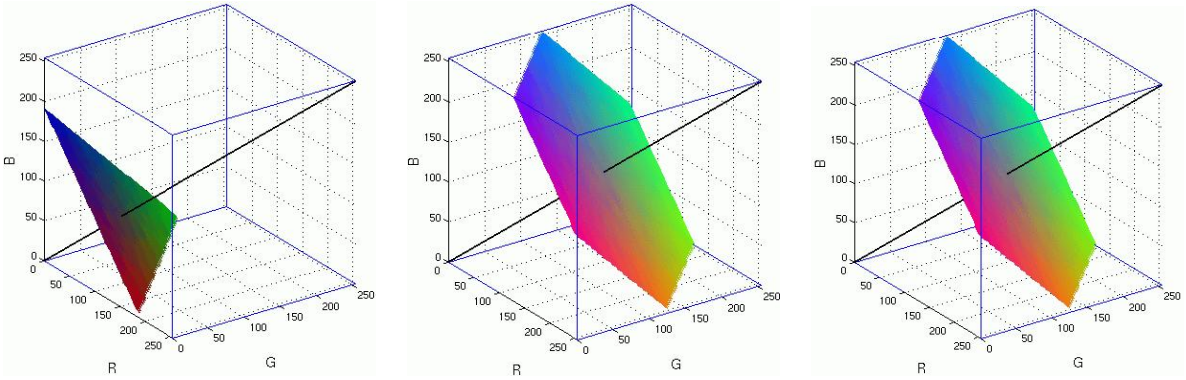
D.3 Modelo de Cores HSB

Este modelo de cor pertence à família de brilho-cromaticidade, e, portanto, usa as seguintes propriedades como suas coordenadas: (a) brilho, que é uma descrição relativa de quanta luz há da cor. Se o brilho é alto, diz-se que a cor é brilhante; (b) matiz, que se refere à tonalidade específica da cor, e, assim é a propriedade que distingue vermelho do laranja, do azul, e assim por diante; (c) saturação, que se refere à *pureza* da cor, e, assim é a propriedade que distingue vermelho de rosa. O nome do modelo e a abreviação HSB vêm das três propriedades já mencionadas: *Hue* (matiz), *Saturation* (saturação) e *Brigthness* (brilho). Para um melhor entendimento do modelo de cor HSB, é de ajuda entender suas relações no cubo RGB. Assim, é útil identificar as partes do cubo RGB que correspondem a valores constantes de matiz, saturação e brilho.

D.3.1 Brilho

É um atributo da percepção visual, na qual uma área parece emitir mais ou menos luz, dando uma descrição relativa de quanta luz está vindo da cor. O conceito de brilho independe da cor da luz, Assim, cores de igual brilho no cubo RGB são aquelas tal que três componentes de cor somam o mesmo valor. Portanto, um determinado nível de brilho é representado no cubo RGB como um plano perpendicular ao eixo neutro ($R + G + B =$ constante). Assim, o brilho é definido como a média aritmética das coordenadas de R , G e B.

$$I = \frac{R+G+B}{3},$$



(a) (b) (c)

Figura D.2: Planos de brilho constante no cubo RGB. (a) brilho a 25%; brilho a 50%; brilho a 75%.

Na Figura D.2 são apresentados três exemplos de planos perpendiculares ao eixo neutro no cubo RGB. O cubo da Figura D.2.a tem um plano obscuro, o cubo na Figura D.2.b contém um plano de brilho intermediário e o cubo na Figura D.2.c apresenta o plano mais claro. Note-se como as cores de cada plano parecem igualmente brilhantes. Os três cubos da Figura D.2 mostram planos de cores a 25%, 50% e 75% de brilho, onde branco puro e preto puro têm um brilho de 0% e 100%, respectivamente, e são pontos individuais.

Já que o brilho é uma propriedade física da radiação, e não corresponde diretamente à percepção humana da cor, conta-se com representações do brilho que compensam este fato, como são luminosidade, valor, e luminância, definidos formalmente como

$$\begin{aligned} \text{luminosidade} \quad & L = \frac{\max(R,G,B) + \min(R,G,B)}{2}, \\ \text{valor} \quad & V = \max(R,G,B), \\ \text{luminância} \quad & Y = 0,30R + 0,59G + 0,11B, \end{aligned}$$

onde luminância é a representação do brilho que melhor corresponde à percepção humana, já que os valores $0,30$, $0,59$ e $0,11$ são as ponderações aproximadas que correspondem à sensibilidade do olho humano para as cores vermelho, verde e azul. A Figura D.3 apresenta as diferentes definições para o brilho projetadas sobre o eixo neutro. Também é indicada a posição no cubo RGB de uma cor representativa, para o qual $(R,G,B) = (220,60,120)$, e as localizações no eixo neutro correspondentes ao valor, V, à luminosidade, L, e à luminância, Y. Como pode ser visto na Figura D.3, o valor é o mais brilhante. A luminância

e luminosidade são variáveis, podendo qualquer uma delas ser mais escura, dependendo da escolha da coordenada RGB.

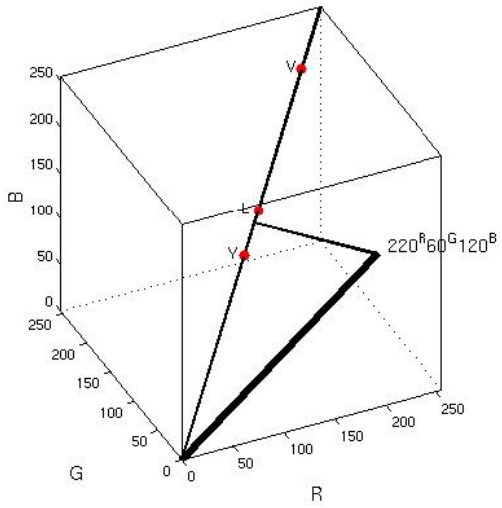
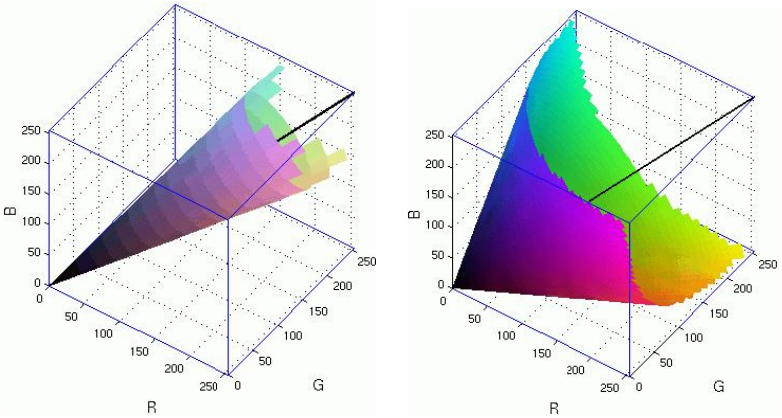


Figura D.3: Projeções sobre o eixo neutro no cubo RGB do ponto $(R, G, B) = (220, 60, 120)$.

D.3.2 Saturação

A saturação é a policromia relativa de uma área em relação ao seu brilho. A policromia pode ser descrita no contexto do cubo RGB e o eixo neutro. Como já foi mencionado, o eixo neutro é a diagonal do cubo que vai de $(R, G, B) = (0, 0, 0)$ até $(R, G, B) = (255, 255, 255)$, e que consiste do preto, branco, e os tons de cinza entre eles. Assim, este eixo não tem informação da cor (isto é, matiz). A policromia de qualquer ponto do cubo RGB, então, é proporcional à distância perpendicular de tal ponto em relação ao eixo neutro. Consequentemente, pontos mais próximos do eixo são menos coloridos (isto é, próximos ao cinza) e aqueles que estão mais longe são mais coloridos. Então, a saturação de um ponto no cubo RGB é a razão entre sua policromia e seu brilho. Isto significa que as superfícies de saturação constante do cubo RGB são cones centradas no eixo neutro.

A Figura D.4 apresenta duas instâncias do cubo RGB. O cubo da Figura D.4.a mostra o cone correspondente a 20% de saturação e da Figura D.4.b corresponde a 70% de saturação. As cores do cone mais à direita parecem mais vivas, pois são mais saturadas. As cores do cone mais à esquerda são muito mais pálidas, porque as suas cores são mais próximas ao eixo neutro. Quando uma cor é mais neutra, é chamada pastel.

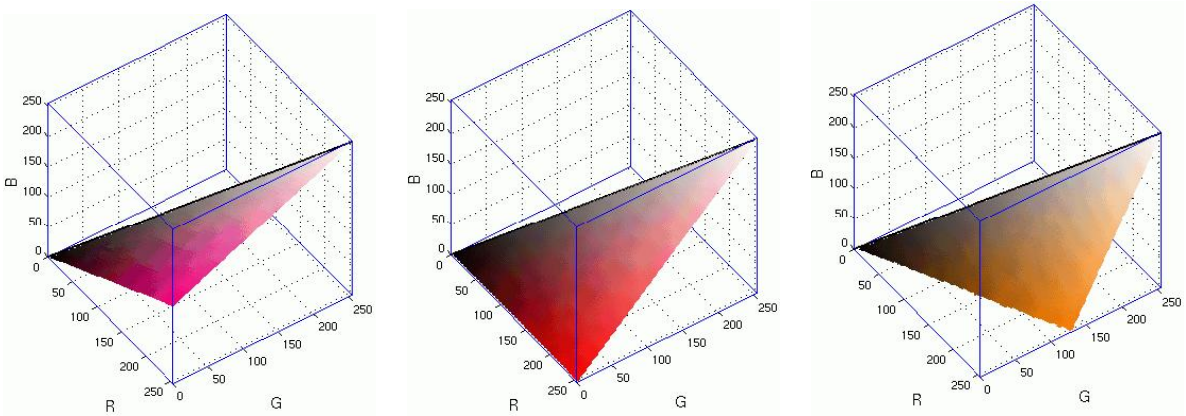


(a) (b)

Figura D.4: Cones de saturação constante no cubo RGB. (a) Saturação à 20% e (b) saturação à 70%.

D.3.3 Matiz

A definição de matiz está relacionada com o que coloquialmente pensa-se como cor. O matiz de um ponto no cubo RGB é definida como a posição angular do ponto em relação ao eixo neutro. Olhando para os cantos do cubo RGB, pode-se ver que o vermelho, amarelo, verde, ciano, azul e magenta, são distribuídos igualmente, em ângulo, em torno do eixo neutro. Assim, a cunha definida pelo eixo neutro e um ponto qualquer da superfície do cubo é um plano de matiz constante.



(a) (b) (c)

Figura D.5: Cunha de matiz constante no cubo RGB. (a) Matiz a 330^o, (b) Matiz a 0^o, (c) Matiz a 30^o.

A Figura D.5 ilustra três instâncias do cubo RGB com fatias diferentes de matiz constante. Devido ao fato do matiz ser uma função do ângulo, seu intervalo vai de 0^o até 360^o. O canto vermelho do cubo é definido como um matiz de 0^o, que é também um matiz de 360^o,

como se observa na Figura D.5.b. O cubo da Figura D.5.a apresenta um matiz de 330^o, que corresponde à cor púrpura, e na Figura D.5.c um matiz de 30^o, que corresponde à cor laranja. Note-se que, embora a matiz de cada cunha seja constante, o brilho e a saturação variam em toda a gama de valores possíveis.

Com o sistema de coordenadas HSB em mente, várias observações podem ser feitas sobre as regiões de cor no cubo RGB. A primeira é que os vértices ciano, magenta e amarelo do cubo representam cores mais vivas do que vermelho, verde e azul, devido ao fato destes últimos serem projetados na parte baixa no eixo neutro. Da mesma forma, todas as cores na pirâmide definida pelos vértices C, Y, M, W correspondem às cores mais claras, e a pirâmide definida pela origem e os vértices R, G e B correspondem às cores mais escuras. Cores perto do eixo neutro no cubo apresentam uma aparência em tons pastéis, porque elas são menos saturadas, e cores mais distantes deste eixo parecem mais vivas.

D.4 Conversão do RGB para HSB

Seja uma coordenada de uma cor $\mathbf{p}_0$ no cubo RGB (ver Figura D.6.a):

$$\mathbf{p}_0 = \begin{bmatrix} r_0 \\ g_0 \\ b_0 \end{bmatrix} = r_0\mathbf{u}_r + g_0\mathbf{u}_g + b_0\mathbf{u}_b,$$

onde

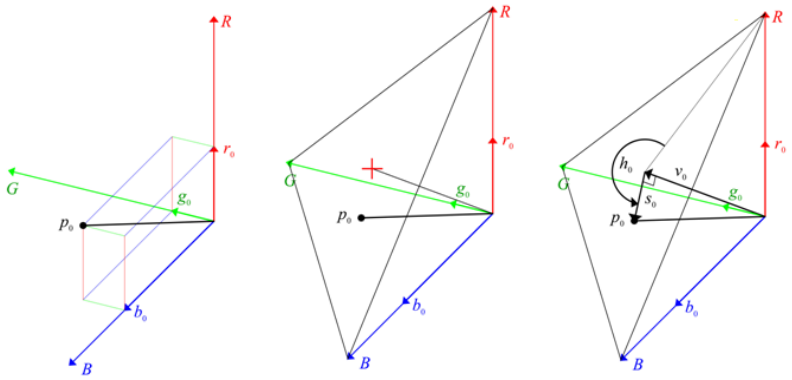
$$\mathbf{u}_r = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \mathbf{u}_g = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad \mathbf{u}_b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

É conhecido que todas as cores obtidas pela combinação de três cores estão dentro de um triângulo cujos vértices são definidos por três cores primárias. Assim o ponto $\mathbf{p}_0$ está no triângulo definido pelo plano

$$R + G + B = (r_0 + g_0 + b_0).$$

Tal plano intercepta os eixos R, G e B em $r_0 + g_0 + b_0$, contendo p_0 (ver Figura D.6.b), e é sobre este plano que são definidas as componentes do modelo de cor HSB matiz e saturação. Assim, observando a Figura D.6.c, tem-se que:

- O matiz, h_0, do ponto de cor $\mathbf{p}_0$ é o ângulo do vetor formado por $\mathbf{p}_0$ e o centro do triângulo em relação ao eixo vermelho. Portanto quando $h_0 = 0^o$, a cor é vermelha e quando $h_0 = 60^o$ a cor é amarela e assim sucessivamente;
- a saturação, s_0, do ponto de cor $\mathbf{p}_0$ é o grau em que a cor não é diluída pelo branco, e é proporcional à distância desde $\mathbf{p}_0$ ao centro do triângulo. Quando $\mathbf{p}_0$ é mais distante do centro do triângulo, a cor tem maior saturação;
- o brilho, v_0, do ponto de cor $\mathbf{p}_0$ é medido em relação a uma linha perpendicular ao triângulo, passando pelo seu centro. Assim: (a) um brilho ao longo desta linha que está abaixo do triângulo vai desde o escuro até preto; (b) um brilho acima do triângulo vai desde o claro até o branco.



(a) (b) (c)

Figura D.6: (a) Coordenada de uma cor $\mathbf{p}_0$ no cubo RGB. (b) Plano $R+G+B = (r_0+g_0+b_0)$ que intercepta os eixos R,G e B em $r_0 + g_0 + b_0$ e contém à cor p_0. (c). Componentes do modelo de cor HSB matiz e saturação sobre o plano $R + G + B = (r_0 + g_0 + b_0)$.

Combinando matiz, saturação e brilho gera-se um sólido quase-piramidal, onde

- qualquer ponto da superfície representa uma cor pura saturada;
- o matiz de qualquer cor é determinado por seu ângulo em relação ao eixo vermelho, e seu brilho pela distância perpendicular a partir do ponto preto (isto é, quanto maior a distância, a partir do preto, maior é o brilho da cor);

- comentários semelhantes aplicam-se a pontos no interior do sólido, sendo a única diferença que as cores tornam-se menos saturadas à medida que se aproximam do eixo neutro.

Para qualquer uma das três componentes r_0, g_0, e b_0, cada uma no intervalo $[0, 255]$, o vetor na direção do centro do triângulo é

$$\begin{aligned} \mathbf{v}_0 &= \frac{1}{3}(r_0 + g_0 + b_0)(\mathbf{u}_r + \mathbf{u}_g + \mathbf{u}_b) \\ &= \frac{1}{3}(r_0 + g_0 + b_0) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \frac{\sqrt{3}}{3}(r_0 + g_0 + b_0)\mathbf{u}_v \\ &= v_0\mathbf{u}_v. \end{aligned} \tag{D.1}$$

onde $\mathbf{u}_v = \frac{\sqrt{3}}{3}\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$, e o brilho no modelo de cor HSB é definido como

$$v_0 = ||\mathbf{v}_0|| = \frac{\sqrt{3}}{3}(r_0 + g_0 + b_0) = \sqrt{3}I_0,$$

onde $I_0 = \frac{1}{3}(r_0 + g_0 + b_0)$.

O próximo passo é obter h_0 e s_0. Para obter h_0 é necessário definir (a) o vetor $\mathbf{s}_0$, formado por $\mathbf{p}_0$ e o centro do triângulo e (b) o vetor $\mathbf{x}$, formado pelo centro do triângulo e o canto superior do triângulo em relação ao eixo vermelho.

$$
\begin{aligned}
\mathbf{s}_0 &= \mathbf{p}_0 - \mathbf{v}_0 \\
&= \begin{bmatrix} r_0 \\ g_0 \\ b_0 \end{bmatrix} - \frac{1}{3}(r_0+g_0+b_0)\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} r_0 \\ g_0 \\ b_0 \end{bmatrix} - I_0\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} r_0 - I_0 \\ g_0 - I_0 \\ b_0 - I_0 \end{bmatrix}, \\
\mathbf{x} &= (r_0+g_0+b_0)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \mathbf{v}_0 = (r_0+g_0+b_0)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{1}{3}(r_0+g_0+b_0)\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\
&= \frac{1}{3}(r_0+g_0+b_0)\begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix} = \frac{\sqrt{6}}{3}(r_0+g_0+b_0)\mathbf{u}_x = \sqrt{6}I_0\mathbf{u}_x,
\end{aligned}
$$

onde $\mathbf{u}_x = \frac{\sqrt{6}}{6}\begin{bmatrix} 2 & -1 & -1 \end{bmatrix}^T$. Finalmente define-se s_0 e h_0 como

$$
\begin{aligned}
s_0 &= ||\mathbf{s}_0|| = \sqrt{(r_0-I_0)^2+(g_0-I_0)^2+(b_0-I_0)^2}, \\
h_0 &= \cos^{-1}\left(\frac{\mathbf{s}_0\mathbf{x}}{||\mathbf{s}_0||||\mathbf{x}||}\right).
\end{aligned}
$$

Usualmente

- s_0 é normalizado para estar no intervalo de $(0,1)$;
- h_0 é deslocado para estar em $(0,2\pi)$.

D.5 Conversão de HSB para RGB

Para determinar a conversão de HSB para RGB é necessário expressar o ponto $\mathbf{p}_0$ através de um sistema de coordenadas localizado no plano do triângulo que contém o ponto $\mathbf{p}_0$, para o qual é necessário considerar que:

- o plano é perpendicular ao vetor $\mathbf{u}_v$,

- o plano contém o vetor $\mathbf{u}_x$ e, portanto, $\mathbf{u}_v$ é perpendicular a $\mathbf{u}_x$, e assim o vetor definido através do produto externo $\mathbf{u}_y$ definido como

$$\mathbf{u}_y = \mathbf{u}_v \times \mathbf{u}_x = \frac{\sqrt{2}}{2} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix},$$

  também está contido no plano;

- portanto, os vetores unitários $\mathbf{u}_x$, $\mathbf{u}_y$ e $\mathbf{u}_v$ constituem uma base ortonormal em relação ao plano que contém o triângulo.

Assim, todo ponto $\mathbf{p}_0$ pode ser representado como a soma dos vetores

$$\mathbf{p}_0 = \mathbf{s}_0 + \mathbf{v}_0 \ , \ \mathbf{s}_0 \perp \mathbf{v}_0. \tag{D.2}$$

Neste caso, o vetor $\mathbf{s}_0$, por estar no plano do triângulo, pode ser representado como uma combinação linear dos vetores $\mathbf{u}_x$ e $\mathbf{u}_y$, e, assim tem-se

$$\begin{aligned} \mathbf{s}_0 &= (\mathbf{u}_x \mathbf{s}_0)\, \mathbf{u}_x + (\mathbf{u}_y \mathbf{s}_0)\, \mathbf{u}_y \\ &= s_0 \cos(h_0)\mathbf{u}_x + s_0 \sin(h_0)\mathbf{u}_y. \end{aligned} \tag{D.3}$$

Substituindo (D.2) e (D.3) em (D.1), tem-se

$$\begin{aligned} \mathbf{p}_0 &= \mathbf{s}_0 + \mathbf{v}_0 \\ &= s_0 \cos(h_0)\mathbf{u}_x + s_0 \sin(h_0)\mathbf{u}_y + v_0 \mathbf{u}_v \\ &= s_0 \cos(h_0) \frac{\sqrt{6}}{6} \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix} + s_0 \sin(h_0) \frac{\sqrt{2}}{2} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} + v_0 \frac{\sqrt{3}}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \end{aligned}$$

Finalmente, tem-se que as relações entre os valores de r_0, g_0 e b_0 e v_0, s_0 e h_0, a saber,

$$
\begin{aligned}
r_0 &= \frac{\sqrt{6}}{6} s_0 \cos(h_0) + \frac{\sqrt{3}}{3} v_0 \\
g_0 &= \frac{\sqrt{2}}{2} s_0 \sin(h_0) - \frac{\sqrt{6}}{6} s_0 \cos(h_0) + \frac{\sqrt{3}}{3} v_0 \\
b_0 &= -\frac{\sqrt{6}}{6} s_0 \cos(h_0) - \frac{\sqrt{2}}{2} s_0 \sin(h_0) + \frac{\sqrt{3}}{3} v_0
\end{aligned}
$$

D.6 Diferença de cores através do HSB

Sabe-se que o modelo de cor HSB permite representar cada cor do cubo RGB através de suas componentes de brilho e cromaticidade, permitindo assim uma análise em função das características próprias das cores. Nesse sentido, para ter uma melhor compreensão da diferença de cores, expressa-se as distancias estudadas na Seção 2.8, d_{0^o}, d_{90^o} e λ, através do modelo de cor HSB. Neste contexto, os píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ são expressos considerando a representação HSB definida pela Equação (D.2), ou seja

$$\mathbf{I}_l(t) = \mathbf{s}_I(t) + \mathbf{v}_I(t) \ , \ \mathbf{s}_I(t) \perp \mathbf{v}_I(t), \tag{D.4}$$

$$\mathbf{B}_l(t) = \mathbf{s}_B(t) + \mathbf{v}_B(t) \ , \ \mathbf{s}_B(t) \perp \mathbf{v}_B(t). \tag{D.5}$$

Todos os parâmetros relevantes desta representação são apresentados na Figura D.7.a, onde na Figura D.7.b tem-se uma vista desconsiderando a matiz e na Figura D.7.c tem-se uma vista desconsidernado o brilho.

D.6.1 Análise das magnitudes d_{0^o} e d_{90^o}

Substituindo (D.4) e (D.5) em (2.49), obtém-se

$$
\begin{aligned}
d_{0^o}^2 &= \left(||\mathbf{I}_l(t)|| - ||\mathbf{B}_l(t)||\right)^2 \\
&= \left(\sqrt{s_I^2 + v_I^2} - \sqrt{s_B^2 + v_B^2}\right)^2 .
\end{aligned} \tag{D.6}
$$

Igualmente, ao substituir (D.4) e (D.5) em (2.50) obtém-se

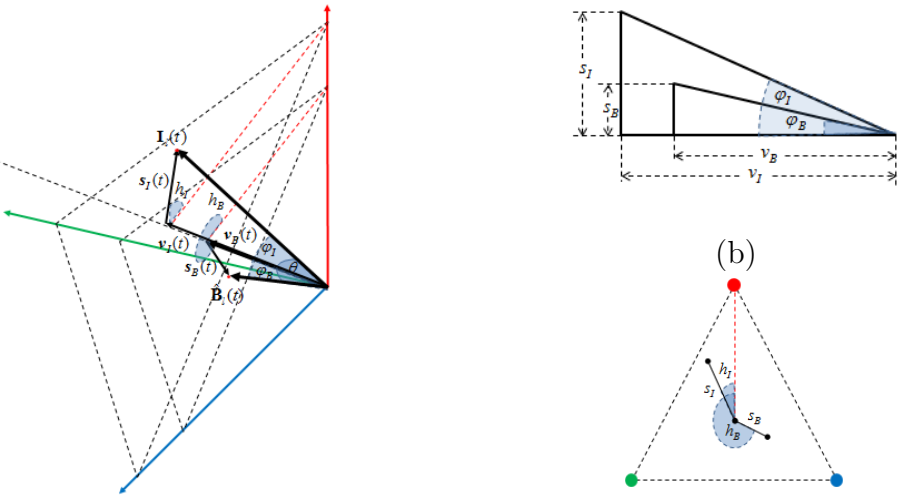


Figura D.7: Representação dos píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ no modelo de cor HSB: (a) vista espacial;(b) desconsiderando a matiz;(c) desconsiderando o brilho.

$$\begin{aligned} d^2_{90^o} &= ||\mathbf{I}_l(t)||^2 + ||\mathbf{B}_l(t)||^2 \\ &= s_I^2 + v_I^2 + s_B^2 + v_B^2. \end{aligned} \tag{D.7}$$

A partir (D.6) e (D.7) pode-se concluir que d_{0^o} e d_{90^o} dependem unicamente das componentes de saturação e brilho dos píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$.

D.6.2 Análise do ângulo de abertura θ

Substituindo (D.4) e (D.5) em (2.46) tem-se

$$\begin{aligned} \cos(\theta) &= \frac{\mathbf{I}_l^T(t)\mathbf{B}_l(t)}{||\mathbf{I}_l(t)||||\mathbf{B}_l(t)||} \\ &= \frac{(\mathbf{s}_I(t)+\mathbf{v}_I(t))^T(\mathbf{s}_B(t)+\mathbf{v}_B(t))}{||\mathbf{s}_I(t)+\mathbf{v}_I(t)||||\mathbf{s}_B(t)+\mathbf{v}_B(t)||} \\ &= \frac{\mathbf{s}_I^T(t)\mathbf{s}_B(t)+\mathbf{s}_I^T(t)\mathbf{v}_B(t)+\mathbf{v}_I^T(t)\mathbf{s}_B(t)+\mathbf{v}_I^T(t)\mathbf{v}_B(t)}{\sqrt{\mathbf{s}_I^2(t)+\mathbf{v}_I^2(t)}\sqrt{\mathbf{s}_B^2(t)+\mathbf{v}_B^2(t)}}. \end{aligned} \tag{D.8}$$

Desenvolvendo cada um dos termos do numerador de (D.8) obtém-se

$$
\begin{aligned}
\mathbf{s}_I^T(t)\mathbf{s}_B(t) &= (s_I\cos(h_I)\mathbf{u}_x + s_I\sin(h_I)\mathbf{u}_y)^T(s_B\cos(h_B)\mathbf{u}_x + s_B\sin(h_B)\mathbf{u}_y) \\
&= s_I\cos(h_I)\mathbf{u}_x^T(s_B\cos(h_B)\mathbf{u}_x + s_B\sin(h_B)\mathbf{u}_y) + \\
&\quad s_I\sin(h_I)\mathbf{u}_y^T(s_B\cos(h_B)\mathbf{u}_x + s_B\sin(h_B)\mathbf{u}_y) \\
&= s_Is_B\left(\cos(h_I)\cos(h_B)||\mathbf{u}_x||^2 + \sin(h_I)\sin(h_B)||\mathbf{u}_y||^2\right) + \\
&\quad s_Is_B\left(\cos(h_I)\sin(h_B) + \sin(h_I)\cos(h_B)\right)\mathbf{u}_x^T\mathbf{u}_y \\
&= s_Is_B\left(\cos(h_I)\cos(h_B) + \sin(h_I)\sin(h_B)\right) \\
&= s_Is_B\cos(h_I - h_B), \qquad \text{(D.9)} \\
\mathbf{s}_I^T(t)\mathbf{v}_B(t) &= (s_I\cos(h_I)\mathbf{u}_x + s_I\sin(h_I)\mathbf{u}_y)^T v_B\mathbf{u}_v \\
&= s_Iv_B\cos(h_I)\mathbf{u}_x^T\mathbf{u}_v + s_Iv_B\sin(h_I)\mathbf{u}_y^T\mathbf{u}_v = 0, \qquad \text{(D.10)} \\
\mathbf{v}_I^T(t)\mathbf{s}_B(t) &= \mathbf{s}_B^T(t)\mathbf{v}_I(t) = (s_B\cos(h_B)\mathbf{u}_x + s_B\sin(h_B)\mathbf{u}_y)^T v_I\mathbf{u}_v \\
&= s_Bv_I\cos(h_B)\mathbf{u}_x^T\mathbf{u}_v + s_Bv_I\sin(h_B)\mathbf{u}_y^T\mathbf{u}_v = 0, \qquad \text{(D.11)} \\
\mathbf{v}_I^T(t)\mathbf{v}_B(t) &= v_Iv_B\mathbf{u}_v^T\mathbf{u}_v \\
&= v_Iv_B. \qquad \text{(D.12)}
\end{aligned}
$$

Substituindo (D.9)-(D.12) em (D.8) tem-se

$$
\begin{aligned}
\cos(\theta) &= \frac{s_Is_B\cos(h_I - h_B) + v_Iv_B}{\sqrt{s_I^2 + v_I^2}\sqrt{s_B^2 + v_B^2}} \\
&= \left(\frac{s_I}{\sqrt{s_I^2 + v_I^2}}\right)\left(\frac{s_B}{\sqrt{s_B^2 + v_B^2}}\right)\cos(h_I - h_B) + \left(\frac{v_I}{\sqrt{s_I^2 + v_I^2}}\right)\left(\frac{v_B}{\sqrt{s_B^2 + v_B^2}}\right) \\
&= \sin(\varphi_I)\sin(\varphi_B)\cos(h_I - h_B) + \cos(\varphi_I)\cos(\varphi_B) \\
&= \frac{1}{2}\left(\cos(\varphi_I - \varphi_B) - \cos(\varphi_I + \varphi_B)\right)\cos(h_I - h_B) + \frac{1}{2}\left(\cos(\varphi_I + \varphi_B) + \cos(\varphi_I - \varphi_B)\right) \\
&= \frac{1}{2}\left(\cos(\varphi_I - \varphi_B)\cos(h_I - h_B) - \cos(\varphi_I + \varphi_B)\cos(h_I - h_B) + \cos(\varphi_I + \varphi_B) + \cos(\varphi_I - \varphi_B)\right) \\
&= \frac{1}{2}\left(\cos(\varphi_I - \varphi_B)\left(1 + \cos(h_I - h_B)\right) + \cos(\varphi_I + \varphi_B)\left(1 - \cos(h_I - h_B)\right)\right) \\
&= \cos(\varphi_I - \varphi_B)\cos^2\left(\frac{h_I - h_B}{2}\right) + \cos(\varphi_I + \varphi_B)\sin^2\left(\frac{h_I - h_B}{2}\right). \qquad \text{(D.13)}
\end{aligned}
$$

Definindo as diferenças angulares $\varphi_I - \varphi_B$ e $h_I - h_B$ como

$$
\begin{aligned}
\Delta\varphi &= \varphi_I - \varphi_B \qquad \text{(D.14)} \\
\Delta h &= h_I - h_B, \qquad \text{(D.15)}
\end{aligned}
$$

e substituindo-as em (D.13), obtém-se

$$\cos(\theta) = \cos(\Delta\varphi)\cos^2\left(\frac{\Delta h}{2}\right) + \cos(2\varphi_B + \Delta\varphi)\sin^2\left(\frac{\Delta h}{2}\right). \tag{D.16}$$

A Equação (D.16) depende das variáveis φ_B, $\Delta\varphi$ e Δh, onde

- o valor que $\Delta\varphi$ assume deve-se: (a) apenas à variação da saturação (ver Figura D.8.a), (b) apenas à variação do brilho (ver Figura D.8.b) ou (c) à variação conjunta tanto da saturação como do brilho (ver Figura D.8.c);
- por outro lado, o valor que Δh assume deve-se à variação das matizes. Portanto, esta variação será maior se a relação das cores dos píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ for mais de caráter complementar (ver Figura D.9.a) que análogo (ver Figura D.9.b); tendo um valor de zero se a relação é monocromática (ver Figura D.9.c).

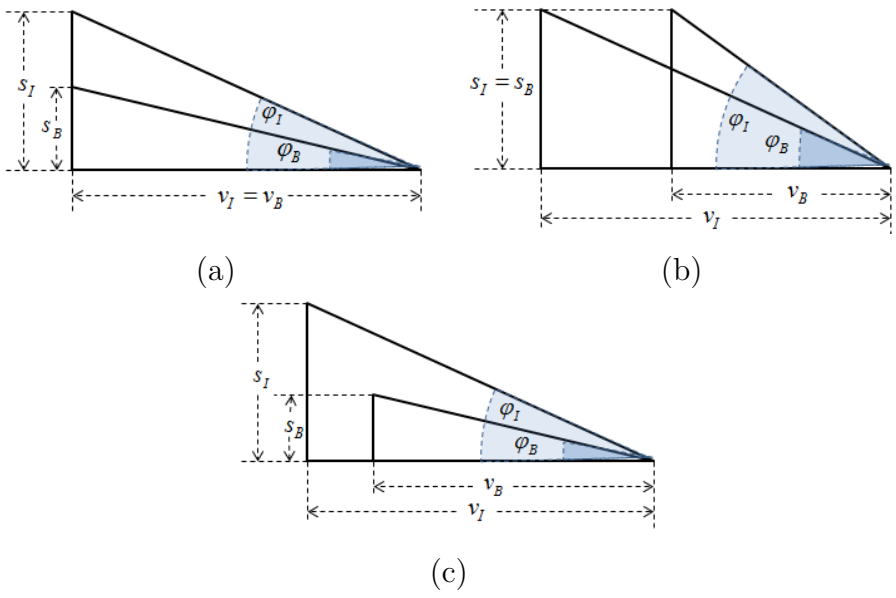


Figura D.8: Possíveis casos que dão origem ao $\Delta\varphi$: (a) devido unicamente a uma variação da saturação;(b) devido unicamente a uma variação do brilho;(c)a uma variação conjunta tanto da saturação como do brilho.

Portanto, em termos gerais pode-se dizer que o $cos(\theta)$ depende das componentes de brilho e cromaticidade dos píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$. Agora se é considerada, em forma restrita, a problemática da detecção de variações, onde para o caso de um pixel $\mathbf{I}_l(t)$ pertencente ao fundo ele fica numa pequena vizinhança centrada em $\mathbf{B}_l(t)$, como é representado na Figura D.10.a, tem-se que à medida que a vizinhança diminui, a relação $\Delta\varphi \approx 0$ será valida

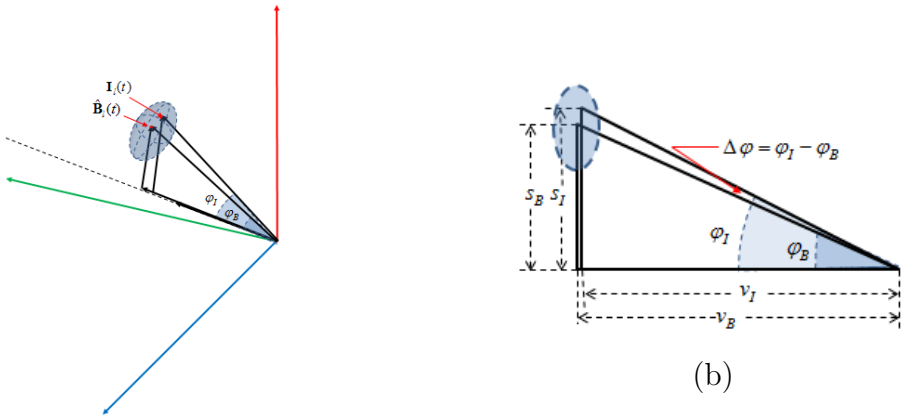
(a) (b) (c)

Figura D.9: Relações dos matizes: (a) complementarias;(b) análogas;(c) monocromáticas.

(uma representação gráfica é apresentada na Figura D.10.b), e, assim, a Equação (D.16) é aproximada por

$$\begin{aligned} \cos(\theta) &\approx \cos^2\left(\frac{\Delta h}{2}\right) + \cos(2\varphi_B)\sin^2\left(\frac{\Delta h}{2}\right) \\ &= 1 - (1 - \cos(2\varphi_B))\sin^2\left(\frac{\Delta h}{2}\right) \\ &= 1 + 4\left(\sin(\varphi_B)\sin\left(\frac{\Delta h}{2}\right)\right)^2. \end{aligned} \tag{D.17}$$

A Equação (D.17) indica que para o caso de detecção de variações, e considerando um pixel $\mathbf{B}_l(t)$ constante, o $\cos(\theta)$ dependerá somente do valor da diferença de matizes, Δh. Assim, as componentes de cromaticidade de $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$, especificamente a diferença de matizes, Δh, define o valor de $\cos(\theta)$.



(a) (b)

Figura D.10: Píxeis $\mathbf{I}_l(t)$ e $\mathbf{B}_l(t)$ considerado a problemática da detecção de variações (a) representação espacial; (b)representação desconsiderando a matiz dos píxeis.

Referências Bibliográficas

[1] CAVIAR Test Case Scenarios. "http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/", 2007.

[2] Aach, T., Dümbgen, L., and Mester, R. Bayesian illumination invariant change detection using a total least squares test statistic. In *Actes/Proceedings 18e Colloque GRETSI sur le Traitement du Signal et des Images, Toulouse, France* (2001), pp. 587–590.

[3] Aach, T., Dumbgen, L., Mester, R., and Toth, D. Bayesian illumination-invariant motion detection. In *Image Processing, 2001. Proceedings. 2001 International Conference on* (2001), vol. 3, IEEE, pp. 640–643.

[4] Aach, T., and Kaup, A. Bayesian algorithms for adaptive change detection in image sequences using Markov random fields. *Signal Processing: Image Communication 7*, 2 (1995), 147–160.

[5] Black, J., Ellis, T., and Rosin, P. A novel method for video tracking performance evaluation. In *In Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS* (2003), pp. 125–132.

[6] Blackman, S. S. Multiple hypothesis tracking for multiple target tracking. *Aerospace and Electronic Systems Magazine, IEEE 19*, 1 (2004), 5–18.

[7] Bouwmans, T., Baf, F. E., and Vachon, B. Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey. *Science 3* (2008), 219–237.

[8] Brown, L. M., Senior, A. W., li Tian, Y., Connell, J., Hampapur, A., fe Shu, C., Merkl, H., and Lu, M. Performance evaluation of surveillance systems under varying conditions. In *In: Proceedings of IEEE PETS Workshop* (2005), pp. 1–8.

[9] Brutzer, S., Höferlin, B., and Heidemann, G. Evaluation of Background Subtraction Techniques for Video Surveillance. In *Computer Vision and Pattern Recognition (CVPR)* (2011), IEEE, pp. 1937–1944.

[10] Cavallaro, A. Change detection based on color edges. *, 2001. ISCAS 2001. The 2001 IEEE 2* (2001), 141–144.

[11] Chen, H., Liu, T., and Fuh, C. Probabilistic tracking with adaptive feature selection. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on* (2004), vol. 2, IEEE, pp. 736–739.

[12] Cheung, S.-c. S. Robust techniques for background subtraction in urban traffic video. *Proceedings of SPIE* (2004), 881–892.

[13] Collins, R., a.J. Lipton, and Kanade, T. Introduction to the special section on video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 8 (aug 2000), 745–746.

[14] Collins, R. T., and Liu, Y. On-line selection of discriminative tracking features. *Proceedings Ninth IEEE International Conference on Computer Vision* (2003), 346–352 vol.1.

[15] Correia, P., and Pereira, F. Change detection-based video segmentation for survillance applications, 2004.

[16] Cristani, M., Bicego, M., and Murino, V. Integrated region-and pixel-based approach to background modelling. In *In Proc. of IEEE Workshop on Motion and Video Computing* (2002), pp. 3–8.

[17] Cristani, M., Farenzena, M., Bloisi, D., and Murino, V. Background subtraction for automated multisensor surveillance: a comprehensive review. *EURASIP Journal on Advances in Signal Processing 2010* (2010), 43.

[18] Cucchiara, R., Grana, C., Piccardi, M., and Prati, a. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 10 (oct 2003), 1337–1342.

[19] Cutler, R., and Davis, L. View-based detection and analysis of periodic motion. *Proceedings. Fourteenth International Conference on Pattern Recognition 1* (1998), 495–500.

[20] Elgammal, A., Harwood, D., and Davis, L. Non-parametric model for background subtraction. *Computer* (2000), 751–767.

[21] Elhabian, S., El-Sayed, K., and Ahmed, S. Moving object detection in spatial domain using background removal techniques-state-of-art. *Recent patents on computer science 1*, 1 (2008), 32–54.

[22] Feris, R., Hampapur, A., Zhai, Y., Bobbitt, R., Brown, L., Vaquero, D., Tian, Y., Liu, H., and Sun, M. Case Study: IBM Smart Surveillance System. *Intelligent Video Surveillance: System and Technology* (2009), 47–76.

[23] Fisher, R. The PETS04 surveillance ground-truth data sets. *Evaluation of Tracking and Surveillance* (2004), 1–5.

[24] Fisher, R. CAVIAR Test Case Scenarios. `http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/`, 2011.

[25] Förstner, W. 10 pros and cons against performance characterization of vision algorithms. *... on Peformance Characteristics of Vision Algorithms* (1996).

[26] Friedman, N., and Russell, S. Image segmentation in video sequences: A probabilistic approach. *Proceedings of the Thirteenth conference on ...* (1997), 1–13.

[27] Gao, X., Boult, T., Coetzee, F., and Ramesh, V. Error analysis of background adaption. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on* (2000), vol. 1, IEEE, pp. 503–510.

[28] Greenhill, S., Venkatesh, S., and West, G. Adaptive model for foreground extraction in adverse lighting conditions. *PRICAI 2004: Trends in Artificial Intelligence* (2004), 805–811.

[29] Hall, D., Pesnel, S., Emonet, R., Crowley, J., Nascimento, J., Ribeiro, P., Moreno, P., Victor, J., Andrade, E., List, T., and Fisher, R. Comparison of Target Detection Algorithms using Adaptive Background Models. *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 0 (2005), 113–120.

[30] Haritaoglu, I., Cutler, R., Harwood, D., and Davis, L. S. Backpack: Detection of People Carrying Objects Using Silhouettes. *Computer Vision and Image Understanding 81*, 3 (Mar. 2001), 385–397.

[31] Haritaoglu, I., and Harwood, D. W4: Who, when, where, what: a real time system for detecting and tracking people. 3rd IEEE Int. *Conf. Automatic Face and Gesture Recognition,* (1998), 877–892.

[32] Haritaoglu, I., Harwood, D., and Davis, L. W/sup 4/: real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 8 (2000), 809–830.

[33] Heikkila, J. A real-time system for monitoring of cyclists and pedestrians. *Image and Vision Computing 22*, 7 (jul 2004), 563–570.

[34] Heikklä, M., and Pietikäinen, M. A texture-based method for modeling the background and detecting moving objects. *IEEE transactions on pattern analysis and machine intelligence 28*, 4 (Apr. 2006), 657–62.

[35] Jonker, R., and Volgenant, A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing 38* (November 1987), 325–340.

[36] Kingsland, S. E. *Modeling nature : episodes in the history of population ecology / Sharon E. Kingsland.* University of Chicago Press, Chicago, 1985.

[37] Klare, B., and Sarkar, S. Background subtraction in varying illuminations using an ensemble based on an enlarged feature set. *Computer Vision and Pattern Recognition ...* (June 2009), 66–73.

[38] Koller, D., Weber, J., Huang, T., Malik, J., Ogasawara, G., Rao, B., and Russell, S. Towards robust automatic traffic scene analysis in real-time. *Proceedings of 1994 33rd IEEE Conference on Decision and Control*, 3776–3781.

[39] Lazarevic-McManus, N., Renno, J., and Jones, G. a. Performance evaluation in visual surveillance using the F-measure. *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks - VSSN '06* (2006), 45.

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

[41] Lien, C. Targets Tracking in the Crowd. *cdn.intechopen.com* (2011).

[42] Martinez-Contreras, F., Orrite-Urunuela, C., Herrero-Jaraba, E., Ragheb, H., and Velastin, S. a. Recognizing Human Actions Using Silhouette-based HMM. *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance* (sep 2009), 43–48.

[43] Mason, M., and Duric, Z. Using histograms to detect and track objects in color video. In *Applied Imagery Pattern Recognition Workshop, AIPR 2001 30th* (2001), IEEE, pp. 154–159.

[44] McFarlane, N., and Schofield, C. Segmentation and tracking of piglets in images. *Machine Vision and Applications 8*, 3 (May 1995), 187–193.

[45] Mittal, A., and Paragios, N. Motion-based background subtraction using adaptive kernel density estimation. *..., 2004. CVPR 2004. Proceedings of the ...* (2004).

[46] Nascimento, J., and Marques, J. New performance evaluation metrics for object detection algorithms. In *IEEE Workshop on Performance Analysis of Video Surveillance and Tracking (PETS 2004)* (2004).

[47] Nascimento, J., and Marques, J. Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia 8*, 4 (aug 2006), 761–774.

[48] Nguyen, H. T., and Smeulders, A. W. Robust tracking using foreground-background texture discrimination. *Int. J. Comput. Vision 69* (September 2006), 277–293.

[49] Noriega, P., and Bernier, O. Real time illumination invariant background subtraction using local kernel histograms. *British Machine Vision Association (BMVC)* (2006), 1–10.

[50] Oberti, F., Teschioni, a., and Regazzoni, C. ROC curves for performance evaluation of video sequences processing systems for surveillance applications. *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348) 2*, 949–953.

[51] Ojala, T., Pietikainen, M., and Harwood, D. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision Image Processing., Proceedings of the 12th IAPR International Conference on* (oct 1994), vol. 1, pp. 582 –585 vol.1.

[52] Oliver, N. M., Rosario, B., and Pentland, A. P. A bayesian computer vision system for modeling human interactions. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 22*, 8 (2000), 831–843.

[53] Parks, D. H., and Fels, S. S. Evaluation of Background Subtraction Algorithms with Post-Processing. *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance* (sep 2008), 192–199.

[54] Pathan, S., Al-Hamadi, A., and Michaelis, B. Intelligent feature-guided multi-object tracking using Kalman filter. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on* (2009), IEEE, pp. 1–6.

[55] Piccardi, M. Background subtraction techniques: a review. *2004 IEEE International Conference on Systems Man and Cybernetics IEEE 4*, C (2004), 3099–3104.

[56] Radke, R. J., Andra, S., Al-Kofahi, O., and Roysam, B. Image change detection algorithms: a systematic survey. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 14*, 3 (mar 2005), 294–307.

[57] Rosin, P. L., and Ioannidis, E. Evaluation of global image thresholding for change detection. *Pattern Recognition Letters 24*, 14 (Oct. 2003), 2345–2356.

[58] Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision 77*, 1-3 (May 2008), 157–173.

[59] Schlogl, T., Beleznai, C., Winter, M., and Bischof, H. Performance evaluation metrics for motion detection and tracking. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on* (aug. 2004), vol. 4, pp. 519 – 522 Vol.4.

[60] Shimkin, N. Kinematic Models for Target Tracking. Israel, 2009.

[61] Sobral, A., Oliveira, L., Schnitman, L., and Souza, F. D. Highway traffic congestion classification using holistic properties. *10th IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA'2013)* (fev. 2013).

[62] Song, K., and Tai, J. Real-time background estimation of traffic imagery using group-based histogram. *Journal of Information Science and Engineering 423* (2008), 411–423.

[63] Stauffer, C., and Grimson, W. Adaptive background mixture models for real-time tracking. *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, 246–252.

[64] Stefano, L. D., and Neri, G. Analysis of pixel-level algorithms for video surveillance applications. *Image Analysis and* (2001).

[65] Stenger, B., Ramesh, V., Paragios, N., Coetzee, F., and Buhmann, J. Topology free hidden Markov models: application to background modeling. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001 1*, 294–301.

[66] Tai, J. Background segmentation and its application to traffic monitoring using modified histogram. *Networking, Sensing and Control, 2004* (2004), 13–18.

[67] Toyama, K., Krumm, J., Brumitt, B., and Meyers, B. Wallflower: principles and practice of background maintenance. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, September (1999), 255–261 vol.1.

[68] Villegas, P., and Marichal, X. Perceptually-weighted evaluation criteria for segmentation masks in video sequences. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 13*, 8 (aug 2004), 1092–103.

[69] Vlahos, J. Surveillance society: New high-tech cameras are watching you. *Popular Mechanics* (January 2008), 64–69.

[70] Wallis, S. Competition between choices over time. Tech. rep., University College London, 2010.

[71] Wang, H., and Suter, D. A re-evaluation of mixture of Gaussian background modeling. *Acoustics, Speech, and Signal Processing, ... 2* (2005), 1017–1020.

[72] Wang, H., and Suter, D. Background subtraction based on a robust consensus method. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 01* (Washington, DC, USA, 2006), ICPR '06, IEEE Computer Society, pp. 223–226.

[73] Wauthier, F. Motion Tracking Project Synopsis. http://www.cs.berkeley.edu/ flw/-tracker/, 2010.

[74] Wesolkowski, S., and Jernigan, E. Color edge detection in RGB using jointly euclidean distance and vector angle. In *Vision interface* (1999), vol. 99, pp. 19–21.

[75] Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A. Pfinder: Real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 19*, 7 (1997), 780–785.

[76] Wu, H., and Zheng, Q. Self-evaluation for video tracking systems. *In Proceedings of the 24th Army Science Conference* (november. 2004).

[77] Xiong, Q., and Jaynes, C. Multi-resolution background modeling of dynamic scenes using weighted match filters. In *Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks* (New York, NY, USA, 2004), VSSN '04, ACM, pp. 88–96.

[78] Yao, J., and Odobez, J.-M. Multi-Layer Background Subtraction Based on Color and Texture. *2007 IEEE Conference on Computer Vision and Pattern Recognition* (June 2007), 1–8.

[79] Yilmaz, A., Javed, O., and Shah, M. Object tracking: A Survey. *ACM Computing Surveys 38*, 4 (dec 2006), 1–44.

[80] Zhang, Y.-J. A review of recent evaluation methods for image segmentation. *Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat.No.01EX467) 1* (2001), 148–151.

[81] Zhang, Y.-J. A Summary of Recent Progresses for Segmentation Evaluation. *Advances in image and video segmentation* (2006), 422–439.

[82] Zhang, Y.-J., Fritts, J. E., and Goldman, S. a. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding 110*, 2 (May 2008), 260–280.

[83] Zheng, J., Wang, Y., Nihan, N., and Hallenbeck, M. Extracting roadway background image: Mode-based approach. *Transportation Research Record: Journal of the Transportation Research Board 1944*, -1 (2006), 82–88.