Detector De Humanos Em Imagens Digitais Utilizando Extratores De Características Indicadoras De Movimento

Trabalho de Conclusão de Curso

Engenharia de Computação

Aluno: Rodrigo Fonseca de Santa Cruz Oliveira

Orientador: Prof. Dr. Carmelo José Albanez Bastos Filho

UNIVERSIDADE
DE PERNAMBUCO

Rodrigo Fonseca de Santa Cruz Oliveira

Detector De Humanos Em Imagens Digitais Utilizando Extratores De Características Indicadoras De Movimento

Monografia apresentada para obtenção do Grau de Bacharel em Engenharia de Computação pela Universidade de Pernambuco

Orientador:
Prof. Dr. Carmelo José Albanez Bastos Filho

Graduação em Engenharia de Computação
Escola Politécnica de Pernambuco
Universidade de Pernambuco

Recife - PE, Brasil

Junho de 2014

ESCOLA POLITÉCNICA
DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 2 de 7 de 2014, às 14:00 horas, reuniu-se para deliberar a defesa da monografia de conclusão de curso do discente RODRIGO FONSECA DE SANTA CRUZ OLIVEIRA, orientado pelo professor Carmelo José Albanez Bastos Filho, sob título Detector De Humanos Em Imagens Digitais Utilizando Extratores De Características Indicadoras De Movimento, a banca composta pelos professores:

Bruno José Torres Fernandes

Carmelo José Albanez Bastos Filho

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

☒ Aprovada ☐ Aprovada com Restrições* ☐ Reprovada

e foi-lhe atribuída nota: 10,0 (dez)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O discente terá 7 dias para entrega da versão final da monografia a contar da data deste documento.

BRUNO JOSÉ TORRES FERNANDES

CARMELO JOSÉ ALBANEZ BASTOS FILHO

* Este documento deverá ser encadernado juntamente com a monografia em versão final.

De acordo

Recife

13/ 06 / 2014

Prof. Dr. Carmelo Bastos-Filho
Orientador

Resumo

Detectar humanos em imagens digitais é uma tarefa recorrente e importante para o desenvolvimento de várias aplicações. Essa tarefa é desafiadora dado que a aparência humana pode variar amplamente. No entanto, alguns detectores de humanos em imagens digitais, como por exemplo o *Aggregated Chanel Features* (ACF), já foram desenvolvidos. Esses detectores normalmente utilizam características indicadoras do formato do objeto que são boas evidências da presença humana. Porém, esses detectores costumam gerar considerável quantidade de alarmes falsos (e.g. pedaço de background erroneamente classificado como humano) que tornam proibitiva sua utilização em aplicações reais.

Neste trabalho de conclusão de curso foi desenvolvido um detector de humanos em imagens digitais que consiste em uma extensão do ACF. Essa extensão se diferencia do ACF pela utilização de características indicadoras de movimento. Para utilizar essas características, uma interface para adição de características extraídas de sequência de imagens foi desenvolvida e integrada ao ACF. Entre as caraterísticas testadas, a escolhida para essa extensão foi a característica indicadora de movimento *Internal Motion Histogram Central Difference* (IMHcd) devido sua melhor performance nos experimentos realizados. A IMHcd consiste em calcular a diferença de fluxo óptico em várias direções. Ela se mostrou eficaz na captura do movimento de partes do objeto em relação ao corpo do objeto. O movimento de membros e articulações dos humanas em relação ao seu corpo são bons exemplos desse tipo de movimento que são boas evidências para detecção de humanos.

A técnica desenvolvida foi avaliada e comparada ao ACF original. Essa avaliação utilizou o banco de dados *Caltech Pedestrian Data Set* e métricas baseadas em *miss-rate* e falso positivos por imagens (FPPI) que são métricas consagradas da área de detectores de humanos em imagens digitais. Essas simulações utilizaram o procedimento de simulação proposto nesse trabalho que consiste em uma melhoria do procedimento de simulação padrão para o banco de dados utilizado. O procedimento proposto e utilizado nas simulações tem o intuito de ser informativo e não tendencioso.

As simulações mostraram que a extensão proposta alcançou melhor performance de detecção em comparação com o ACF original. Ela se mostrou mais precisa que o ACF original principalmente em cenários pouco restritivos, baixo limiar de detecção, onde o *miss-rate* é baixo mas há uma grande geração de alarmes falsos. A extensão proposta também apresentou melhores taxas de falsos positivos por imagens (FPPI) em valores específicos de *miss-rate*. Essa diferença fica ainda mais evidente quando escolhemos valores de referência de *miss-rate* mais baixos, que caracterizam um cenário pouco restritivo com baixo limiar de detecção. Portanto, a utilização de características indicadoras de movimento se mostrou um boa ferramenta para redução de alarmes falsos.

Abstract

Human detection in digital images is one of the most important tasks for the development of many applications. This is a challenging task because human appearance may vary widely. However, some humans detectors have been developed such as the Aggregated Chanel Features (ACF). Detectors usually use shape features because they are good evidence of human presence, and they provide the best results in the literature. Nevertheless, those detectors often generate a considerable amount of false alarms (e.g. Background pieces misclassified as human), it makes the employment of them in some applications almost unfeasible.

This monograph describes the development of a new human detector. This new detector is in fact an extension of the ACF. This extension uses motion features that is the main difference between it and the original ACF. In addition, an interface to include motion features to the original ACF's feature set was developed. This interface improves the ACF's capabilities because it can use that interface to work with any feature that extracts information from sequences of images. The Internal Motion histogram Central Difference (IMHcd) was chosen among tested motion features because it presented the best results in the performed experiments. The IMHcd basically computes optical flow differences in many directions, consequently, this feature extractor presents good performance to capture movement of body parts in relation to the object body movement. The human members and articulations are good examples for this kind of movement, and it is a good source of information for the human recognition.

The developed technique was evaluated and compared to the original ACF. This evaluation used the Caltech Pedestrian Data Set and detection performance metrics based on miss-rate and false positives per images (FPPI) that are well known metrics in the human detection field. These simulations followed the simulation procedure proposed in this monograph. The proposed simulation procedure is an improvement of the standard simulation procedure suggested for use with the Caltech Pedestrian Data Set. The proposed simulation procedure aims to be informative and unbiased.

The Simulations implied that the proposed detector had better results in detection performance than the original ACF. It was more accurate than the original ACF, mainly, in less restrictive scenarios, with a low detection threshold, where there is a low miss-rate and a high FPPI. The proposed technique also presented lower FPPI in specifics reference levels of miss-rate than the original ACF. These results are even clearer when we look at low miss-rate reference levels which characterize a less restrictive scenario with a low detection threshold. Therefore, motion features are good tools for false alarm reduction in human detection.

Agradecimentos

Primeiramente aos meus pais, Lusanira Maria da Fonseca de Santa Cruz e Lincoln de Santa Cruz Oliveira Filho, por ter me ensinado como a educação é importante não só no caráter profissional como também no ético. Eu não teria chegado até aqui sem seus conselhos e palavras motivantes. Também as minhas irmãs, Nara e Mariana, por sempre apoiarem as minhas decisões durante a graduação.

À minha companheira, namorada e amiga, Flávia Vasconcelos, por ter me ajudado diretamente não só com este trabalho como também durante toda a graduação. Por ter sido uma das grandes responsáveis pelo meu amadurecimento pessoal e ter dividido sonhos e frustrações durante a graduação.

Ao Professor Carmelo J. A. Bastos-Filho, por ter me orientado com excelência não só durante este trabalho de conclusão de curso, como também durante todos os anos de Iniciação Científica. E aos demais professores e membros da direção da Universidade de Pernambuco que sempre buscaram trazer boas oportunidades para minha carreira.

Aos meus companheiros de classe, os quais contribuíram diretamente com o que me tornei hoje. Agradeço pelas contribuições acadêmicas, pelas longas tardes e noites de estudo coletivo, pela comemoração de todo objetivo alcançado e pelos conselhos e suporte nas inevitáveis derrotas que ocorreram durante a graduação. Agradeço também pelas idas a praia, as partidas de futebol e toda conversa jogada fora. Todo esse companheirismo foi fundamental para minha formação profissional e pessoal.

"(...) Embora ninguém possa voltar atrás e fazer um novo começo,
qualquer um pode começar agora e fazer um novo fim."

Chico Xavier

Sumário

Lista de Figuras ix

Lista de Tabelas xi

Lista de Algoritmos xi

Lista de Abreviaturas e Siglas xii

1 Introdução 1

1.1 Motivação e Caracterização do Problema 1

1.2 Hipóteses e Objetivos 3

1.3 Organização do Documento 4

2 Detectores de Humanos em Imagens Digitais 5

2.1 Fundamentos e conceitos básicos 5

2.1.1 Geração de janelas candidatas 6

2.1.2 Extração de características 7

2.1.3 Classificação binária 8

2.2 *Aggregated channel features (ACF)* 9

2.2.1 Pirâmide Rápida de Características 10

2.2.2 Conjunto de Características 13

2.2.3 Classificador 16

3 Caracteríticas Indicadoras de Movimento 18

3.1 Movimento em Imagens Digitais 18

3.2 Fluxo Óptico 19
3.3 Extratores de Características Indicadoras de Movimento 22

4 Uma Nova Extensão do ACF 23

4.1 Problemas Encontrados no ACF 23
4.2 Alterações propostas 25
4.3 Novo Conjunto de Características 26
4.3.1 Entropia do Histograma da Orientação do Gradiente (EHOG) 26
4.3.2 *Motion Boundary Histogram* (MBH) 27
4.3.3 *Internal Motion Histogram Central Difference* (IMHcd) 28
4.3.4 *Weak Stabilized Temporal Difference* (WSTD) 30

5 Resultados e Discussão 31

5.1 Arranjo Experimental 31
5.1.1 *Caltech Pedestrian Data Set* 31
5.1.2 Metodologia de Avaliação 33
5.1.3 Métricas 34
5.2 Parâmetros de Simulação 36
5.3 Experimentos 38
5.3.1 Análise da Performance dos Extratores de Características Indicadoras de Movimento 38
5.3.2 Análise das Melhorias Alcançadas pela Extensão Proposta 40

6 Conclusões e Trabalhos Futuros 46

6.1 Contribuições e Conclusões 46
6.2 Trabalhos Futuros 47

Referências 49

Lista de Figuras

1 Exemplo de detecção perfeita em uma imagem do banco de dados descrito em [1]. Os quadrados verdes são as posições da imagem onde há a presença de um ser humano. 2

2 Ilustração das subtarefas que geralmente, norteiam os detectores de humanos em imagens digitais. 6

3 Exemplos de máscaras para extração de características *Haar Like* 8

4 Ilustração do *pipeline* do ACF extraída de [2]. 10

5 Comparação entre a abordagem tradicional e a abordagem proposta pelo ACF. Imagem extraída de [2]. **Na parte superior:** Na abordagem padrão o canal de características é calculado diretamente sobre a imagem redimensionada para escala desejada. **Na parte inferior:** Na abordagem usada pelo ACF, o canal de características é calculado diretamente na imagem original e aproximado para a escala desejada utilizando a equação 2.1. 12

6 Canais de características LUV. A figura 6a é a imagem utilizada para extração dessas características. As figuras 6b, 6c e 6d são respectivamente as três componentes L, U e V. 14

7 Ilustração das etapas do canal de características magnitude do gradiente normalizado. 15

8 Ilustração das seis camadas do canal histograma da orientação do gradiente. . 16

9 Resultado da aplicação do ACF. 24

10 Exemplos de alarmes falsos produzidos pelo ACF original. 24

11 Visualização do canal de características EHOG. A imagem 11b é o canal EHOG da imagem 11a. 27

12 Visualização do canal de características MBH. As imagens 12c e 12d são a magnitude do gradiente do fluxo óptico na direção x e y respectivamente. 12a e 12b são os *frames* utilizados para extração dessa caraterística. 28

13 Máscara utilizada para subtração do fluxo óptico dentro de um bloco de 3x3 células. As setas e os valores -1 e +1 mostram como ocorrem as subtrações. . 29

14 Visualização das seis camadas do canal de características IMHcd. 12a e 12b são os *frames* utilizados para extração dessa caraterística. 29

15 Visualização do canal de características WSTD. 12a e 12b são os *frames* utilizados para extração dessa caraterística. 30

16 Curva *miss-rate* x FPPI para o ACF configurado com cada uma das novas características propostas. 39

17 Box plot das médias logarítmicas do *miss-rate* em nove diferentes valores de FPPI log espaçados de 10^{-2} à 10^{0} para o ACF configurado com cada uma das novas características propostas. 40

18 Curva *miss-rate* x FPPI para o ACF original e o ACF com a inclusão da característica IMHcd. 41

19 Box plot das médias logarítmicas do *miss-rate* em nove diferentes valores de FPPI log espaçados de 10^{-2} à 10^{0} para o ACF original e a nova extensão proposta neste trabalhos. 42

20 Box plot dos valores de FPPI para *miss-rate* igual a 30% 43

21 Box plot dos valores de FPPI para *miss-rate* igual a 20% 44

22 Box plot dos valores de FPPI para *miss-rate* igual a 10% 44

23 A imagem 23a exibe as detecções do ACF original e a imagem 23b exibe as detecções da nova extensão (ACF + IMHcd). 45

Lista de Tabelas

1 Matriz de Confusão 36

2 Parâmetros do banco de dados *Caltech Pedestrian Data Set* 37

3 Parâmetros do ACF. 37

4 Parâmetros do EHOG 37

5 Parâmetros do MBH e IMHcd. 38

6 Parâmetros do WSTD. 38

Lista de Algoritmos

1 Pseudocódigo do Classificador utilizado pelo ACF. 16

2 Pseudocódigo do algoritmo Lucas-Kanade. 22

Lista de Abreviaturas e Siglas

ACF – *Aggregated Channel Features*

AdaBoost – *Adaptative Boosting*

CSS – *Color Self-Similarity*

EHOG – *Entropia do Histograma de Orientação do Gradiente*

FPPI – *False positives per images*

FPS – *Frames* por segundo

HOG – *Histogram of Oriented Gradients*

IMHCD – *Internal Motion Central Difference*

LBP – *Local Binary Pattern*

MBH – *Motion Boundary Histogram*

MR – *Miss-rate*

SVM – *Support Vector Machine*

WSTD – *Weak Stabilized Temporal Difference*

1 *Introdução*

"Se eu vi mais longe,
foi por estar de pé no ombro de gigantes."
– Isaac Newton.

Neste trabalho de conclusão de curso, é desenvolvida uma extensão de um detector de humanos em imagens digitais através da inclusão de características indicadoras de movimento. A importância do padrão de movimento intrínseco ao ser humano é o fator que guiou a escolha de utilizar características indicadoras de movimento para detecção de humanos.

Este capítulo apresenta a introdução desta monografia, e está organizado em 3 Seções. Na Seção 1.1, é apresentada a motivação para a realização deste trabalho bem como o problema abordado pelo mesmo. Em seguida, na Seção 1.2 são apresentados os objetivos gerais e específicos, assim como a hipótese para solução do problema proposto. Por fim, na Seção 1.3, é descrita a estrutura do restante da monografia.

1.1 Motivação e Caracterização do Problema

Um dos principais objetivos da engenharia moderna é possibilitar a interação entre as máquinas e as pessoas ao seu redor. Para tal interação, a detecção de humanos é uma tarefa recorrente e fundamental em várias aplicações como robôs autônomos [3], proteção de pedestres [4], assistência hospitalar [5], segurança [6], entre outras. Mais especificamente, podemos definir a detecção de humanos em imagens digitais como: "Dado uma imagem ou sequencia de imagens, onde estão as pessoas na imagem ?" A figura 1 é um exemplo do resultado ideal da aplicação de um detector. Portanto, esse é o objetivo final na detecção de humanos em imagens digitais.

Detectar humanos em imagens é uma tarefa desafiadora dado que a aparência humana pode variar amplamente. Variações de dimensão (altura e biótipo), cor e textura (roupas e pele), *background*, Iluminação, perspectiva e oclusão são fatores que dificultam a correta

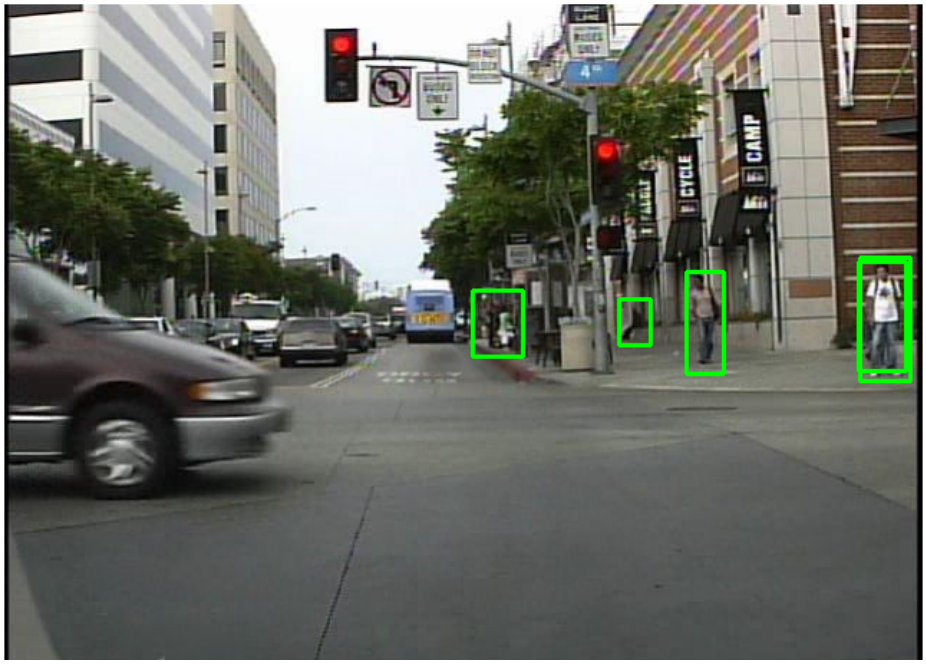


Figura 1: Exemplo de detecção perfeita em uma imagem do banco de dados descrito em [1]. Os quadrados verdes são as posições da imagem onde há a presença de um ser humano.

caracterização da aparência humana devido a vasta gama de possibilidades. Além disso, as aplicações destes algoritmos exigem execução em tempo real, o que gera mais um desafio que é a restrição do tempo de processamento do algoritmo de detecção.

Entretanto, inúmeros detectores de humanos em imagens digitais já foram desenvolvidos. Papageorgiou e Poggio propuseram um dos primeiros detectores de humanos em imagens monocular utilizando uma abordagem baseada em *sliding window*, características *Haar wavelets* e *Support Vector Machine (SVM)* como descrito em [7]. Dalal e Triggs propuseram a característica de forma baseada em gradiente chamada *Histogram of Oriented Gradients (HOG)* como descrito em [8]. A HOG alcançou melhores resultados que características baseadas em intensidade e se tornou um característica muito comum em todos os detectores modernos.

Detectores utilizando outras características indicadoras de forma foram propostos em [9, 10], no entanto não se mostraram melhores que a HOG individualmente. Então, conjuntos ricos e compostos por diferentes características passaram a ser utilizados e estudados. Características de cor como *Color Self-Similarity (CSS)* [11] e de textura como *Local Binary Pattern (LBP)* [12] entre outras também foram combinadas. Em [13], Dollar et al. propuseram um *framework* unificado para integrar diferentes características que vinham sendo desenvolvidas e utilizadas .

No entanto, incorporar características de movimento tem se mostrado raro e desafiador dado a não estaticidade da câmera. As características de movimento alcançam bons desempenhos em cenários com câmera estática como descrito em [14], mas são de difícil integração em cenários com câmeras móveis. Porém, instintivamente é inegável que características de movimento são informações extremamente úteis para identificação de humanos. Os movimentos das articulações e de pernas e braços do ser humano são um padrão muito característico do ser humano e podem ser boas fontes de informações para detecção de humanos.

Com a diversificação e o aumento do numero de características foi necessário o desenvolvimento de mecanismos que acelerassem a extração das características. Em [15] e [2], foi demostrado que é possível fazer o cálculo exato das características em algumas escalas e aproximar o seu valor em escalas próximas, assim minimizando o tempo de processamento e principalmente o tempo de detecção. A redução do tempo de detecção também foi contemplada por abordagens que usam classificadores em cascata para agilizar a classificação.

Porém, como descrito em [1], os detectores considerados o estado da arte em detectores de humanos em imagens digitais apresentam considerável taxa de falso positivos (e.g. pedaço de background erroneamente classificado como humano) que consistem em alarmes falsos. Portanto, este trabalho de conclusão de curso tem como objetivo desenvolver um detector que consiste em uma extensão de um detector de humanos em imagens digitais existente na literatura. O detector desenvolvido utilizará características indicadoras de movimento em conjunto com as características já altamente difundidas na literatura. Acredita-se que com a extração de características de movimento será possível reduzir a quantidade de alarmes falsos e aumentar a precisão de detecção com a mínima degradação da velocidade de detecção.

1.2 Hipóteses e Objetivos

Existem diversos detectores de humanos em imagens digitais utilizando os mais diversos extratores de características. No entanto, eles ainda apresentam considerável taxa de falso positivos (alarmes falsos) que podem inviabilizar certas aplicações. Portanto, a inclusão de características indicadoras de movimento pode amenizar tal problema, visto que o ser humano tem um padrão de movimento bem característico.

Devido a esta hipótese este trabalho de conclusão de curso tem como objetivo desenvolver uma extensão do detector de humanos em imagens digitais *Aggregated Channel Features (ACF)* descrito em [2]. Esse novo detector deve ser capaz de reduzir a taxa de falso positivos, melhorar a precisão de detecção e não degradar o tempo de detecção de seu detector

base. A extensão proposta utilizará características indicadoras de movimento calculadas sobre sequências de imagens. A extensão desenvolvida será testada no desafiador banco de imagens *Caltech Pedestrian Dataset* descrito em [1] e comparada com o ACF original, para comprovar as melhorias alcançadas.

1.3 Organização do Documento

Este trabalho está organizado em 6 capítulos. No Capítulo 2 são abordados aspectos relativos à detectores de humanos em imagens digitais e a descrição do ACF. Em seguida, no Capítulo 3, é realizada uma explanação de extratores de características indicadoras de movimento. No Capítulo 4 é apresentada a contribuição deste trabalho de conclusão de curso: A extensão do *Aggregated Channel Features* através da utilização de características indicadoras de movimento. Em seguida, no Capítulo 5, os experimentos e resultados são apresentados. Por fim, no Capítulo 6, são discutidas as principais conclusões deste trabalho, como também propostas para trabalhos futuros.

2 *Detectores de Humanos em Imagens Digitais*

"Os problemas significativos que enfrentamos não podem ser resolvidos no mesmo nível de pensamento em que estávamos quando os criamos."

– Albert Einstein.

Neste capítulo são apresentados os conceitos básicos dos detectores de humanos em imagens digitais como: geração de janelas candidatas, conjunto de características, classificação, entre outros. Também será descrito detalhadamente o detector ACF (*Aggregated Channel Features*) e suas propriedades que determinaram a sua escolha como detector base para extensão proposta neste trabalho de conclusão de curso.

Na seção 2.1 serão descritos os conceitos básicos de funcionamento dos detectores de humanos em imagens digitais. Em seguida, na seção 2.2 será descrito o detector ACF.

2.1 Fundamentos e conceitos básicos

Como descrito na seção 1.1, um detector de humanos em imagens digitais recebe uma imagem e retorna a localização precisa de cada ser humano presente na imagem. Essa localização é normalmente estruturada como uma tupla de quatro elementos: distâncias à lateral esquerda e a parte superior da imagem, como também, largura e altura do quadrado que cobre exatamente o corpo do ser humano. Essas quatro medidas são especificadas em *pixels*. Essa estrutura de dados é conhecida como *Bounding Box*. Portanto, o objetivo principal dos detectores de humanos em imagens digitais é retornar os *Bounding Boxes* dos humanos, dado uma imagem ou uma sequência de imagens.

Para alcançar tal objetivo, os detectores não podem utilizar abordagens tradicionais devido os desafios citados na seção 1.1. Portanto várias técnicas oriundas da inteligência artificial, aprendizagem de maquina além de visão computacional são aplicadas. Desta forma, os detec-

tores funcionam em duas etapas: Treinamento e Detecção. Devido a complexidade da tarefa, essas duas etapas geralmente usam a abordagem "dividir para conquistar" que consiste em quebrar a tarefa maior em uma sequencia de tarefas menores. Contudo, tanto o treinamento como a detecção podem ser divididos nas seguintes etapas: Geração de janelas candidatas, extração de características e classificação binária. A figura 2 mostra cada uma dessas subtarefas, assim como suas entradas e saídas.

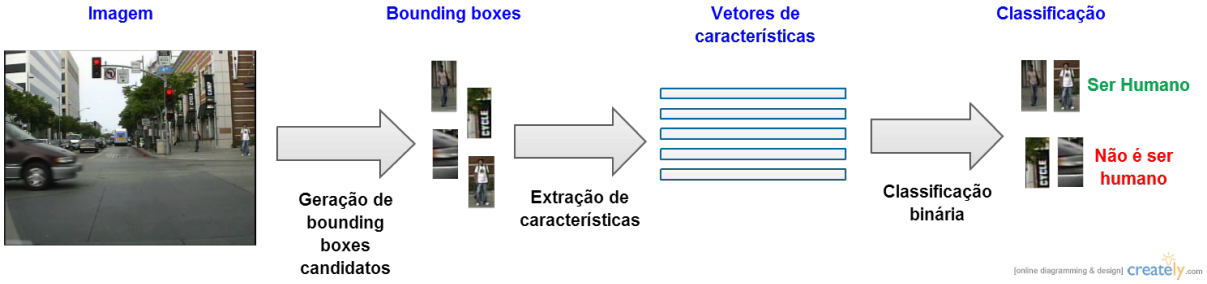


Figura 2: Ilustração das subtarefas que geralmente, norteiam os detectores de humanos em imagens digitais.

2.1.1 Geração de janelas candidatas

A geração de janelas candidatas consiste em repartir a imagem em quadrados que podem ou não conter seres humanos. Usualmente, a imagem é repartida em quadrados de tamanho fixo (normalmente, um parâmetro do detector). Para executar esta tarefa, técnicas de processamento de imagens digitais como segmentação, pontos de interesse e uma associação de *sliding window* com imagem pirâmide são utilizadas.

A segmentação consiste em separar os objetos do *background* de uma imagem. Técnicas como limiarização, clusterização de *pixels* entre outras podem ser aplicadas como descrito em [16]. Utilizando essa abordagem os objetos gerados são passados para a etapa de extração de características. Desta forma, a quantidade de janelas candidatas geradas é menor em comparação com outras abordagens. Isso reduzirá o tempo necessário para extração de características, já que diminuirá a quantidade de aplicações dos extratores de características. Porém, as ineficiências do algoritmo de segmentação prejudicam o desempenho do detector e podem elevar o tempo de processamento, pois os algoritmos de segmentação precisos são custosos. O detector descrito em [17] utiliza essa abordagem.

Semelhante à segmentação, a abordagem que utiliza pontos de interesse consiste em utilizar ferramentas como o SURF, descrito em [18], para localizar pontos de referência chamados de pontos de interesse. Em seguida é recortado uma área pré-definida ao redor de cada um desses pontos de interesse. Essas regiões recortadas são passadas para a etapa de extração

de características. Como na segmentação, essa bordagem gera menos janelas candidatas, mas costuma ser mais precisa que muitas técnicas de segmentação. Porém, as técnicas de localização de pontos de interesse são ainda mais custosas. O detector descrito em [19] utiliza essa abordagem.

Por fim, a abordagem de geração de janelas candidatas mais utilizada é a associação de *sliding window* com imagem pirâmide. Essa técnica se mostra mais promissora em imagens de baixa à media resolução que as descritas anteriormente, como descrito em [1]. Essa abordagem consiste em gerar imagens com diferentes resoluções, formando uma pirâmide, e executar o *sliding window* com janela de tamanho fixo em cada uma dessas imagens que foram redimensionadas. Perceba que a utilização da imagem pirâmide permite a detecção em várias escalas. Contudo, essa abordagem gera um exagerada quantidade de janelas candidatas para a etapa de extração de características, mas praticamente não introduz erro ao detector.

2.1.2 Extração de características

O processo de extração de características consiste em extrair e eficientemente representar informações contidas nas imagens. O principal objetivo deste processo é tornar a informação contida na imagem útil para tarefas subsequentes como a classificação. Portanto, nos detectores de humanos em imagens digitais, os extratores de características recebem as janelas candidatas e as transformam em vetores ou canais de características [13]. Essas características são as entradas para o processo de classificação.

Os detectores de humanos em imagens digitais mais modernos utilizam conjuntos de características heterogêneos, compostos por características que indicam diferentes informações contidas na imagem. As características comumente utilizadas são: *Haar Like*, *Color Self-Similarity*, *Local Binary Pattern* e *Histogram of Oriented Gradients*. Normalmente, essas características são utilizadas em conjunto para aumentar a precisão do detector.

Haar like: Consiste em uma forma rápida e eficiente de calcular diferenças de intensidade de *pixels* em regiões da imagem [20]. As máscaras exibidas na figura 3 são usadas para extração dessas características e cada máscara é um tipo diferente de característica *Haar like*. Em cada máscara, a soma dos *pixels* sobre a região preta é subtraída da soma dos *pixels* da região branca e essas operaçõe se repetem por toda imagem seguindo o algoritmo *sliding window*. Essas características são de baixo custo computacional devido ao uso de imagem integral para o cálculo eficiente das diferenças entre regiões. Elas também são usadas em várias outras aplicações de reconhecimento de padrões como detecção de face [20].

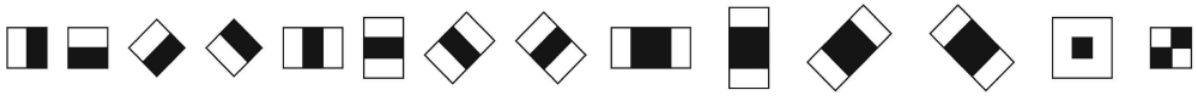


Figura 3: Exemplos de máscaras para extração de características *Haar Like*

Color Self-Similarity (CSS): Proposta em [11], consiste em codificar similaridades de cores em regiões diferentes da imagem. A abordagem tradicional para calcular essa característica é dividir a imagem em blocos sem sobreposição. Então, em cada bloco, histogramas de cor no modelo *HSV* são construídos com interpolação. Por fim, similaridades são calculadas através da intersecção desses histogramas onde cada intersecção produz um vetor de características. As *CSS* são características de cor e são boas para identificar regiões continuas na imagem.

Local Binary Pattern (LBP): Proposta em [12], é uma característica de textura que foi originalmente proposta para detecção de face. *LBP* consiste em dividir a imagem em células não sobrepostas. Cada *pixel* nestas células são comparados com seus 8-vizinhos e uma *string* binária é gerada de acordo com essa comparação. Posteriormente, essa *string* binária é transformada em um número inteiro e para cada célula é criada um histograma de acordo com a frequência desses números. Por fim, os histogramas são normalizados e concatenados para formar o vetor de características. Essas características são utilizadas no detector proposto em [21].

Histogram of Oriented Gradients (HOG): Proposta em [8], é uma característica indicadora de forma e é a base de muitos detectores modernos. A extração das características HOG consiste na a aplicação do gradiente, na divisão da imagem em células sem sobreposição e na construção de histogramas da orientação dos vetores do gradiente ponderado pela magnitude do mesmo para cada célula. Por fim, as células são agrupadas em blocos com superposição onde é executada uma normalização local. Outras variações dessa característica são encontradas na literatura. Individualmente, a HOG obtém os melhores resultados de detecção de humanos em imagens digitais como descrito em [1].

2.1.3 Classificação binária

Classificação é uma das tarefas mais frequentes no campo de Visão Computacional. Ela está presente em praticamente qualquer aplicação que consiste na identificação de entidades como detecção de humanos, detecção de face entre outras. Ela consiste em atribuir *labels*, valores inteiros definidos pelo número finito de classes distintas, para as amostras a serem classificadas. Já a classificação binária, é uma subclasse da classificação onde só existe duas

classes distintas.

Na detecção de humanos em imagens digitais, geralmente, a tarefa de classificação ocorre após a extração de características. Ela consiste em utilizar os vetores de características gerados, para treinar o classificador ou para classificar a respectiva janela candidata em "humano" ou "não humano". Os classificadores comumente utilizados para tal tarefa são o SVM ou AdaBoost devido suas características de aprender em espaço multidimensionais. As Redes neurais artificiais também são opções válidas mas não são frequentemente utilizadas pelos detectores de humanos em imagens digitais devido à alta dimensionalidade apresentada nessa tarefa.

Support Vector Machine(SVM) é uma técnica de aprendizagem de máquina que visa separar as amostras em um espaço hiper dimensional usando um hiperplano que tem a maior margem possível [22]. O hiperplano traçado pela SVM depende do kernel utilizado. Existe kernel linear e não linear. Porem, o kernel linear é mais utilizado na detecção de humanos devido a alta dimensionalidade e pouca disponibilidade de dados das aplicações de detecção de humanos em imagens digitais. Nessa abordagem o conjunto de caraterísticas é considerado um espaço onde um hiperplano linear é suficiente para classificação perfeita. O SVM é utilizado em vários detectores como em [23].

Adaboost é um técnica de combinação de classificadores fracos para formar um classificador forte. Existem várias opções de classificadores fracos e inúmeras formas de combinações. Por exemplo, utilizar várias árvores de decisão com diferentes subconjuntos de características é uma forma comum de utilizar tal técnica. Outra forma de utilizar essa técnica está descrita em [24]. Contudo, o *adaboost* é muito utilizado na detecção de humanos devido a sua capacidade de aprender em espaços multidimensionais, de executar implícita seleção de características e ser menos suscetível a *overfitting*. O detector descrito em [2] usa essa abordagem.

2.2 *Aggregated channel features (ACF)*

O *Aggregated Channel Features (ACF)* é um detector de humanos em imagens digitais descrito em [2]. O ACF foi desenvolvido com o intuito de utilizar conjuntos heterogêneos de características de forma eficiente. Outros detectores utilizavam esses conjuntos, mas apresentavam alto custo computacional a medida que o conjunto de característica era enriquecido. Consequentemente, foi criado um *trade-off* entre a precisão de detecção gerado por conjuntos de características elaboradas e diversas e a velocidade de detecção. Surpreendentemente, o ACF propôs um mecanismo de aproximação do cálculo de características em densas escalas.

Esse mecanismo permitiu ao ACF alcançar a taxa de detecção de 30 fps com ótima precisão de detecção.

O ACF pode ser explicado utilizando o *framework* descrito na seção 2.1 com a adição de algumas particularidades. As janelas candidatas são geradas utilizando a abordagem *sliding window* com imagem pirâmide. Porém, a pirâmide não é calculada para todas as escalas e o mecanismo de aproximação de características é utilizado, conferindo maior velocidade de detecção. O conjunto de característica é heterogêneo e contém informações de cores e formas dispostas no *framework* decrito em [13]. Por fim, para a classificação é utilizado uma técnica derivada do *AdaBoost* para combinar árvores de decisão.

A imagem da figura 4 descreve em detalhes como funciona o ACF. Dada um imagem de entrada ou um *patch* extraído da imagem, canais de características são extraídos pela função Ω. Esse mecanismo será explicado em detalhes na seção 2.2.2. Então, cada um desses canais são divididos em blocos de tamanho fixo (parâmetro do ACF) e são redimensionados para que todos os canais de características tenham as mesmas dimensões. Então esses canais são agrupados em um único vetor. Por fim, esse vetor é utilizado para treinar árvores de decisão combinadas por uma abordagem AdaBoost que será detalhada na seção 2.2.3.

Figura 4: Ilustração do *pipeline* do ACF extraída de [2].

O código do ACF é publicamente distribuído sobre a licença BSD simplificada e está disponível online. O código foi desenvolvido na linguagem e ferramenta de simulação Matlab e está inserido na *toolbox* desenvolvido por Piotr Dollar e disponível em [25] sobre a mesma licença. Porém, esse software só opera com características calculadas sobre uma imagem estática. Portanto, foi necessário profundas mudanças com o intuito da utilização do ACF neste trabalho de conclusão de curso. Essas mudanças visam a extração de características de movimento as quais são calculadas a partir de sequências de imagens. Todas as mudanças propostas serão descritas no Capítulo 4.

2.2.1 Pirâmide Rápida de Características

A velocidade de detecção e a precisão de um detector de humanos em imagens digitais são vitais para a aplicação do mesmo em problemas reais. No entanto, esses dois fatores

formam um *trade-off*. Por exemplo, um detector com boa precisão deve utilizar um elevado e diversificado número de extratores de características. No entanto, a aplicação de diversos extratores de características em pedaços da imagem em várias escalas torna o detector lento e talvez inviabilize sua aplicação. Portanto, é preciso desenvolver um mecanismo que permita utilizar vastos conjuntos de características sem degradar a velocidade de detecção.

Como descrito em [2], existe uma relação analítica e empiricamente testada entre características calculadas na imagem original e em suas imagens redimensionadas. Essa descoberta sugere a possibilidade de aproximar o cálculo de características em imagens redimensionadas a partir do valor dessa característica na imagem original. Portanto, essas conclusões são a base da pirâmide rápida de características que é a maior contribuição do ACF na área de detecção de humanos em imagens digitais.

O ACF utiliza canais de características $C = \Omega(I)$ que são calculados sobre a imagem I. Os *pixels* em C são características calculadas a partir dos *pixels* ou *regiões* correspondente na imagem original I, assim mantendo o *layout* da imagem. Porém, C pode ter dimensões diferentes de I e varias camadas. Segundo [13], muitos tipos de características podem ser escritas dessa forma, como gradientes, histogramas entre outras.

A abordagem tradicional é calcular o canal de características na escala s diretamente sobre a imagem redimensionada para escala s, ignorando a informação contida no mesmo canal de característica calculado sobre a imagem original. O ACF propõem a seguinte aproximação:

$$C_s = R(C, s) \cdot s^{-\lambda_\Omega}, \tag{2.1}$$

onde C_s é o canal de característica na escala s e C é o canal de característica da imagem original. $R(C, s)$ significa redimensionar o canal de característica calculado com a imagem original para a escala s. s é a escala na qual queremos calcular as características e λ_Ω é o parâmetro de aproximação que deve ser estimado empiricamente para cada tipo de característica. A figura 5 demonstra a compração entre essas abordagens. Essa relação não é apenas valida para a imagem original mas também para qualquer par de janelas w_s e w em I_s e I, respectivamente.

Pirâmide de características é uma representação multi escala de uma imagem onde os canais são computados para cada escala s utilizando a imagem redimensionada. Normalmente, s incia em 1 e possui 4 a 12 escalas por oitavo (um oitavo é o intervalo entre uma escala e outra escala com metade ou o dobro do seu valor). Já a pirâmide rápida de características é uma forma eficiente de construir a pirâmide de características utilizado o método de aproximação baseado na equação 2.1.

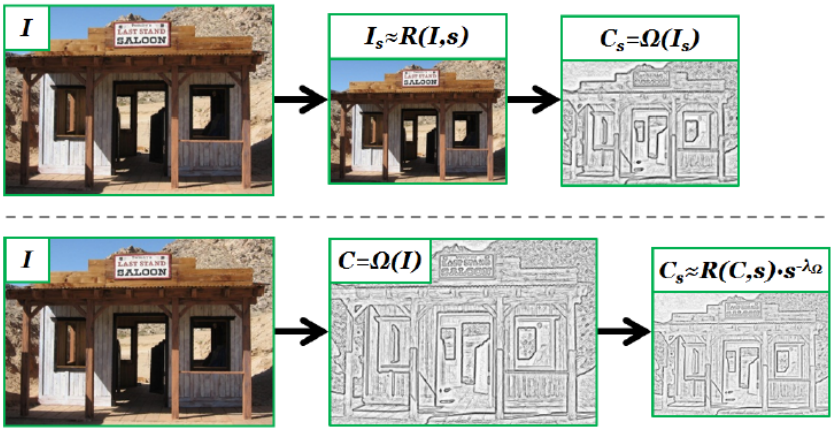


Figura 5: Comparação entre a abordagem tradicional e a abordagem proposta pelo ACF. Imagem extraída de [2]. **Na parte superior:** Na abordagem padrão o canal de características é calculado diretamente sobre a imagem redimensionada para escala desejada. **Na parte inferior:** Na abordagem usada pelo ACF, o canal de características é calculado diretamente na imagem original e aproximado para a escala desejada utilizando a equação 2.1.

Na piramide rápida de características os canais de características são calculados diretamente uma vez por oitavo. Mas, nas escalas intermediárias esses canais são calculados de acordo com a equação 2.2, que é inspirada na equação 2.1.

$$C_s = R(C_{s'}, s/s') \cdot (s/s')^{-\lambda_\Omega}, \tag{2.2}$$

onde s é a escala intermediária e s' é a escala do oitavo mais próximo. Então, os canais de características C_s das escalas intermediarias s são o redimensionamento $R(C_{s'}, s/s')$ do canal calculado diretamente com a imagem redimensionada para o oitavo mais próximo s' multiplicado pelo fator de aproximação $(s/s')^{-\lambda_\Omega}$.

Contudo, o parâmetro λ_Ω deve ser estimado empiricamente para cada canal de característica utilizado. Utilizar um equação analítica para tal próposito, apesar de possível, pode ser complexo devido a sofisticação e diversidade das características existentes. Então, o ACF utiliza uma um metodologia empírica de aproximação.

As características utilizadas para o treinamento do classificador no *framework* ACF é definida pela equação 2.3 e calculada sobre o canal C_s:

$$f_\Omega(I_s) = \frac{1}{h_s w_s k} \sum_{ijk} C_s(i,j,k), \tag{2.3}$$

onde h_s, w_s são as dimensões do canal C_s que possui k camadas.

Ruderman e Bialek em [26], estudaram como estatísticas calculadas sobre a imagem se

comportam em função da sua escala. Eles chegaram a equação 2.4 que norteia essa relação e é chamada de *power law*.

$$E[\phi(I_{s_1})]/E[\phi(I_{s_2})] = (s_1/s_2)^{-\lambda_\phi} \tag{2.4}$$

Colocando no contexto do ACF, esse resultado pode ser descrito pela equação 2.5, substituindo a estatística ϕ pela característica f_Ω e a constante λ_ϕ pela constante λ_Ω.

$$f_{\Omega(I_{s_1})}/f_{\Omega(I_{s_2})} = (s_1/s_2)^{-\lambda_\Omega} + \varepsilon \tag{2.5}$$

Então, estimando λ_Ω para uma dada característica Ω, calculamos:

$$\mu_s = \frac{1}{N}\sum_{i=1}^{N} f_\Omega(I_s^i)/f_\Omega(I_1^i), \tag{2.6}$$

para N imagens I^i e múltiplos valores de s, onde I_s^i é obtida pelo redimensionamento de $I_1^i = I^i$ para escala s. Essas N imagens são exemplos de humanos e não humanos. De acordo com 2.5 e 2.6, $\mu_s = s^{-\lambda_\Omega} + E[\varepsilon]$. Então, nosso objetivo é escolher o valor de λ_Ω de forma que $E[\varepsilon] \approx 0$ para que a equação 2.5 seja válida.

Então, a equação 2.6 é calculada para cada imagem por três oitavos com oito escalas por oitavo, totalizando 24 medições por imagem, e plotada contra o parâmetro s. Segundo [2], em todos os casos μ_s segue a equação 2.4, com todas as medições sobre a linha do gráfico log-log $s \times \mu_s$.

Portanto, μ_s pode ser descrito como $\mu_s = a_\Omega s^{-\lambda_\Omega}$. Por fim, o método dos mínimos quadrados é utilizado para calcular os parâmetros a_Ω e λ_Ω da curva descrita por $log_2(\mu_s) = log_2(a_\Omega) - \lambda_\Omega log_2(s)$ para as 24 medições de μ_s por imagem calculadas anteriormente. Esse mecanismo é repetido para cada tipo de canal de características Ω.

Portanto, o mecanismo descrito reduz a quantidade de vezes que aplicamos os extratores de características, pois no ACF eles são utilizados uma vez por oitavo. Essa redução é substituída pela aproximação em escalas intermediárias aos oitavos. Esse mecanismo confere velocidade com baixa degradação da qualidade das características.

2.2.2 Conjunto de Características

No *framework* ACF as características são formadas por conjuntos de canais de características como descrito em [13] e explicado anteriormente. Mais especificamente, o conjunto de características do ACF padrão é composto por 10 canais de características: Magnitude

do gradiente normalizada (1 camada), Histograma da orientação do gradiente (6 camadas) e esquema de cores LUV (3 camadas). Onde, os dois primeiros são características de forma e o ultimo característica de cor.

Antes e depois do cálculo das características algumas operações são realizadas para melhorar a qualidade das características extraídas. Antes do cálculo dos canais a imagem é suavizada com um filtro cuja a mascara é [1 2 1]/4. Depois de tal suavização, os canais de características são calculados. Em seguida, eles são divididos em blocos de 4x4 *pixels* e os *pixels* dentro de cada bloco são somados (esse processamento é realizado pela aplicação da equação 2.3). Por fim, os canais são suavizados novamente com o filtro [1 2 1]/4.

Modelo de cor LUV: Essa é uma característica indicadora de cor e possui três camadas: L (luminância), U (verde-laranja) e V (azul-amarelo). LUV é um espaço de cor resultante da transformação do modelo de cor CIE XYZ através de uma normalização que provê "percepção uniforme". Isso significa que duas cores igualmente distantes no espaço de cor têm percepções igualmente distantes. Nesse modelo de cor, todos os canais possuem valores entre 0 e 1. A figura 6 mostra o resultado da extração dessas características.

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

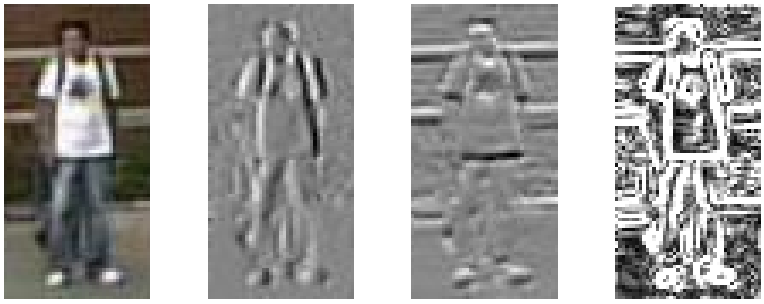
Figura 6: Canais de características LUV. A figura 6a é a imagem utilizada para extração dessas características. As figuras 6b, 6c e 6d são respectivamente as três componentes L, U e V.

Magnitude do gradiente normalizada: Essa é uma característica de forma que possui um único canal. Essa característica consiste em calcular o gradiente para cada pixel na direção horizontal e vertical utilizando a equação 2.7. Esse processo gera um vetor na direção x e outro na direção y que representam a variação dos *pixels* e capturam as bordas das formas na imagem. Então, a magnitude é calculada utilizando a formula 2.8 para cada *pixel*. Por fim, é executada uma normalização para evitar valores destoantes que possam confundir o classificador que será usado na classificação da imagem. No ACF essa característica é calculada sobre a componente luminância da imagem no modelo de cor LUV.

$$\frac{\partial I(x,y)}{\partial x} = \frac{I(x+1,y) - I(x-1,y)}{2}, \quad \frac{\partial I(x,y)}{\partial y} = \frac{I(x,y+1) - I(x,y-1)}{2} \tag{2.7}$$

$$M(x,y) = \sqrt{\left(\frac{\partial I(x,y)}{\partial x}\right)^2 + \left(\frac{\partial I(x,y)}{\partial y}\right)^2} \tag{2.8}$$

A figura 7 mostra o resultado de cada etapa do cálculo do canal de característica magnitude do gradiente normalizada. A figura 7a é a imagem de um humano na qual será extraída essa característica. A figura 7b e 7c são respectivamente as imagens de visualização do gradiente na direção horizontal e vertical. Por fim, a figura 7d exibe o *layout* final dessa característica extraída sobre a imagem 7a.



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

Figura 7: Ilustração das etapas do canal de características magnitude do gradiente normalizado.

Histograma da orientação do gradiente: Essa é uma característica indicadora de forma que possui 6 camadas. Essa característica é descrita em [13] e consiste em uma adaptação das característica HOG proposta em [8]. O primeiro passo no cálculo dessa característica é calcular os ângulos e a magnitude do gradiente em cada *pixel* da imagem. Depois, para cada célula de tamanho $c \times c$ *pixels* é construído um histograma de 6 bins, onde os bins são intervalos de ângulos e a frequência é ponderada pela magnitude. A magnitude do gradiente é dividida entre os bins mais próximos de sua orientação proporcionalmente a distância a cada bin. Por fim, esse canal consiste no mapa de células $c \times c$ *pixels* onde cada célula possui 6 camadas. Cada camada em uma dada célula é o valor da frequência em um determinado bin de orientação. A figura 8 exibe a visualização das seis camadas desse canal de característica aplicado na figura 7a. No ACF essa característica é calculada sobre a componente luminância da imagem no modelo de cor LUV.

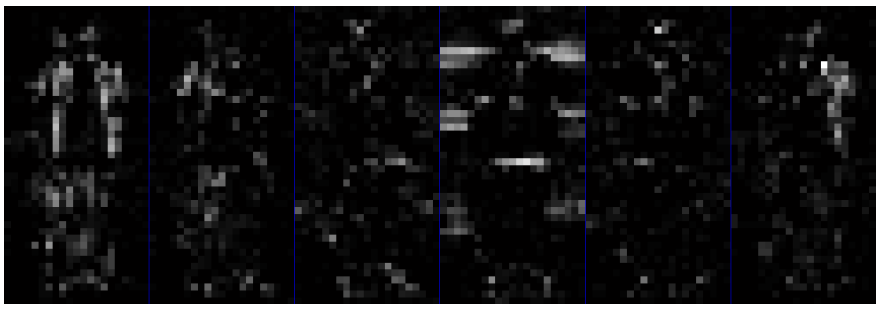


Figura 8: Ilustração das seis camadas do canal histograma da orientação do gradiente.

2.2.3 Classificador

Para a classificação de humanos o ACF utiliza uma classificador que é formado por um conjunto de classificadores fracos definidos por árvores de decisão. Essas árvores de decisão são combinadas através da técnica *AdaBoost*, assim criando um novo classificador mais preciso. Esse mecanismo é conhecido como *Boosting trees*. A principal ideia do *Boosting trees* é treinar uma sequência de árvores de decisões simples, onde cada árvore sucessiva é treinada para classificar corretamente as amostras erroneamente classificadas pelas árvores de decisão predecessoras. Contudo, no ACF esse classificador final é otimizado utilizando as ideias propostas em [27].

O pseudocódigo do algoritmo 1 explica como ocorre o treinamento do classificador utilizado pelo ACF, dado N amostras de treinamento $(x_1, y_1), ..., (x_N, y_N)$, onde $x_i \in R^n$ e $y_i \in \{-1, 1\}$. O classificador retornado é $H(X) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$, onde $h_t(x)$ é um classificador fraco, α_t sua importância no processo geral de classificação e T é o numero de iterações do *AdaBoost*.

Algoritmo 1: Pseudocódigo do Classificador utilizado pelo ACF.

1 Inicialize os pesos $w_i = 1/N$;
2 **para** $t = 1, ..., T$ **faça**
3 Treine o classificador fraco h_t;
4 Calcule o erro $\epsilon_t = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \cdot 1\{h_t(x_i) \neq y_i\}$;
5 Calcule $\alpha_t = \frac{1}{2} ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$;
6 Atualize os pesos $w_i = \frac{w_i exp(-\alpha_t y_i h_t(x_i))}{\sum_{j=1}^{N} w_j exp(-\alpha_t y_j h_t(x_j))}$
7 Retorne o classificador final $H(X) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$;

No pseudocódigo do algoritmo 1, w_i é o peso atribuído a amostra i, esse peso é atualizado de acordo com o acerto das árvores que foram treinadas anteriormente. As funções $sign(X)$

e $1\{E\}$ são definidas nas equações 2.9.

$$sign(X) = \begin{cases} -1, & \text{se } X < 0\ , \\ +1, & \text{se } X \geq 0\ . \end{cases} \qquad 1\{E\} = \begin{cases} 1, & \text{se E for verdadeiro, }, \\ 0, & \text{caso contrario.} \end{cases} \tag{2.9}$$

O classificador fraco no *framework* ACF são árvores binárias de decisão com profundidade dois. Cada nó dessas árvores dividem as amostras utilizando a regra de decisão definida pela equação 2.10. Onde k_j é o índice da característica $x\left[k_j\right]$ utilizada para dividir as amostras, τ_j é o limiar de divisão e $p_j \in \{-1, +1\}$ é a polaridade. Todos esse parâmetros são relativos ao o nó j e sua regra de decisão $h_j(X)$. Por fim, nos nós terminais, a classificação é realizada com a atribuição do label mais frequente na partição.

$$h_j(X) = p_j sign(x\left[k_j\right] - \tau_j) \tag{2.10}$$

A partir dessa classificação podemos calcular o erro de classificação da árvore pela equação 2.11. Então o treinamento da árvore deve minimizar o objetivo $\arg\min_{p,\tau,k} \epsilon^{(k)}$, ou seja, selecionar os parâmetros $\{p^*, \tau^*, k^*\}$ para cada nó j que resultam em menor erro $\epsilon^{(k)}$. Em [28, 29] são descritos procedimentos iterativos para formação dessas árvores através da minimização do objetivo proposto.

$$\epsilon^{(k)} = \frac{1}{\sum_{i=1}^{N} w_i} \left[\sum_{i=1}^{N} w_i \cdot 1\{y_i = +p_m^{(k)}\} + \sum_{i=1}^{N} w_i \cdot 1\{y_i = -p_m^{(k)}\} \right] \tag{2.11}$$

Portanto, o ACF utiliza 2048 árvores binárias de decisão com profundidade dois sobre 10 canais de características que são agrupados e somados em células de 4x4 *pixels* para cada canal, como descrito pela equação 2.3. Por exemplo, na análise de uma janela 128 x 64 *pixels* são geradas $\frac{128 \times 64 \times 10}{4 \times 4} = 5120$ características de treinamento. Esse mecanismo combina a simplicidade das árvores de decisão com a poder de aprendizado em espaços multidimensionais e a implícita seleção de características do *AdaBoost*.

3 Caracteríticas Indicadoras de Movimento

"Deve-se mais aprender em profundidade,
do que em largura."
– Quintiliano.

Neste capítulo é abordado o tópico extração de características indicadoras de movimento a partir de sequências de imagens digitais. Conceitos básicos desse tópico serão explicados: tipos de movimento, extratores, fluxo óptico entre outros conceitos. Também é explicado como essas características podem ajudar na detecção de humanos em imagens digitais.

Na Seção 3.1 são descritos os conceitos básicos de percepção de movimento em imagens digitais. Em seguida, na Seção 3.2 são descritos o conceito e técnicas para o cálculo de fluxo óptico. Por fim, na Seção 3.3 são citados extratores de características existentes na literatura.

3.1 Movimento em Imagens Digitais

Dada uma sequência de imagens é possível reconhecer quais entidades na imagem estão em movimento. Essas informações podem ser de grande valor para tarefas de detecção de objetos "vivos". Porém, essas informações são difíceis de serem extraídas e representadas, devido a dinamicidade da câmera e a grande quantidade de elementos em movimento entre outros fatores. Contudo, técnicas para extração, filtragem e representação de movimento em sequências de imagens já foram propostas. [30, 31, 32] são exemplos de técnicas para extração de características indicadoras de movimento.

Em vídeo ou sequência de imagens, existem basicamente três tipos de movimento: Movimento da câmera, movimento central do objeto e movimento de partes do objeto. O movimento da câmera consiste do movimento gerado pela movimentação da câmera ou por mudanças no *background*. O movimento central do objeto consiste na movimentação do cen-

troide do objeto, que pode ser representado pelo movimento do objeto como um todo. Já o movimento de partes do objeto é caracterizado pelo movimento das extremidades ou articulações do objeto em relação ao próprio objeto, como por exemplo o movimento das pernas do ser humano durante uma caminhada.

Nem todo tipo de movimento é útil para tarefas de detecção de objetos em imagens digitais. Por exemplo, o movimento de câmera não acrescenta informação para detecção de objetos e é uma grande fonte de ruído. Já o movimento central do objeto apesar de acrescentar informação, ele também gera ruído devido a grande quantidade de entidades capazes de executar tal movimento. Porém, o movimento de partes de objetos é uma boa fonte de informação para tarefas de detecção, visto que cada entidade possui um padrão particular de movimento. Portanto, os extratores de características indicadoras de movimento têm como objetivo filtrar certos tipos de movimentos, extrair informações oriundas desses movimentos, e representar essas informações em um formato útil para tarefas de detecção de objetos.

3.2 Fluxo Óptico

O fluxo óptico é uma boa ferramenta para extração de movimento contido em vídeo ou sequência de imagens. Ele consiste em representar a translação de objetos, estruturas ou *pixels* na imagem em relação ao tempo. Portanto, o objetivo da estimativa de fluxo óptico é calcular aproximações para o movimento de *pixels* em função do tempo. [33, 34] são técnicas para estimativa do fluxo óptico.

O fluxo óptico tenta representar o movimento dos elementos da imagem capturadas em instantes diferentes em cada posição da imagem. Considere $I(x, y, t)$ como a intensidade dos *pixels* ou regiões em função do espaço $\vec{x} = (x, y)^T$ e tempo t. Também considere $v = (v_x, v_y)^T$ como o vetor velocidade 2D. Então, assuma que a intensidade dos *pixels* são constantes de um *frame* para o outro como descrito na equação 3.1. Essa restrição é irreal, mas é útil para derivação e funciona bem na prática.

$$I(x, y, t) = I(x + v_x, y + v_y, t + 1) \tag{3.1}$$

Então, aproximamos o *frame* no instante seguinte pela série de Taylor de primeira ordem como mostrado na equação 3.2. Onde $I_x = \frac{\partial I(x,y,t)}{\partial x}$, $I_y = \frac{\partial I(x,y,t)}{\partial y}$ e $I_t = \frac{\partial I(x,y,t)}{\partial t}$ são respectivamente as derivadas parciais da imagem em relação a x, y e t. Observe que os

termos de alta ordem são desconsiderados.

$$I(x+v_x, y+v_y, t+1) \approx I(x,y,t) + v_x I_x(x,y,t) + v_y I_y(x,y,t) + I_t(x,y,t) \tag{3.2}$$

Então combinando 3.1 e 3.2 chegamos na equação 3.3, conhecida como *gradient constraint equation*. Perceba que estou usando as abreviações I_x, I_y e I_t para as derivadas parciais.

$$\begin{aligned} I(x,y,t) &= I(x,y,t) + v_x I_x + v_y I_y + I_t \\ v_x I_x + v_y I_y + I_t &= 0 \\ v_x I_x + v_y I_y &= -I_t \end{aligned} \tag{3.3}$$

Porém, essa equação não pode ser resolvida analiticamente já que existem duas variáveis desconhecidas e apenas uma equação. Então podemos estender a equação 3.3 para regiões contínuas na imagem assumindo que todos os *pixels* dessa região possuem o mesmo vetor velocidade $\vec{v}$. Então definimos o erro quadrático do vetor de velocidade para a janela w como na equação 3.4. Onde $g(\vec{x})$ é uma função ponderadora que distribui pesos para os *pixels* dentro da janela w. Normalmente, $g(\vec{x})$ é uma gaussiana que atribui maiores pesos para os *pixels* no centro da janela.

$$E(v_x, v_y) = \sum_{\vec{x}}^{w} g(\vec{x}) \left[v_x I_x + v_y I_y + I_t \right]^2 \tag{3.4}$$

Para encontrar os valores de v_x e v_y que minimize $E(v_x, v_y)$ podemos calcular derivada parcial de $E(v_x, v_y)$ em relação a v_x e v_y, e em seguida, igualar a zero:

$$\frac{\partial E(v_x, v_y)}{\partial v_x} = \sum_{\vec{x}}^{w} 2g(\vec{x}) \left[v_x I_x + v_y I_y + I_t \right] I_x = \sum_{\vec{x}}^{w} g(\vec{x}) \left[v_x I_x^2 + v_y I_y I_x + I_t I_x \right] = 0 \tag{3.5}$$

$$\frac{\partial E(v_x, v_y)}{\partial v_y} = \sum_{\vec{x}}^{w} 2g(\vec{x}) \left[v_x I_x + v_y I_y + I_t \right] I_y = \sum_{\vec{x}}^{w} g(\vec{x}) \left[v_x I_x I_y + v_y I_y^2 + I_t I_y \right] = 0 \tag{3.6}$$

Em seguida, podemos desenvolver as equações 3.5 e 3.6 utilizando o procedimento descrito pelas equações 3.7 para a variável v_x. Esse procedimento tem como objetivo colocar as equações no formato matricial para facilitar sua manipulação. O mesmo procedimento deve

ser repetido para a variável v_y.

$$\sum_{\vec{x}}^{w} g(\vec{x}) \left[v_x I_x^2 + v_y I_y I_x + I_t I_x \right] = 0 \tag{3.7}$$

$$\sum_{\vec{x}}^{w} g(\vec{x}) v_x I_x^2 + \sum_{\vec{x}}^{w} g(\vec{x}) v_y I_y I_x + \sum_{\vec{x}}^{w} g(\vec{x}) I_t I_x = 0$$

$$v_x \cdot \sum_{\vec{x}}^{w} g(\vec{x}) I_x^2 + v_y \cdot \sum_{\vec{x}}^{w} g(\vec{x}) I_y I_x = - \sum_{\vec{x}}^{w} g(\vec{x}) I_t I_x$$

$$\left[\begin{array}{cc} \sum_{\vec{x}}^{w} g(\vec{x}) I_x^2 & \sum_{\vec{x}}^{w} g(\vec{x}) I_y I_x \end{array} \right] \times \left[\begin{array}{c} v_x \\ v_y \end{array} \right] = - \sum_{\vec{x}}^{w} g(\vec{x}) I_t I_x$$

Então, as equações 3.5 e 3.6 podem ser escritas em conjunto na forma matricial como mostrado pela equação 3.8. Por fim, a estimativa dos mínimos quadrados para esse tipo de equação é $\vec{v} = \boldsymbol{M^{-1}\vec{b}}$.

$$\underbrace{\left[\begin{array}{cc} \sum_{\vec{x}}^{w} g(\vec{x}) I_x^2 & \sum_{\vec{x}}^{w} g(\vec{x}) I_y I_x \\ \sum_{\vec{x}}^{w} g(\vec{x}) I_x I_y & \sum_{\vec{x}}^{w} g(\vec{x}) I_y^2 \end{array} \right]}_{\text{Matriz M}} \times \left[\begin{array}{c} v_x \\ v_y \end{array} \right] = \underbrace{\left[\begin{array}{c} - \sum_{\vec{x}}^{w} g(\vec{x}) I_t I_x \\ - \sum_{\vec{x}}^{w} g(\vec{x}) I_t I_y \end{array} \right]}_{\text{Matriz } \vec{b}} \tag{3.8}$$

Contudo, existem um conjunto de impedimentos práticos para implementação desse algoritmo, conhecido como *Lucas-kanade* e proposto em [35]. A inversão de matriz M^{-1} é implementado como a pseudo-inversa de *Moore-Penrose*. As derivadas parciais I_x, I_y e I_t são calculadas a priori e armazenadas, além de serem computadas em versões suavizadas das imagens para melhorar a aproximação. O pseudocódigo 2 mostra como ocorre a implementação do algoritmo descrito nessa seção. Esse algoritmo para cálculo do fluxo óptico recebe duas imagens, o tamanho das janelas w a serem utilizadas e retorna o vetor velocidade 2D $\vec{v} = (v_x, v_y)^T$.

Se os movimentos não forem pequenos o método descrito não terá boa precisão, devido o uso da série de Taylor para a aproximação de $I(x + v_x, y + v_y, t + 1)$. Então, uma extensão desse método, que utiliza uma abordagem que consiste em repetir o algoritmo *lucas-kanade* em imagens redimensionadas para múltiplas escalas, pode ser utilizado para alcançar uma estimativa de fluxo óptico mais precisa.

Algoritmo 2: Pseudocódigo do algoritmo Lucas-Kanade.

1 Aplique o filtro de suavização gaussiano nas imagens Im_1 e Im_2;
2 Calcule os gradiente I_x e I_y que são as derivadas parciais do frame Im_1 no instante t;
3 Calcule I_x^2, I_xI_y e I_y^2;
4 Execute a convolução de I_x^2, I_xI_y e I_y^2 com a mascara $g(\vec{x})$ para cada janela w;
5 Agrupe os resultados de cada convolução construindo a matriz M da equação 3.8;
6 Calcule a derivada parcial $It = Im_2 - Im_1$;
7 Calcule I_xI_t e I_yI_t utilizando as parcelas já calculadas;
8 Monte a matriz e o vetor $\vec{b}$ da equação 3.8;
9 Faça $\vec{v} = M^{-1} \times \vec{b}$;
10 Retorne $\vec{v} = (v_x, v_y)^T$;

3.3 Extratores de Características Indicadoras de Movimento

Extratores de características indicadoras de movimento são técnicas para filtrar, extrair e representar informações de movimento em sequencia de imagens. Eles basicamente operam em duas etapas: Extração e Representação. Existem basicamente duas abordagens para extrair características indicadoras de movimento: características baseadas em fluxo óptico e características baseadas em subtração de *frames*.

As características baseadas em fluxo óptico consistem em estimar o fluxo óptico entre *frames* subsequentes e calcular diferenças entre os fluxos. Com esse mecanismo, é possível filtrar o movimento da câmera e capturar o movimento do objeto e de partes do objeto. Extratores de características como [31, 30] utilizam essa abordagem. Essas técnicas possuem consideráveis limitações, pois calcular o fluxo óptico com precisão é custoso.

As características baseadas em subtração de *frames* consistem em realizar subtração de *frames* subsequentes. Essas técnicas funcionam bem com câmeras estáticas mas produzem muito ruído para câmeras móveis. Para superar tal limitação, técnicas como [36] são utilizadas par eliminação do *background*. Essa abordagem também é utilizada em operações *multiframes* já que quanto mais distantes temporalmente dois *frames*, mais evidentes ficam os movimentos até certo limiar. O extrator de característica proposto em [32] usa esse tipo de abordagem.

4 *Uma Nova Extensão do ACF*

"A mente que se abre a uma nova idéia
jamais voltará ao seu tamanho original."
– Albert Einstein

Neste Capítulo é apresentada a proposta deste trabalho de conclusão de curso, um detector de humanos em imagens digitais que consiste em uma nova extensão do detector ACF [2]. Essa nova extensão se diferencia do ACF original por ser capaz de extrair características de sequências de imagens. Assim, tornando possível a utilização de características indicadoras de movimento.

Na Seção 4.1 são discutidos os problemas encontrados no ACF e em sua precisão de detecção. Em seguida, na Seção 4.2 é explicada quais alterações são propostas para solucionar esses problemas encontrados. Por fim, o novo conjunto de características proposto é explicado em detalhes.

4.1 Problemas Encontrados no ACF

O *Agreggated Channel Features* [2] é , atualmente, considerado o melhor detector de humanos em imagens digitais devido sua precisão e velocidade de detecção. Essas características foram determinantes na sua escolha para base da nova extensão. No entanto, ele ainda gera uma considerável quantidade de alarmes falsos, principalmente quando seu classificador é ajustado par ser pouco restritivo. Pedaços de *background* e placas de sinalização entre outras entidades são usualmente detectados incorretamente como humanos.

A figura 9 exibe o resultado da aplicação do detector ACF original. Os retângulos azuis são as localizações corretas dos humanos e os retângulos verdes são as detecções do ACF. Podemos perceber que apesar do ACF encontrar todos os humanos na imagem, ele produz um excessivo número de alarmes falsos. Esses alarmes falsos podem ser proibitivos para muitas aplicações. Por fim, a figura 10 exibe um conjunto de alarmes falsos comumente produzido

pelo ACF original.

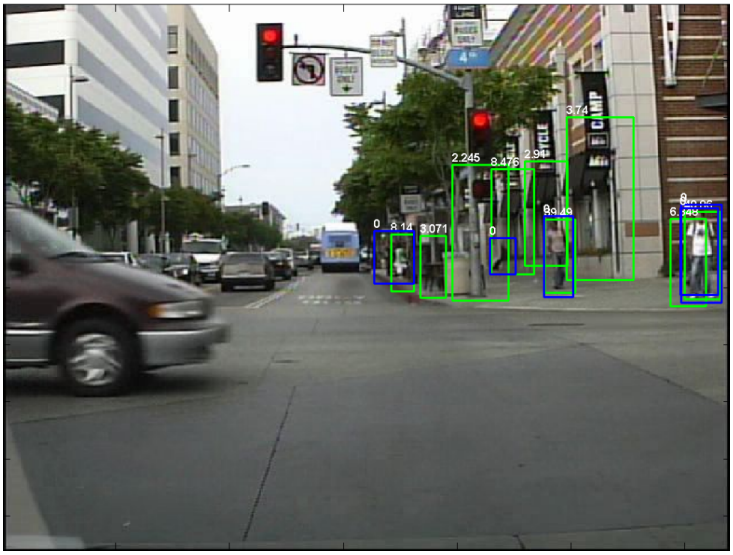


Figura 9: Resultado da aplicação do ACF.

Figura 10: Exemplos de alarmes falsos produzidos pelo ACF original.

O ACF também é um *framework* robusto onde novos extratores de características podem ser acrescentados facilmente. O código para utilização desse *framework* está disponível online em [25]. No entanto, o ACF só possui interfaces para características extraídas de um único *frame* (características estáticas). Isso é considerado uma limitação visto que existem várias características robustas que podem ser úteis no processo de detecção, mas são extraídas a

partir de uma sequência de *frames*. Características indicadoras de movimento são um bom exemplo de características que são extraídas de sequências de *frames*.

Contudo, apesar do ACF ser preciso e rápido, ele produz considerável número de alarmes falsos. Além disso, ele também possui uma estrutura limitada que só permite a inclusão de caraterística estáticas. Então, a extensão proposta por esse trabalho tem como objetivo superar essas limitações do ACF.

4.2 Alterações propostas

A extensão proposta por esse trabalho tem como objetivo reduzir a quantidade de alarmes falsos produzidos pelo ACF original. Além disso, essa nova extensão deve fornecer interfaces para características extraídas de sequências de frames. Desta forma, a nova extensão será mais precisa e mais flexível.

Como mostrado na figura 10, muitos falsos positivos são ocasionados pela incorreta classificação de pedaços de *background* ou entidades inanimadas. Além disso, é notório que o movimento do corpo humano tem um padrão muito particular que o diferencia de várias outras entidades presentes nas imagens. Portanto, essas informações guiaram a decisão de utilizar características de movimento para a detecção de humanos em imagens digitais.

Características indicadoras de movimento foram adicionadas e combinadas com as já utilizadas pelo ACF. As características implementadas são adaptações das características propostas por [31, 37]. Essas características foram adaptadas para o formato de canais de características proposta em [13] e utilizado no ACF. Além disso, uma característica estática baseada em entropia também foi desenvolvida e combinada com todas as outras. Na seção 4.3 essas novas características serão explicadas em detalhes.

Com o objetivo de reduzir a quantidade de alarmes falsos, características de movimento foram adicionadas ao conjunto de características originais utilizadas pelo ACF. No entanto, para inclusão dessas características foi necessário realizar alterações profundas no código do ACF disponível em [25]. Então, foi desenvolvida a interface "Pseq" que permite a adição de extratores que extraem características a partir de sequências de imagens. Essa interface foi parametrizadas da seguinte forma:

- ***Span***: Quantidade total de *frames* anteriores a serem considerados.
- ***Skip***: Quantidade de *frames* a serem ignoradas por frame selecionado.

- ***Handle***: Função que recebe uma sequência de imagens e um conjunto de parâmetros, e retorna canais de características no modelo descrito em [13].

- ***Params***: Conjunto de parâmetros customizáveis.

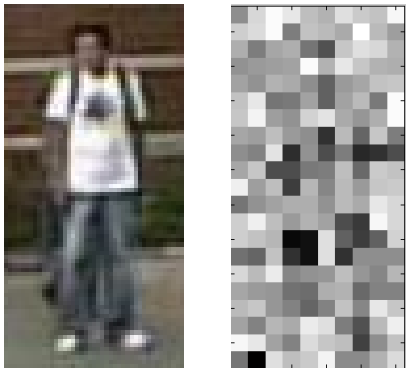
A interface "PSeq" ao analisar um *frame* , ela agrupa "*Span*" *frames* espaçados no tempo pelo parâmetro "*Skip*" em uma lista. Essa lista de *frames* juntamente com o frame em análise e os parâmetros "*Params*" são passados para a função "*Handle*" que extrai as características dessa sequência de imagens e retorna vários canais de características. "Pseq" é estruturada de forma bem semelhante a interface de adição de características estáticas já existente no *framework* do ACF. Porém, ela é voltada para extração de características em sequências de imagens. Assim tornando a nova extensão um *framework* mais maleável que o ACF original.

4.3 Novo Conjunto de Características

Com o objetivo de reduzir a quantidade de alarmes falsos gerados pelo ACF foram utilizadas as seguintes características indicadoras de movimento: *Motion Boundary Histogram* (MBH), *Internal Motion Histogram Central Difference* (IMHcd) e *Weak Stabilized Temporal Difference* (WSTD). Também foi desenvolvida uma característica estática a qual denominamos Entropia do Histograma da Orientação do Gradiente (EHOG). Essas características utilizadas foram combinadas com as já existentes no ACF: Histograma de Orientação do Gradiente, Magnitude do Gradiente e modelo de cor "LUV". Portanto, nas próximas seções deste capítulo cada nova característica utilizada será descrita em detalhes.

4.3.1 Entropia do Histograma da Orientação do Gradiente (EHOG)

EHOG é uma característica indicadora de forma que possui 1 camada. Essa característica pode ser considerada uma extensão da HOG onde calculamos a entropia de *Shannon*, para medir a uniformidade dos histogramas. Quanto maior a entropia do histograma de orientação do gradiente, mais uniforme é o histograma e consequentemente, as direções do gradiente nessa região são mais diversas. Seguindo a mesma lógica, quanto menor a entropia do histograma de orientação do gradiente, menos diversas são as direções do gradiente nessa região. Logo, essa característica pode, teoricamente, distinguir regiões de formas continuas de regiões de formatos diferentes. Como por exemplo, o *background* e as pessoas em uma imagem. A figura 11 exibe a visualização desse canal de característica.



(a) (b)

Figura 11: Visualização do canal de características EHOG. A imagem 11b é o canal EHOG da imagem 11a.

O primeiro passo para extração dessa característica é calcular os histogramas das orientações do gradiente para cada célula de $c \times c$ *pixels* da imagem. Esses histogramas são calculados seguindo o mesmo procedimento da característica HOG [8] que foi descrito na seção 2.2.2. Esse procedimento resulta em um mapa de células $c \times c$ *pixels* onde cada célula possui um histograma com 6 bins e cada bin é um intervalo de ângulos. Esses histogramas são normalizados individualmente dividindo a frequência de cada bin pelo somatório das frequências de cada bin do histograma. Essa normalização torna cada histograma uma variável aleatória onde os eventos são os bins e a distribuição é a frequência normalizada desses bins. Por fim, é calculado a entropia de *Shannon* de cada uma dessas variáveis aleatória utilizando a equação 4.1. Assim cada bloco $c \times c$ *pixels* produz um valor que é a entropia do seu histograma. Na equação 4.1, $p(x_i)$ é a probabilidade do bin x_i no histograma da orientação do gradiente e consequentemente é a probabilidade do intervalo do bin x_i está presente na região onde foi calculado o histograma.

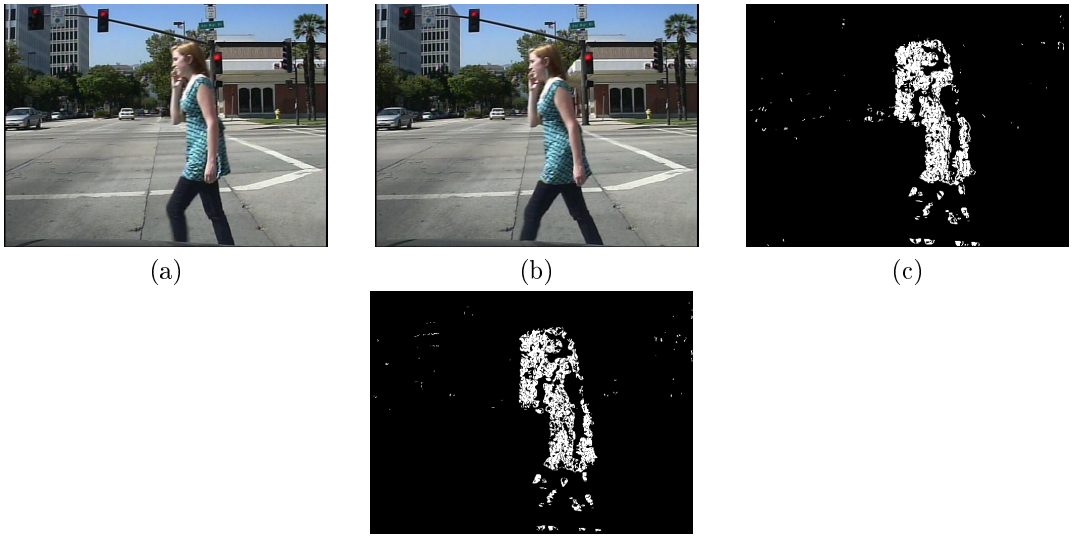
$$H(x) = -\sum_{i=1}^{6} p(x_i) ln(p(x_i)) \tag{4.1}$$

4.3.2 *Motion Boundary Histogram* (MBH)

MBH é uma característica indicadora de movimento que possui doze camadas e foi proposta em [31]. Ela pode ser considerada uma extensão da HOG para detectar as bordas do movimento. Ela consegue filtrar parte do movimento da câmera e capturar o movimento central do objeto. Esse movimento é capturado na direção horizontal e vertical e é representado por meio de histograma.

O primeiro passo para a extração dessa característica é calcular o fluxo óptico $\vec{v} = (v_x, v_i)$ de cada pixel entre as os *frames* F_{t-1} e F_t. Assim, gerando I_x e I_y que são os valores da velocidade na direção x e na direção y para cada pixel. Assim, I_x e I_y podem ser tratados como se fossem duas imagens diferentes onde o valor dos *pixels* são os valores de sua velocidade. Em seguida, o mesmo procedimento realizado para calcular a característica HOG descrita na sessão 2.2.2 é utilizada em I_x e I_y. Desta forma doze histogramas são construídos para cada bloco de $c \times c$ *pixels*. Desses doze histogramas, seis são da componente I_x e seis são da componente I_y. Ao final todos os histogramas são agrupados para formar as doze camadas desse canal de características. O fluxo óptico é calculado com a técnica proposta por *Lucas-kanade* em [35] e descrita em 3.2. Essa técnica foi utilizada devido ao baixo custo computacional requerido.

A figura 12 exibe uma exemplo de visualização desse canal de características. Note que o humano que está em movimento é completamente detectado (*pixels brancos*). Note também que o ruído gerado pelo movimento da câmera foi reduzido, mas o movimento relativo de partes do objeto não fica completamente evidenciado.



(a) (b) (c)

(d)

Figura 12: Visualização do canal de características MBH. As imagens 12c e 12d são a magnitude do gradiente do fluxo óptico na direção x e y respectivamente. 12a e 12b são os *frames* utilizados para extração dessa caraterística.

4.3.3 *Internal Motion Histogram Central Difference* (IMHcd)

IMHcd é uma característica indicadora de movimento que possui seis camadas e foi proposta em [31]. Ela consegue filtrar o movimento da câmera e o movimento central do objeto, assim capturando o movimento relativo de partes do objeto. Na detecção de humanos, essa característica tem como objetivo capturar o movimento de braços, pernas e articulações dos

humanos em movimento na sequência de imagens.

O primeiro passo para a extração dessa característica é calcular o fluxo óptico $\vec{v} = (v_x, v_i)$ de cada pixel entre os *frames* F_{t-1} e F_t. Assim, gerando I_x e I_y que são os valores da velocidade na direção x e na direção y para cada pixel. I_x e I_y são tratadas como imagens onde os *pixels* possuem o valor de sua velocidade na respectiva direção. Em seguida, é utilizada a máscara da figura 13 em I_x e I_y separadamente. Essa máscara calcula diferenças entre fluxos. Ela divide a imagem em células de 3x3 *pixels* e agrupa essas células em blocos não sobrepostos de tamanho 3x3 células. Em cada bloco os *pixels* de cada célula são subtraídos do valor do respectivo *pixel* da célula central. Após a aplicação da mascara de diferença central, as duas parcelas I_x e I_y são usadas para calcular a magnitude M e a os ângulos O. Por fim, M e O são utilizados para construir histogramas, nos moldes dos histogramas da HOG, para cada célula não central em cada bloco. Esses histogramas são as seis camadas desse canal de características.

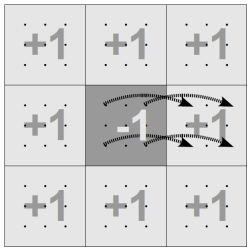


Figura 13: Máscara utilizada para subtração do fluxo óptico dentro de um bloco de 3x3 células. As setas e os valores -1 e +1 mostram como ocorrem as subtrações.

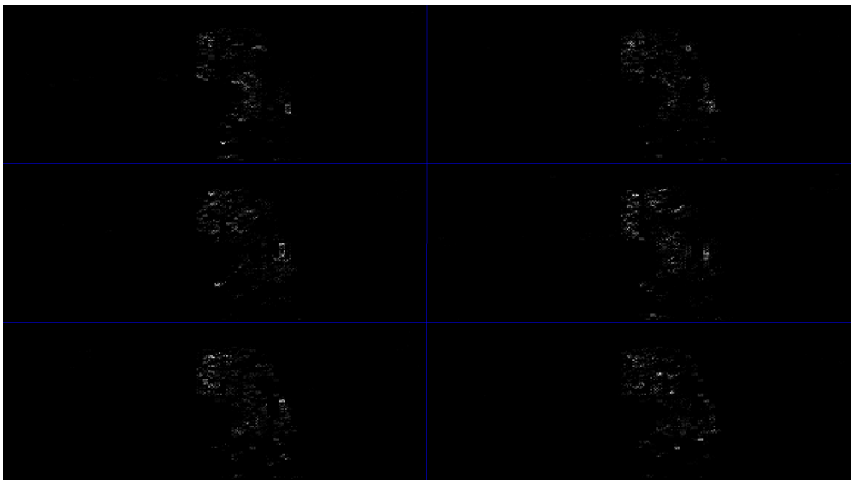


Figura 14: Visualização das seis camadas do canal de características IMHcd. 12a e 12b são os *frames* utilizados para extração dessa caraterística.

A figura 14 exibe a visualização das seis camadas desse canal de característica. Note que em todas as camadas é possível identificar pequenos movimentos (*pixels brancos*) que são relativos ao movimento de partes do objeto em relação ao seu corpo. Note também que a imagem tem suas dimensões alteradas devido o descarte da célula central de cada bloco.

4.3.4 *Weak Stabilized Temporal Difference* (WSTD)

WSTD é uma característica indicadora de movimento e foi proposta em [37]. Essa técnica é uma abordagem híbrida entre a utilização do fluxo óptico e a diferença temporal. Ela consegue filtrar o movimento da câmera e o movimento central do objeto, assim capturando o movimento relativo de partes do objeto. Portanto, essa característica pode capturar o movimento de braços, pernas e articulações dos humanos em movimento.

O WSTD é calculado da seguinte forma: Dado dois *frames* capturados em instantes diferentes I_t e I_{t-1}, o fluxo óptico $\vec{v} = (v_x, v_i)$ de cada pixel da imagem I_{t-1} para I_t é calculado utilizando a técnica [35] que está descrita na seção 3.2. Em seguida, o fluxo óptico $\vec{v}$ é usado para distorcer a imagem I_{t-1} para a imagem I_t produzindo a imagem distorcida $I_{t-1,t}$. Esse processo é chamado de estabilização de sequências de imagens, cujo o objetivo é eliminar os efeitos do movimento da câmera. Então, calculamos o canal de características $C = I_t - I_{t-1,t}$. Por fim, um processo de normalização por blocos é executado em C, assim gerando o canal de característica WSTD.

Esse canal de característica pode ser estendido para maiores diferenças temporais. Essa característica pode ser calculada entre o *frame* I_t e os *frames* $I_{t-(1 \cdot skip)}$, $I_{t-(2 \cdot skip)}$ até $I_{t-(span \cdot skip)}$. Onde $span$ é a quantidade de *frames* utilizados e $skip$ é o espaçamento temporal entre eles. Portanto, esse canal terá $span$ camadas.

A figura 15 exibe a visualização desse canal de característica. O WSTD consegue capturar o contorno do movimento relativo de partes do objeto. Note que o movimento da câmera é significantemente reduzido devido ao processo de estabilização da sequencia de *frames*.

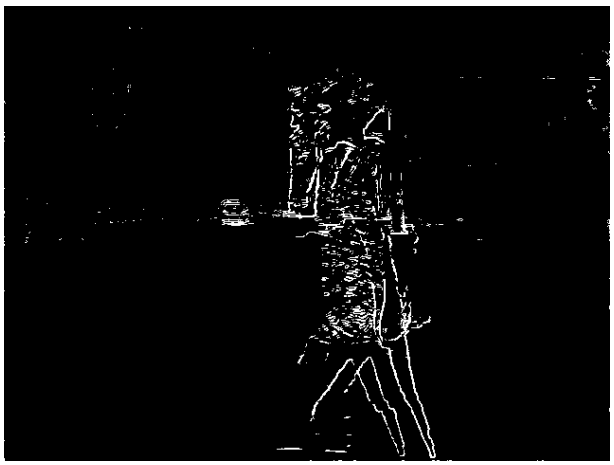


Figura 15: Visualização do canal de características WSTD. 12a e 12b são os *frames* utilizados para extração dessa caraterística.

5 Resultados e Discussão

"O mundo e o universo são lugares extremamente belos,
e quanto mais os compreendemos mais belos eles parecem."
– Richard Dawkins

Este Capítulo tem como objetivo apresentar o arranjo experimental utilizado para avaliar a performance e desempenho do detector de humanos em imagens digitais proposto. Neste Capítulo também é realizado um estudo comparativo da performance das características indicadoras de movimento no contexto da detecção de humanos. Por fim, é realizada uma comparação do detector ACF com o detector proposto. Nesta comparação são destacadas as melhorias alcançadas, assim como a contribuição gerada por esse trabalho de conclusão de curso.

5.1 Arranjo Experimental

5.1.1 *Caltech Pedestrian Data Set*

A utilização de técnicas inteligentes para o desenvolvimento de detectores de humanos em imagens digitais gerou a necessidade da criação de bancos de dados para essa tarefa. Essas coleções de dados são muito importantes para o desenvolvimento e avaliação dos detectores de humanos. [8, 38] são exemplos de bancos de dados utilizados no treinamento e avaliação de detectores de humanos em imagens digitais. Esses bancos existentes podem ser agrupados em duas categorias: Banco de dados de pessoas que consistem em imagens não sequenciais de pessoas em várias poses e em diferentes ambientes; Banco de dados de pedestres que consiste em imagens sequenciais em cenários cotidianos.

O banco de dados escolhido para a utilização nos experimentos deste trabalho é o *Caltech Pedestrian Data Set* proposto em [1]. Esse banco de dados de pedestres consiste em 10 horas de vídeos gravados por uma câmera instalada na frente de um automóvel que foi dirigido

por áreas urbanas com regular tráfico de pedestres e outros automóveis. Ele foi escolhido para o arranjo experimental devido sua praticidade de manipulação, qualidade das anotações e por exibir imagens sequenciais. Sua escolha foi guiada, principalmente, pela necessidade de imagens sequenciais para extração de características indicadoras de movimento. Ele é o banco que mais se aproxima do problema de detecção de humanos em imagens de cenas cotidianas além de ser um dos principais *Benchmark* para avaliação de detectores de humanos em imagens digitais modernos. Observe que o termo pedestre e humanos são utilizados com o mesmo sentido na área de detecção de humanos e nesse trabalho.

As dez horas de vídeo gravados geraram aproximadamente um milhão de *frames*. Dessas dez horas de vídeo, segmentos menores de vídeo foram anotados, assim produzindo aproximadamente duzentos e cinquenta mil *frames* de resolução 640×480 anotados. Foram anotados quase trezentos e cinquenta mil janelas das quais foram observados aproximadamente dois mil e trezentos pedestres diferentes e cento e vinte seis mil pedestres oclusos.

No trabalho proposto por Dollar, Wojek, Schiele e Perona em [1] várias estatísticas em relação a esse banco de dados foram reportadas:

- **Presença de pedestres:** 50% dos *frames* não possuem pedestres enquanto que 30% possui dois ou mais pedestres. Além disso, os pedestres costumam ficar visíveis em média por cinco segundos.

- **Altura dos pedestres:** A altura dos pedestres nas imagens está relacionada a distância do pedestre à câmera no momento em que ele foi filmado. Os pedestres anotados foram divididos em três grupos: "Próximo" , "médio" e "distante". O grupo "próximo" é composto por pedestres cuja altura é maior ou igual a 80 *pixels*. Esses pedestres estavam mais próximos à câmera quando filmados e constituem 15% dos pedestres anotados. O grupo "médio" é composto por pedestres cuja altura está entre 30 e 80 *pixels*. Esses pedestres estavam moderadamente próximos à câmera quando filmados e constituem 69% dos pedestres anotados. Por fim, o grupo "distante" é composto por pedestres cuja altura é menor que 30 *pixels*. Esses pedestres estavam distantes da câmera quando filmados e constituem 16% dos pedestres anotados.

- **Oclusão:** Pedestres oclusos são aqueles que pelo menos alguma parte do corpo está encoberto por alguma outra entidade na imagem. Nesse banco de dados 29% dos pedestres nunca estão oclusos, 53% estão oclusos em alguns *frames* e 19% estão oclusos em todos os *frames* em que eles aparecem. Portanto, mais de 70% dos pedestres estão oclusos em pelo menos um *frame*. Isso mostra a importância da oclusão em aplicações

reais. Também foi investigado os tipos de oclusões encontradas nesse banco de dados. Os pedestres oclusos foram agrupados em três grupos: Completamente oclusos (mais de 80% do corpo encoberto), consideravelmente oclusos (entre 35% e 80% do corpo encoberto), parcialmente oclusos (entre 1% e 35% do corpo encoberto). Os pedestres completamente oclusos são 44% dos pedestres oclusos anotados, enquanto os consideravelmente oclusos são 35% e os parcialmente oclusos são 10% dos pedestres oclusos anotados. Note que exite uma faixa de oclusão entre 0% e 1% do corpo encoberto que são 11% dos pedestres oclusos anotados mas que não são considerados oclusos devido a pequena região encoberta.

- **Posição dos Pedestres:** A posição na imagem onde os pedestres anotados normalmente estão localizados foi investigada. Pode se concluir que os pedestres se localizam na região central da imagem se estendendo horizontalmente por toda largura da imagem. Portanto, essa região deve ser a mais importante para o detector de humanos.

Os arquivos do *Caltech Pedestrian Data Set* estão divididos em onze conjuntos. Os *frames* estão inseridos nesses conjuntos assim como suas anotações. As anotações consistem em um arquivo texto para cada imagem, onde cada linha do arquivo é um pedestre identificado na imagem relacionada a esse arquivo. Cada linha possui a localização do ponto superior esquerdo da janela do pedestre no formato de coordenadas cartesianas onde a origem é a ponta superior esquerda da imagem. Cada linha possui também a largura e a altura da janela assim como o tipo do pedestre identificado. O tipo do pedestre identificado pode ser pedestre individual (aproximadamente 1900 amostras), grupos de pessoas que foram difíceis de separar (aproximadamente 300 amostras) e objetos individuais que não é clara a identificação como pedestre (aproximadamente 110 amostras).

5.1.2 Metodologia de Avaliação

Para uma correta avaliação e comparação de detectores de humanos em imagens digitais devemos utilizar protocolos de avaliação. Esses protocolos devem quantificar e ranquear a performance de detectores de forma realística, informativa e não tendenciosa. Em [1] foi proposto um protocolo de avaliação para ser utilizado com o *Caltech Pedestrian Data Set*. Porém, nesse trabalho de conclusão de curso também foi proposto outro protocolo para tornar a avaliação ainda mais segura e precisa.

O *Caltech Pedestrian Data Set* possui 11 conjuntos ($S0 - S10$). O protocolo proposto em [1] sugere agrupar esses conjuntos em duas partições: Treinamento ($S0 - S5$) e Teste

($S6 - S10$). O conjunto treinamento deve ser usado para treinar o detector e a partição teste deve ser usado para avaliar o detector e reportar suas métricas de performance. Esse protocolo sugere também que durante o treinamento as janelas utilizadas como exemplos positivos e negativos devem ser aleatoriamente selecionadas das janelas extraídas das imagens presentes na partição Treinamento. O mesmo deve ocorrer para a avaliação, mas utilizando a partição Teste. Por fim, o procedimento de treino e teste deve ser repetido várias vezes para remover qualquer influência dos mecanismos aleatórios dos detectores.

A metodologia proposta nesse trabalho é uma adaptação da proposta em [1] utilizando uma abordagem de *10-fold cross validation*. Ela consiste em treinar e avaliar o detector 10 vezes alterando os dados utilizados. Inicialmente os dados são aleatoriamente repartidos em 10 coleções distintas e em cada repetição uma coleção é usada para avaliar e reportar as métricas de performance do detector e as outras retantes para treinar o detector. Assim como a metodologia proposta em [1], durante o treinamento as janelas utilizadas como exemplos positivos e negativos devem ser aleatoriamente selecionadas das janelas extraídas das imagens presentes nas coleções de treinamento. O mesmo deve ocorrer para a avaliação mas utilizando a coleção de teste. Contudo, devido ao grande número de janelas que podem ser extraídas das coleções utilizadas para treinamento, um numero máximo de janelas a serem utilizadas no treinamento deve ser definido.

Simulações com ambas metodologias foram executadas. Em ambos os casos as performances foram relativamente as mesmas. Contudo, os resultados discutidos nessa seção foram obtidos utilizando a abordagem proposta nesse trabalho. Porque a abordagem proposta é baseada em *10-fold cross validation* que é reconhecida por não ser tendenciosa em relação aos dados utilizados e aos mecanismos aleatórios dos detectores.

5.1.3 Métricas

Na avaliação de detectores de humanos em imagens digitais muitas métricas criadas originalmente para problemas de classificação binária podem ser empregadas. Essas métricas geralmente são calculadas em termos de verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). Porém, uma janela contendo uma pessoa pode ter várias dimensões diferentes daquela definida no arquivo de anotação ou podem existir várias janelas para a mesma pessoa na imagem. Portanto, é necessário utilizar um procedimento que defina quem são esses valores e como detectar se uma janela do detector corresponde à anotação da imagem. O procedimento de avaliação de detectores utilizado nesse trabalho é descrito em [1].

Um detector de humanos em imagens digitais recebe uma imagem e retorna uma lista de detecções onde cada detecção possui um *bounding box* e um valor que representa a confiabilidade da detecção. Essa lista e o arquivo com anotações onde estão os *bounding boxes* que são as respostas corretas para a imagem são usadas para avaliar os acertos e os erros do detector. Os *bounding boxes* gerados pelo detector são chamados de BB_{dt} e os *bouding boxes* presentes no arquivo de anotação são chamados de BB_{gt}. Então, um casamento entre um BB_{dt} e um BB_{gt} ocorre quando eles se sobrepõe acima de um limiar (limiar de detecção) que geralmente é 0.5. Esse critério mede a área de sobreposição entre os *bouding boxes* e está descrito pela equação 5.1.

$$a_0 = \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \tag{5.1}$$

No entanto, cada BB_{dt} e BB_{gt} só podem ser casados um única vez, logo é necessário uma regra de prioridade para resolver os impasses gerados. Caso um BB_{dt} case com vários BB_{gt}, o casamento com maior a_0 (área sobreposta) tem prioridade. Utilizando essa regra é possível resolver quase todos os impasses, mas caso alguma impasse persista, a confiabilidade do detector é usado para desempate. Desta forma, podemos calcular os seguintes valores que são as entradas da matriz de confusão para tarefas de classificação binária:

- **Falsos Positivos** (FP): é a quantidade de BB_{dt} não casados.
- **Falsos Negativos** (FN): é a quantidade de BB_{gt} não casados.
- **Verdadeiros Positivos** (TP): é a quantidade de casamentos entre BB_{dt} e BB_{gt}.
- **Verdadeiros Negativos** (TN): é quantidade de BB_{gt} no arquivo de anotação da imagem menos a quantidade de falso positivos, falso negativos e verdadeiros positivos

Podemos calcular as quatro entradas da matriz de confusão mostrada na tabela 1, executando o procedimento descrito para cada imagem e somando suas respectivas parcelas. Por fim, usando essa matriz para classificação binária podemos calcular métricas como *precision*, *recall*, *accuracy*, *f-score*, *miss-rate*, taxa de falso positivos entre outras. Contudo, as métricas mais frequentemente utilizadas para avaliar detectores de humanos em imagens digitais são *miss-rate* (MR) e taxa de falso positivos por imagem (FPPI).

O *Miss-rate* (MR) ou taxa de falso negativo é calculado pela equação 5.2. Essa métrica pode ser interpretada como a quantidade de humanos nas imagens que não foram detectados pelo detector. Portanto, bons detectores apresentam baixo *Miss-rate*.

$$MR = \frac{FN}{TP + FN} = 1 - \frac{TP}{TP + FN} \tag{5.2}$$

Tabela 1: Matriz de Confusão

	Classe Prevista	
Classe Real	Verdadeiro	Falso
Verdadeiro	TP	FN
Falso	FP	TN

O falso positivo por imagens (FPPI) é calculado pela equação 5.3. Essa métrica pode ser interpretada como a quantidade média de alarmes falsos gerados pelo detector. Portanto, bons detectores apresentam baixo falso positivos por imagem.

$$FPPI = \frac{FP}{\#imagens} \tag{5.3}$$

Para comparar a performance dos detectores é utilizado um gráfico que se assemelha a curva ROC. Nesse gráfico é plotado MR x FPPI na escala logarítmica variando o limiar da equação 5.1. Quanto maior o limiar maior o MR e menor o FPPI, pois a avaliação do detector ficará mais restritivo em suas detecções. Por outro lado, quando o limiar for pequeno menor será o MR e maior será o FPPI, pois a avaliação do detector será menos restritiva. Note que esse limiar pode inspirar alterações nos parâmetros do classificador de forma que ele mesmo se torne mais restritivo ou menos restritivo. Desta forma, esse tipo de gráfico resume o desempenho do detector em várias configurações. Por fim, nesse gráfico quanto mais próximo a curva passar da origem, melhor será a performance do detector.

Também é utilizada a média logarítmica do valor do miss-rate em nove diferentes valores de FPPI log espaçados de 10^{-2} à 10^{0}. Essa média é usada para resumir a performance do detector à um único valor, assim simplificando a comparação.

5.2 Parâmetros de Simulação

O conjunto de parâmetros utilizados nas simulações foram retirados dos artigos onde os detectores ou extratores de características foram propostos. Desta forma, uma comparação justa e não tendenciosa foi realizada nesse trabalho de conclusão de curso.

A tabela 2 exibe os parâmetros de utilização do banco de dados *Caltech Pedestrian Data Set*. Esses parâmetros restringem a coleção de dados utilizada para treinamento e teste de detectores. Eles podem ser usados para avaliar o desempenho do detector em cenários específicos. Porém, nesse trabalho de conclusão de curso eles foram configurados para um cenário genérico para comparação de detectores como proposto em [1].

Tabela 2: Parâmetros do banco de dados *Caltech Pedestrian Data Set*

Parâmetro	Descrição	Valor
Labels	Tipos de pedestres	Pedestres individuais e grupo de pessoas
Altura	Altura dos pedestres	Superior à 50 *pixels*
Visibilidade	Percentagem do corpo do pedestre visível	Entre 65% e 100% do corpo visível
janelas positivas	Numero máximo de janelas positivas utilizadas no treinamento	2500 janelas

A tabela 3 exibe os parâmetros utilizados para simulação do detector ACF. Esses parâmetros foram os mesmos utilizados nas simulações descritas no artigo [2] que propôs o detector.

Tabela 3: Parâmetros do ACF.

Parâmetro	Descrição	Valor
Janela de Detecção	Tamanho da janela de detecção	128×64 *pixels*
Agrupamento	Tamanho das células de agrupamento de caraterísticas	células 4×4 *pixels*
Classificadores fracos	Número de árvores binárias de profundidade dois utilizadas pelo *AdaBoost*	2048
Escalas por oitavo	Número de subescalas por oitavo	8 escalas por oitavo
Orientações	Número de orientações do histograma do gradiente	6 orientações
Células	Tamanho das células para calcular o histograma	Células 4×4 *pixels*
Ângulos	Intervalo de ângulos do gradiente	$[0, 2\Pi]$

As tabela 4, 5, 6 exibe os parâmetros utilizados para os extratores de características EHOG, MBH, IMHcd e WSTD. Os parâmetros foram selecionados a partir dos valores propostos em seus artigos e algumas simulações preliminares com o ACF buscando melhor *miss-rate*.

Tabela 4: Parâmetros do EHOG

Parâmetro	Descrição	Valor
Orientações	Número de orientações do histograma do gradiente	6 orientações
Células	Tamanho das células para calcular o histograma	Células 4×4 *pixels*
Ângulos	Intervalo de ângulos do gradiente	$[0, 2\Pi]$

Tabela 5: Parâmetros do MBH e IMHcd.

Parâmetro	Descrição	Valor
Span	*Frames* passados a serem considerados	1
Skip	Espaçamento entre os *frames* selecionados	2
Sigma	Tamanho da janela para calcular o fluxo óptico	16
Orientações	Número de orientações do histograma do gradiente	6 orientações
Células	Tamanho das células para calcular o histograma	Células 4×4 *pixels*
Ângulos	Intervalo de ângulos do gradiente	$[0, 2\Pi]$

Tabela 6: Parâmetros do WSTD.

Parâmetro	Descrição	Valor
Span	*Frames* passados a serem considerados	8
Skip	Espaçamento entre os *frames* selecionados	4
Sigma	Tamanho da janela para calcular o fluxo óptico	16

5.3 Experimentos

5.3.1 Análise da Performance dos Extratores de Características Indicadoras de Movimento

Para medir a performance do ACF com as características indicadoras de movimento descritas em 4.3, é necessário realizar simulações utilizando o banco de dados *Caltech Pedestrian Data Set* apresentado na seção 5.1.1. Nesses experimentos o ACF configurado com diferentes conjuntos de características é treinado e testado seguindo o procedimento descrito na seção 5.1.2 e utilizando parâmetros descritos na seção 5.2. Por fim, para comparação de desempenho as métricas e os gráficos descritos na seção 5.1.3 são calculados e reportados.

A curva *Miss-rate* $\times$ FPPI na figura 16 exibe um espécie de curva ROC utilizada para comparar o desempenho de detectores em várias configurações relativas ao limiar de detecção da equação 5.1. Nessa curva, o limiar de detecção começa em aproximadamente 1 e termina em aproximadamente 0. Portanto, no começo da curva a avaliação do detector é mais restritiva, assim alcançando alto *miss-rate*, mas baixo FPPI. Enquanto que no final da curva onde a avaliação do detector é menos restritiva, ele atinge baixo *miss-rate* e alto FPPI. Por fim, a curva mais próxima à origem é a de melhor performance. Maiores detalhes como escala, e definições de *miss-rate* e FPPI estão descritos na seção 5.1.3.

Portanto, a figura 16 exibe a performance do ACF configurado com as novas características propostas. Nesse gráfico podemos observar a performance de cada nova característica em

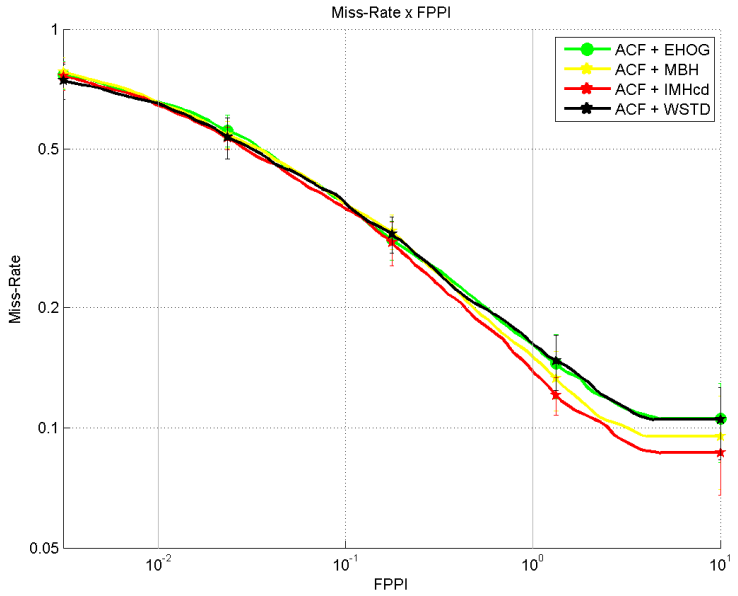


Figura 16: Curva *miss-rate* x FPPI para o ACF configurado com cada uma das novas características propostas.

várias configurações do ACF. Perceba que quando a avaliação do ACF é mais restritiva (no começo da curva), as características possuem precisão semelhante. No entanto, quando o ACF está sendo avaliado de modo pouco restritivo (final da curva) as características apresentam resultados bem distintos. Mais especificamente, se observarmos o valor do *miss-rate* para um certo valor de FPPI essas afirmações ficarão mais evidentes. Por exemplo, para FPPI igual à 10^{-2} temos semelhantes *miss-rate* para todas as características, enquanto que para FPPI igual à 10^{0} notamos menor *miss-rate* para a característica IMHcd seguida por MBH.

Através do gráfico da figura 16 podemos também comparar a quantidade de alarmes falsos (falsos positivos) produzidos por cada nova característica proposta. Se observarmos o valor do FPPI para específicos valores de *miss-rate* podemos perceber qual característica gera menos alarmes falsos. Quando o *miss-rate* está entre $0,3$ e $0,5$, os valores de FPPI são praticamente os mesmos para todas as novas características. Porém, quando o *miss-rate* está entre $0,1$ e $0,2$, a curva do IMHcd apresenta menores valores de FPPI. Perceba que essa diferença ocorre na parte final do gráfico onde o ACF é avaliado de forma menos restritiva. Portanto, essas características podem ser úteis na redução de alarmes falsos utilizando o ACF de forma pouco restritiva. Assim, o ACF poderá alcançar baixo *miss-rate* com aceitável nível de falsos positivos.

Para comparar a performance das novas características podemos também utilizar a métrica média logarítmica dos valores de *miss-rate* em específicos valores de FPPI como descrito na seção 5.1.3. A figura 17 exibe o box plot dessa métrica nas 10 avaliações executadas no

procedimento *10-fold cross-validation* descrito na seção 5.1.2.

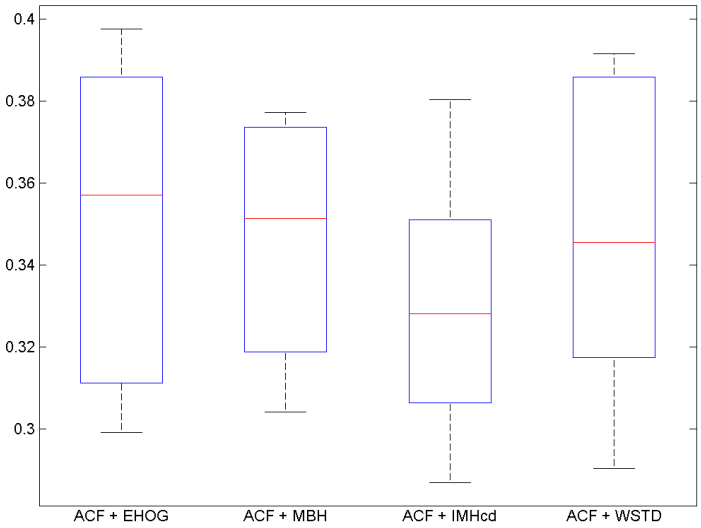


Figura 17: Box plot das médias logarítmicas do *miss-rate* em nove diferentes valores de FPPI log espaçados de 10^{-2} à 10^{0} para o ACF configurado com cada uma das novas características propostas.

O box plot da figura 17 mostra, de forma geral, que a característica IMHcd alcança o melhor desempenho em relação à média logarítmica dos valores de *miss-rate*. Podemos notar que o IMHcd possui melhor mediana e menor dispersão que as outras características. Portanto, o IMHcd confere melhor precisão de detecção que as outras características testadas.

Portanto, os experimentos mostraram através dos resultados descritos nessa seção que a característica IMHcd é a que confere melhor precisão de detecção dentre as testadas. Os resultados mostraram também que características de movimento podem auxiliar o ACF a conseguir menor *miss-rate* e menor taxa de alarmes falsos quando ele estiver em regime pouco restritivo, como o simulado pela avaliação com baixo limiar de detecção. Portanto, O ACF acrescido da característica de movimento IMHcd é a extensão proposta por esse trabalho de conclusão de curso.

5.3.2 Análise das Melhorias Alcançadas pela Extensão Proposta

Para medir o ganho de performance alcançado pelo ACF com a inclusão da característica indicadora de movimento IMHcd descrita em 4.3, foi necessário realizar simulações utilizando o banco de dados *Caltech Pedestrian Data Set* apresentado na seção 5.1.1. Nesses experimentos o ACF com o conjunto original de características foi treinado e testado seguindo o procedimento

descrito na seção 5.1.2 e utilizando parâmetros descritos na seção 5.2. O ACF com inclusão do IMHcd também foi avaliado com os mesmos procedimentos. Por fim, para comparação dessas duas abordagens as métricas e os gráficos descritos na seção 5.1.3 foram calculados e reportados. Nessa seção, o ACF com a adição de IMHcd também será chamado de nova extensão do detector ACF.

A curva ROC exibida na figura 18 descreve de forma geral a performance do detector ACF e da extensão proposta neste trabalho. Como podemos observar a nova extensão (ACF + IMHcd) apresenta melhor precisão de detecção que o ACF original. Essa melhora fica mais evidente no final da curva ROC da figura 18. Nessa região há um baixo *miss-rate* e grande taxa de falsos positivos devido a avaliação com baixo limiar de detecção. Porém, a nova extensão conseguiu atingir baixos *miss-rate* para uma mesma taxa de FPPI. Como por exemplo, o ACF original apresenta *miss-rate* igual a 16% para FPPI igual a 10^0. Enquanto a nova extensão apresenta *miss-rate* igual a 13% com o mesmo FPPI.

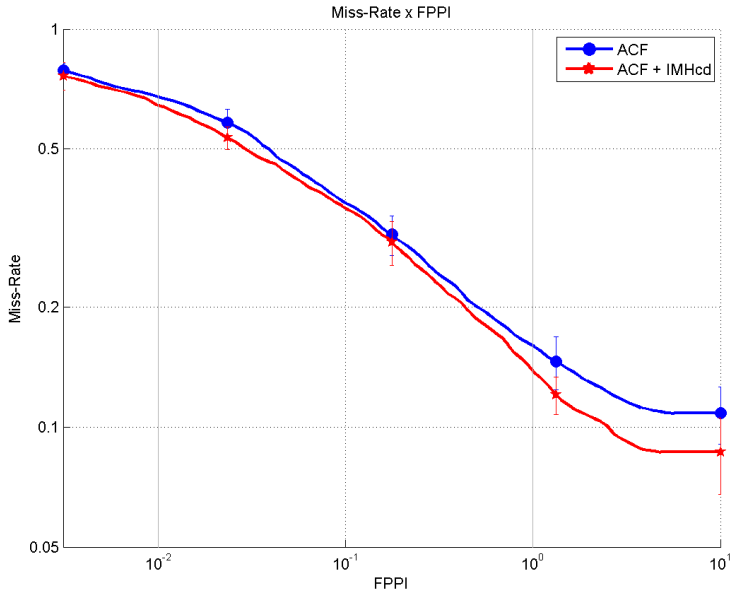


Figura 18: Curva *miss-rate* x FPPI para o ACF original e o ACF com a inclusão da característica IMHcd.

O *box plot* da figura 19 é plotado utilizando as médias logarítmicas dos valores do *miss-rate* em valores específicos de FPPI calculadas em cada um dos 10 *folds* do *10-fold cross-validation* como descrito na seção 5.1.3. Esse box plot exibe de forma resumida o melhor desempenho da nova extensão em relação ao ACF original. Esse melhor desempenho fica caracterizado pela mediana mais baixa e menor dispersão da nova extensão em comparação com o ACF original.

Para medir a redução de alarmes falsos apresentada pela nova extensão foi plotado gráficos *box plot* de valores de FPPI em determinados níveis de referência de *miss-rate* em cada *fold* do

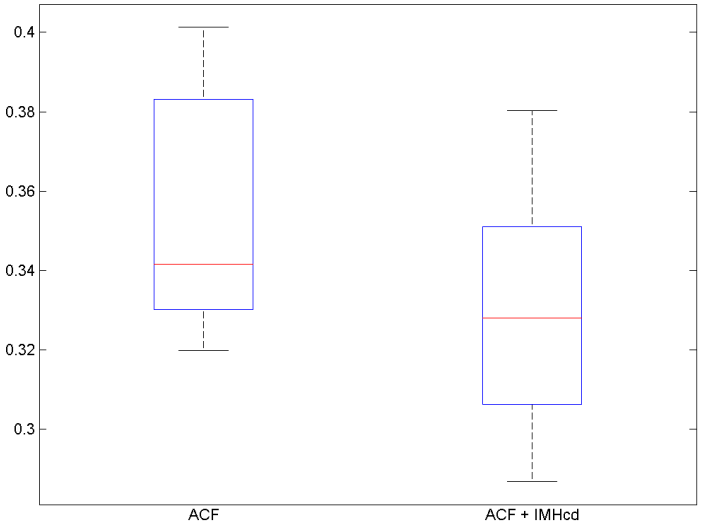


Figura 19: Box plot das médias logarítmicas do *miss-rate* em nove diferentes valores de FPPI log espaçados de 10^{-2} à 10^{0} para o ACF original e a nova extensão proposta neste trabalhos.

10-fold cross validation executado nas simulações. Foram escolhidos três níveis de *miss-rate* para plotar esse gráfico. As figura 20, 21 e 22 utilizam respectivamente 30%, 20% e 10% como valor de referência de miss-rate.

O *box plot* da figura 20 mostra a redução de FPPI para um *miss-rate* igual a 30%. A redução na métrica FPPI consiste na redução de alarmes falsos produzidos pelo ACF original. Portanto, com o *miss-rate* de 30 % a nova extensão produz menos alarmes falsos que o ACF original. Assim, a nova extensão é mais indicada para aplicações onde alarmes falsos são críticos.

O *box plot* da figura 21 mostra a redução de FPPI para um *miss-rate* igual a 20%. Nesse box plot percebemos que ocorre a mesma tendência do box plot anterior, cujo o *miss-rate* de referência é 30%. O *miss-rate* igual a 20% é geralmente alcançado quando o ACF está configurado para ser pouco restritivo, cenário que é simulado pelo baixo limiar de detecção. Então nessa configuração as características de movimento são ainda mais eficientes na redução de alarmes falsos.

O *box plot* da figura 22 mostra a redução de FPPI para um *miss-rate* igual a 10%. Esse *box plot* é o caso extremo de redução de alarmes falsos. Nesse cenário o ACF é muito pouco restritivo em sua detecção e gera um grande numero de alarmes falsos. A nova extensão utilizando a característica de movimento IMHcd é capaz reduzir consideravelmente os falsos

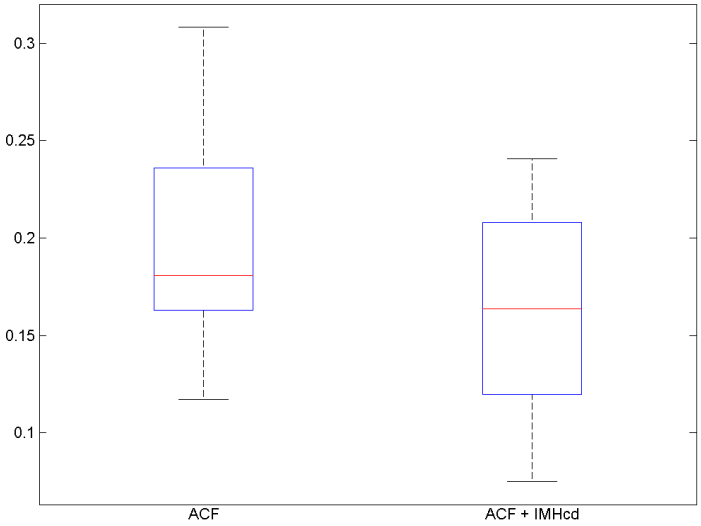


Figura 20: Box plot dos valores de FPPI para *miss-rate* igual a 30% .

positivos gerados pelo ACF original.

A redução de alarmes falsos (falsos positivos) mostrados nos gráficos das figuras 22, 21 e 20 pode ser exemplificada pelas imagens da figura 23. A imagem da esquerda exibe as detecções do ACF original enquanto a da direita são as detecções da nova extensão na mesma imagem. Os retângulos azuis são as detecções corretas enquanto os retângulos verdes são as detecções dos detectores. Podemos observar a redução de 4 falsos positivos comparado essas duas imagens.

A nova extensão proposta consiste na adição da característica de movimento IMHcd ao conjunto original de características utilizadas pelo ACF. Essa nova extensão possui melhor desempenho de detecção que o ACF original. Essa melhora fica mais evidente quando o ACF original é configurado em regime pouco restritivo. Essa melhora é principalmente notada pela redução do numero de alarmes falsos (falsos positivos). Por fim, a nova extensão pode ser configurada em regime pouco restritivo de detecção que mesmo assim apresentará baixo *miss-rate* e aceitável FPPI em comparação com o ACF original.

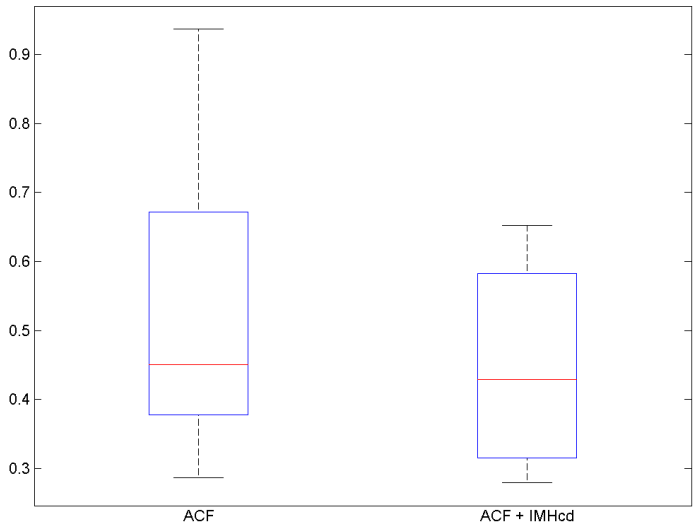


Figura 21: Box plot dos valores de FPPI para *miss-rate* igual a 20% .

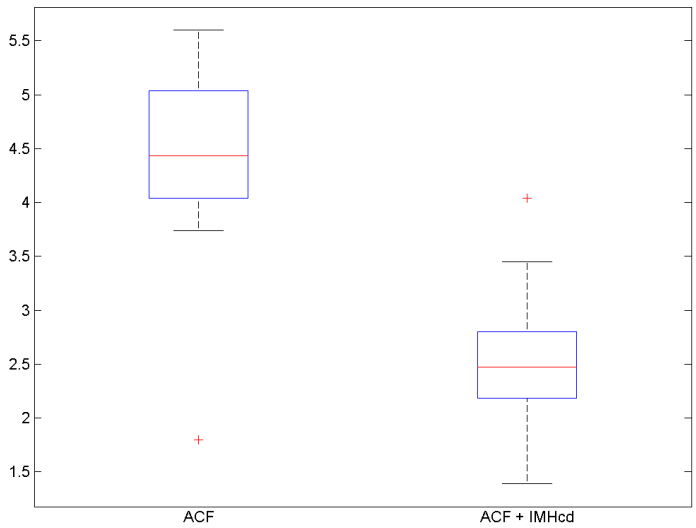
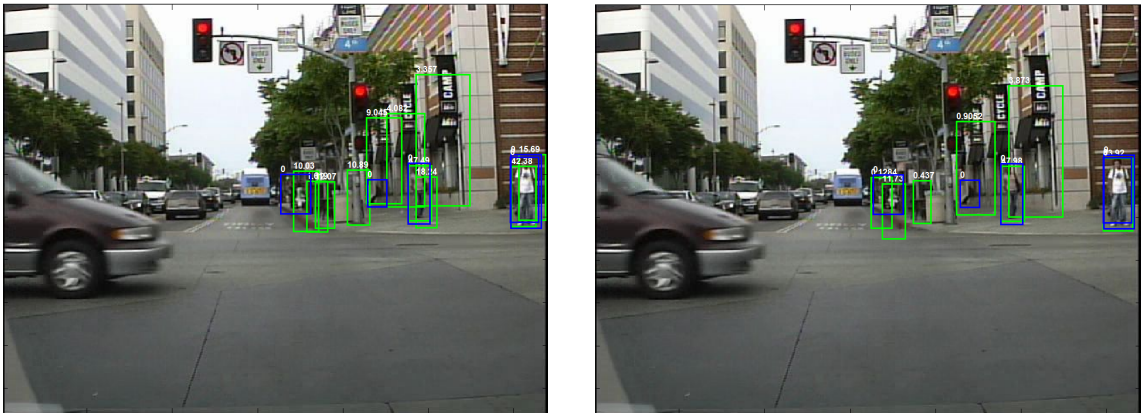


Figura 22: Box plot dos valores de FPPI para *miss-rate* igual a 10% .



(a) (b)

Figura 23: A imagem 23a exibe as detecções do ACF original e a imagem 23b exibe as detecções da nova extensão (ACF + IMHcd).

6 Conclusões e Trabalhos Futuros

"A educação tem raízes amargas
mas os seus frutos são doces."
Aristóteles

Neste trabalho de conclusão de curso foi proposto um nova extensão do detector de humanos em imagens digitais ACF. Essa nova extensão se diferencia do ACF original pela utilização de características indicadoras de movimento. Essa proposta foi desenvolvida com o intuito de melhorar a precisão do detector e reduzir a quantidade de alarmes falsos gerados.

Neste capítulo, serão feitas considerações finais sobre os resultados encontrados e ao comparativo feito entre a extensão proposta e o ACF original. Por fim, os trabalhos futuros serão apresentados.

6.1 Contribuições e Conclusões

A interação entre homem e máquina é um dos principais objetivos da engenharia moderna. Essa interação geralmente se inicia pelo reconhecimento de humanos e a forma mais comum de representar essa tarefa é reconhecimento de humanos em imagens digitais. Essa tarefa é geralmente tratada como um problema de visão computacional e reconhecimento de padrões e suas soluções são úteis para inúmeras aplicações nas áreas de segurança, assistência hospitalar, mobilidade urbana entre outras. Portanto, este trabalho de conclusão de curso contribui para as áreas de visão computacional, reconhecimento de padrões e busca solucionar o problema de detecção de humanos em imagens digitais, o qual vem sendo bastantemente estudado pela academia durante os últimos anos.

A técnica desenvolvida neste trabalho de conclusão de curso é uma técnica para detecção de humanos em imagens digitais. Essa proposta pode ser considerada uma extensão do detector ACF proposto em [2]. Essa extensão se diferencia do ACF original pela utilização de características indicadoras de movimento. Foram testadas alguma características indicadoras

de movimento em simulações utilizando o renomado e desafiador banco de dados *Caltech Pedestrian Data Set*. A característica de movimento IMHcd foi a que alcançou melhores resultados na precisão de detecção e redução de alarmes falsos. Essa característica se mostrou eficiente na captura do movimento de partes do objeto em relação ao corpo do objeto , assim alcançando os melhores resultados em termos de precisão de detecção comparada as outras características testadas. Portanto, a nova extensão proposta é caracterizada pela inclusão da característica IMHcd no conjunto de características utilizado pelo ACF original.

A técnica desenvolvida foi testada e comparada ao ACF original. Esse teste utilizou o banco de dados *Caltech Pedestrian Data Set* e métricas baseadas em *miss-rate* e falso positivos por imagens (FPPI) que são as métricas padrões da área de detectores de humanos em imagens digitais. Essas simulações utilizaram um procedimento de simulação proposto nesse trabalho que consiste em uma melhoria do procedimento de simulação padrão proposto em [1]. O procedimento proposto e utilizado nas simulações e tem o intuito de ser mais informativo e menos tendencioso.

Nos experimentos executados a nova extensão alcançou melhor performance de detecção. Ela se mostrou mais precisa que o ACF original principalmente quando o limiar de detecção é baixo gerando um cenário pouco restritivo onde o *miss-rate* é baixo, mas há uma grande geração de alarmes falsos. Nesse cenário, a nova extensão apresenta resultados ainda melhores que o ACF original. Um estudo comparativo de redução de alarmes falsos também foi realizado. Nesse estudo, a extensão proposta alcança melhores FPPI em valores específicos de *miss-rate* em comparação com ACF original. Essa diferença fica ainda mais evidente quando são escolhidos valores de referência de *miss-rate* mais baixos que caracterizam um cenário pouco restritivo com baixo limiar de detecção.

Além da nova extensão proposta, uma interface para inclusão de características extraídas de sequências de imagens no ACF original foi desenvolvida. Também foi proposto um procedimento de simulação mais informativo e não tendencioso que foi utilizado nas simulações para avaliar os detectores. Essas propostas são as principais contribuições deste trabalho de conclusão de curso.

6.2 Trabalhos Futuros

Como trabalhos futuros pode ser realizado um estudo de tempo de processamento e consumo de memória da nova extensão proposta. Esse mesmo estudo deve avaliar também o ACF original que é pouco avaliado em termos de tempo de processamento e consumo de

memória no artigo que o propõe. Portanto, esse estudo seria um ótimo complemento deste trabalho de conclusão de curso.

Outro importante estudo é avaliar a diferença de performance causada pela utilização de características indicadoras de movimento em conjunto com o mecanismo de aproximação de características chamado de pirâmide rápida de característica utilizado pelo ACF original. Esse mecanismo só foi testado para características extraídas de *frames* estáticos e nunca foi utilizada com características extraídas de sequência de *frames*. Esse estudo pode melhorar os resultados obtidos por este trabalho e tornar esse mecanismo ainda mais genérico e eficiente.

Desenvolver outras formas de representação de características de movimento são melhorias que podem estender este trabalho. A maiorias das características de movimento encontradas na literatura utilizam de alguma forma o histogram para quantizar e representar as informações extraídas. Outas formas de representação como a utilização da transformada *wavelet* podem ser testadas com intuito de melhorar a eficiência dessas características.

Por fim, a implementação da extensão proposta assim como do ACF original em plataformas de hardware como microcontroladores e FPGAs seriam grandes contribuições. Essas implementações viabilizariam várias aplicações em segurança, assistência hospitalar e mobilidade urbana. Como exemplo, a utilização dessas técnicas em UAVs, *Unmanned Aerial Vehicle*, possibilitaria a vigilância de eventos com grande número de pessoas.

Referências

[1] DOLLAR, P. et al. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 34, n. 4, p. 743–761, April 2012. ISSN 0162-8828.

[2] APPEL, R.; PERONA, P.; BELONGIE, S. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 99, n. PrePrints, p. 1, 2014. ISSN 0162-8828.

[3] FONG, T.; NOURBAKHSH, I.; DAUTENHAHN, K. A survey of socially interactive robots. *Robotics and Autonomous Systems*, v. 42, n. 3-4, p. 143–166, 2003. ISSN 0921-8890.

[4] GERONIMO, D. et al. Survey of pedestrian detection for advanced driver assistance systems. In: . Washington, DC, USA: IEEE Computer Society, 2010. v. 32, n. 7, p. 1239–1258. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2009.122>.

[5] FEIL-SEIFER, D.; MATARIĆ, M. J. Defining socially assistive robotics. In: *in Proc. IEEE International Conference on Rehabilitation Robotics (ICORR)*. [S.l.: s.n.], 2005. p. 465–468.

[6] KIM, C.; HWANG, J. neng. Object-based video abstraction for video surveillance systems. *IEEE Trans. on Circuits and Systems for Video Technology*, v. 12, p. 1128–1138, 2002.

[7] PAPAGEORGIOU, C. P. *A Trainable System for Object Detection in Images and Video Sequences*. Tese (Doutorado), 2000. AAI0801978.

[8] DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005.* [S.l.: s.n.], 2005. v. 1, p. 886–893 vol. 1. ISSN 1063-6919.

[9] GAVRILA, D. M. A bayesian, exemplar-based approach to hierarchical shape matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 29, p. 2007.

[10] WU, B.; NEVATIA, R. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, Kluwer Academic Publishers, Hingham, MA, USA, v. 75, n. 2, p. 247–266, November 2007. ISSN 0920-5691. Disponível em: <http://dx.doi.org/10.1007/s11263-006-0027-7>.

[11] WALK, S. et al. New features and insights for pedestrian detection. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. [S.l.: s.n.], 2010. p. 1030–1037. ISSN 1063-6919.

[12] OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 24, n. 7, p. 971–987, Jul 2002. ISSN 0162-8828.

[13] DOLLAR, P. et al. Integral channel features. In: *Proceedings of the British Machine Vision Conference*. [S.l.]: BMVA Press, 2009. p. 91.1–91.11. ISBN 1-901725-39-1.

[14] VIOLA, P.; JONES, M.; SNOW, D. Detecting pedestrians using patterns of motion and appearance. In: *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. [S.l.: s.n.], 2003. v. 2, p. 734–741.

[15] DOLLAR, P.; BELONGIE, S.; PERONA, P. The fastest pedestrian detector in the west. In: *Proceedings of the British Machine Vision Conference*. [S.l.]: BMVA Press, 2010. p. 68.1–68.11. ISBN 1-901725-40-5.

[16] GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. ISBN 013168728X.

[17] GU, C. et al. Recognition using regions. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. [S.l.: s.n.], 2009. p. 1030–1037. ISSN 1063-6919.

[18] BAY, H. et al. Speeded-up robust features (surf). In: . New York, NY, USA: Elsevier Science Inc., 2008. v. 110, n. 3, p. 346–359. ISSN 1077-3142. Disponível em: <http://dx.doi.org/10.1016/j.cviu.2007.09.014>.

[19] LEIBE, B.; SEEMANN, E.; SCHIELE, B. Pedestrian detection in crowded scenes. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. [S.l.: s.n.], 2005. v. 1, p. 878–885.

[20] VIOLA, P.; JONES, M. J. Robust real-time face detection. *International Journal of Computer Vision*, Kluwer Academic Publishers, Hingham, MA, USA, v. 57, n. 2, p. 137–154, maio 2004. ISSN 0920-5691. Disponível em: <http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb>.

[21] WANG, X.; HAN, T.; YAN, S. An hog-lbp human detector with partial occlusion handling. In: *Computer Vision, 2009 IEEE 12th International Conference on*. [S.l.: s.n.], 2009. p. 32–39. ISSN 1550-5499.

[22] CORTES, C.; VAPNIK, V. Support-vector networks. In: . Hingham, MA, USA: Kluwer Academic Publishers, 1995. v. 20, n. 3, p. 273–297. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1022627411411>.

[23] YE, Q. et al. Human detection in images via piecewise linear support vector machines. *Image Processing, IEEE Transactions on*, v. 22, n. 2, p. 778–789, Feb 2013. ISSN 1057-7149.

[24] FREUND, Y.; SCHAPIRE, R. E. A short introduction to boosting. In: *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. [S.l.]: Morgan Kaufmann, 1999. p. 1401–1406.

[25] DOLLÁR, P. *Piotr's Image and Video Matlab Toolbox (PMT)*. Http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html.

[26] RUDERMAN, D. L.; BIALEK, W. Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett.*, American Physical Society, v. 73, p. 814–817, Aug 1994.

[27] APPEL, R. et al. Quickly boosting decision trees – pruning underachieving features early. In: DASGUPTA, S.; MCALLESTER, D. (Ed.). *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. JMLR Workshop and Conference Proceedings, 2013. v. 28, n. 3, p. 594–602. Disponível em: <http://jmlr.org/proceedings/papers/v28/appel13.pdf>.

[28] QUINLAN, J. R. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

[29] LEWIS, R. J.; D, P. An introduction to classification and regression tree (cart). In: *Annual Meeting of the Society of Academic Emergency Medicine in*. [S.l.: s.n.], 2000.

[30] EFROS, A. et al. Recognizing action at a distance. In: *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. [S.l.: s.n.], 2003. p. 726–733 vol.2.

[31] DALAL, N.; TRIGGS, B.; SCHMID, C. Human detection using oriented histograms of flow and appearance. In: *Proceedings of the 9th European Conference on Computer Vision - Volume Part II*. Berlin, Heidelberg: Springer-Verlag, 2006. (ECCV'06), p. 428–441. ISBN 3-540-33834-9, 978-3-540-33834-5.

[32] LAPTEV, I.; LINDEBERG, T. Local descriptors for spatio-temporal recognition. In: *In First International Workshop on Spatial Coherence for Visual Motion Analysis*. [S.l.: s.n.], 2004.

[33] ADELSON, E. H.; BERGEN, J. R. Spatio-temporal energy models for the perception of motion. *J. OPT. SOC. AM. A*, v. 2, n. 2, p. 1861, 1985.

[34] ANANDAN, P. *A Computational Framework and an Algorithm for the Measurement of Visual*. Amherst, MA, USA, 1987.

[35] LUCAS, B. D.; KANADE, T. An iterative image registration technique with an application to stereo vision. In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981. (IJCAI'81), p. 674–679. Disponível em: <http://dl.acm.org/citation.cfm?id=1623264.1623280>.

[36] PICCARDI, M. Background subtraction techniques: a review. In: *Systems, Man and Cybernetics, 2004 IEEE International Conference on*. [S.l.: s.n.], 2004. v. 4, p. 3099–3104 vol.4.

[37] PARK, D. et al. Exploring weak stabilization for motion feature extraction. In: *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2013. (CVPR '13), p. 2882–2889. ISBN 978-0-7695-4989-7. Disponível em: <http://dx.doi.org/10.1109/CVPR.2013.371>.

[38] ENZWEILER, M.; GAVRILA, D. Monocular pedestrian detection: Survey and experiments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 31, n. 12, p. 2179–2195, Dec 2009.