ION FERREIRA PATRUNI

DETECÇÃO DE INTRUSÃO UTILIZANDO FLUXO ÓPTICO E HISTOGRAMA DE GRADIENTES ORIENTADOS

Dissertação apresentada ao Curso de Pós-Graduação em Computação Aplicada da Universidade do Estado de Santa Catarina, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Roberto Silvio Ubertino Rosso Junior
Coorientador: Prof. Dr. André Tavares da Silva

JOINVILLE, SC
2015

P314d Patruni, Ion Ferreira
Detecção de intrusão utilizando fluxo óptico e histograma de gradientes orientados /Ion Ferreira Patruni. – 2015.
131 p. : il. ; 21 cm

Orientador: Roberto Silvio Ubertino Rosso Junior
Coorientador: André Tavares da Silva
Bibliografia: p. 113-118
Dissertação (mestrado) – Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Computação Aplicada, Joinville, 2015.

1.Computação aplicada. 2. Detecção de intrusão de pedestres. 3. Visão computacional. 4. Reconhecimento de padrões. I. Rosso Junior, Roberto Silvio Ubertino. II. Silva, André Tavares da. III. Universidade do Estado de Santa Catarina. Programa de Pós-Graduação em Computação Aplicada. IV. Título.

CDD: 006.6 - 23. ed.

ION FERREIRA PATRUNI

DETECÇÃO DE INTRUSÃO UTILIZANDO FLUXO ÓTICO E HISTOGRAMA DE GRADIENTES ORIENTADOS

Dissertação apresentada ao Curso de Mestrado Acadêmico Computação Aplicada como requisito parcial para obtenção do título de Mestre em Computação Aplicada na área de concentração "Ciência da Computação".

Banca Examinadora

Orientador:
Prof. Dr. Roberto S Ubertino Rosso Junior
CCT/UDESC

Coorientador:
Prof. Dr. André Tavares da Silva
CCT/UDESC

Membros

Prof. Dr. Marcelo da Silva Hounsell
CCT/UDESC

Prof. Dr. Jefersson Alex dos Santos
UFMG

Joinville,SC, 25 de agosto de 2015.

Este trabalho é dedicado a todas as pessoas apaixonadas pela computação gráfica e pelo aprendizado de máquina e que acreditam que os computadores possam executar tarefas desgastantes de forma repetitiva e tomar decisões com quase a mesma precisão que um ser humano.

AGRADECIMENTOS

Inicialmente gostaria de agradecer às minhas queridas esposa e filha, Franciani e Isabeli, não somente pela paciência, compreensão e incentivo, como também pelo auxílio na execução de alguns testes e compilação de dados.

Aos meus pais, Renê e Mara, por terem investido na minha educação e ensinado que no embate entre ciência e religião, a primeira deve prevalecer porque está em constante movimento. Sem falar no incentivo que eles e a minha avó Iris me deram.

Meu agradecimento aos membros da banca Professores Dr. Jefersson Alex dos Santos e Dr. Marcelo da Silva Hounsell pelas sugestões e colaborações.

Gostaria também de agradecer ao acadêmico Rafael Vincence pelo auxílio na programação e execução dos testes.

Meu especial muito obrigado aos meus orientadores Professores Dr. Roberto Silvio Ubertino Rosso Jr. e Dr. André Tavares da Silva, que além de estarem sempre disponíveis nos horários mais improváveis, foram mais do que amigos e me reergueram quando eu estava muito próximo de ter desistido.

Agora, olhando para trás, lembro com muito carinho de todos os mestres que tive ao longo da minha vida, desde o jardim de infância até a faculdade.

Finalmente gostaria de agradecer a todos os colaboradores que integram a força de trabalho da Universidade do Estado de Santa Catarina (UDESC), em todos os níveis.

“O sucesso consiste de se ir de fracasso em fracasso sem perder o entusiasmo”

Winston Churchill

RESUMO

PATRUNI, Ion Ferreira. **Detecção de Intrusão Utilizando Fluxo Óptico e Histograma de Gradientes Orientados.** 131 p. Dissertação (Mestrado em Computação Aplicada) – Universidade do Estado de Santa Catarina. Programa de Pós-Graduação em Computação Aplicada, Joinville, 2015.

Este trabalho trata da detecção de pessoas em imagens de vigilância. Esta busca ocorre em ambientes externos, onde há pessoas e carros transitando conjuntamente. O objetivo é estabelecer uma linha imaginária que, quando ultrapassada por pessoas, gere um alarme ao operador de Circuito Fechado de TV (CFTV). O estudo concentra-se principalmente na utilização conjunta de Histogramas de Gradientes Orientados (HOG) e análise de características estatísticas extraídas do campo de vetores de Fluxo Óptico (FO) para treinamento de Máquinas de Vetores Suportes (SVM) e posterior classificação. Foram realizados alguns experimentos para verificar se a ordem de aplicação dos classificadores baseados em análise do FO e HOG interferem na qualidade final da resposta que é mensurada através das curvas de Características Operacionais do Receptor (ROC) traçadas a partir das matrizes de confusão. O desempenho temporal de cada classificador é medido. Foi criada uma base de dados proveniente de uma câmera de estacionamento, separada em dois grupos: (1) treinamento e (2) testes para controle. O segundo grupo é usado para averiguar se os resultados parciais estão apresentando melhora em relação ao uso isolado das características HOG. Um terceiro grupo de imagens, extraídas de imagens da Internet relacionadas com questões de intrusão, foi escolhido para compor os testes finais efetivos. A abordagem utilizada para avaliar o evento intrusão é baseada na persistência e consistência das classificações, ou seja, envolvendo mais de um grupo de imagens, podendo variar de 3 a 9 quadros para se ter uma geração efetiva de alarme que avise um operador de CFTV. As curvas ROC dos classificadores FO→HOG e HOG→FO ficaram sobreposta para todos os pontos testados. Com relação ao desempenho temporal, a utilização do classificador baseado no FO teve desempenho 3 vezes mais rápido que o classificador baseado em HOG. Para região de interesse de operação dos classificadores usados na detecção de pessoas em vídeos

de vigilância, levando-se em conta a detecção de eventos de intrusão, que é a região inicial da escala das curvas ROC dos classificadores, o classificador baseado em FO→HOG teve melhor desempenho geral. Por fim, 6 simulações foram realizadas no grupo de vídeos separados por eventos nas mais diferentes situações com o objetivo de se avaliar a abordagem proposta de detecção de eventos de intrusão e o classificador FO→HOG teve o melhor desempenho. A adição da informação do movimento, através da análise e classificação de informações extraídas do campo de fluxo óptico para diferenciar pessoas de carros, trouxe melhorias quando se analisa uma sequência de vídeo, gerando mais alarmes verdadeiro-positivos e reduzindo consideravelmente a geração de alarmes falso-positivos.

Palavras-chave: Detecção de Intrusão de Pedestres, Visão Computacional, Reconhecimento de Padrões.

ABSTRACT

PATRUNI, Ion Ferreira. **Detecção de Intrusão Utilizando Fluxo Óptico e Histograma de Gradientes Orientados.** 131 p. MSc Thesis (Mestrado em Computação Aplicada) – Universidade do Estado de Santa Catarina. Programa de Pós-Graduação em Computação Aplicada, Joinville, 2015.

This work is about people detection in surveillance images. This occurs in external environments, where there are transit of both people and cars. The aim is to settle an imaginary line that once overpassed by people an alarm is sent to the CCTV operator. The study is mainly focused in the use of Histograms of Oriented Gradients (HOG) together with analysis of characteristics extracted from vectors fields of Optical Flow (OF) for training of Support Vector Machines (SVM) and subsequent classification. Some experiments are also executed to verify if the order of cascade classifiers based in OF and HOG affects the quality of the final response. This is measured by Receiver Operating Characteristics (ROC) curves plotted from confusion matrixes. Temporal performance of each classifier was measured as well. A new database was created from a parking camera/surveillance system, and the videos were divided into two groups: (1) training and (2) testing for control purposes. The group called testing for control purposes is used to verify if partial results are presenting improvements in comparison to classifications by HOG alone. A third group of images extracted from the Internet related to intrusions situations was selected to compose the effective final test database. The approach used to detect the intrusion event is based on persistence and consistence of individual classifications. That is, involving more than one group of images, varying from 3 to 9 frames until an effective alarm is sent to the CCTV operator. The ROC curves of classifiers OF→HOG and HOG→OF are fully coincident for all tested points. The OF based classifiers are 3 times faster than the HOG based. For the region of interest of classifiers' operation used in people detection for surveillance videos, considering intrusion events, that is the lower portions of x axis of ROC curves, the classifier OF→HOG had the best performance. Finally, 6 simulations were done involving final tests database, split by events in different

situations to evaluate the proposed approach to detect intrusion events and again the OF→HOG classifier had the best performance. Adding movement information by analysing e classifying data extracted from OF field to distinguish people from car improved alarm generation performance, increasing True Positives and reducing False Positives, when analysing a video sequence.

Key-words: Pedestrian Intrusion Detection. Computer Vision. Pattern Recognition.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CFTV	Circuito Fechado de Televisão
DVR	*Digital Video Recorder*
DoG	*Difference of Gaussians*
FO	Fluxo Óptico
FN	Falso-Negativos
FP	Falso-Positivos
FPS	Fotos por Segundo
GPU	*Graphics Processing Unit*
HOG	*Histogram of Oriented Gradients*
IRHT	*Iterative Recursive Hough Transform*
LARVA	*Laboratory for Research on Visual Applications*
LoG	*Laplacian of Gaussian*
MLP	*Multilayer Perceptron*
NVR	*Network Video Recorder*
PC	*Personal Computer*
RE	Retângulos Envolventes
ROC	*Receiver Operating Characteristic*
SIFT	*Scale Invariant Features Transform*
SURF	*Speeded Up Robust Features*
SVM	*Support Vector Machine*
UDESC	Universidade do Estado de Santa Catarina
VP	Verdadeiro-Positivos
VN	Verdadeiro-Negativos

SUMÁRIO

1 INTRODUÇÃO

O uso de imagens de segurança, obtidas por circuitos fechados de televisão (CFTV) são importantes para o acompanhamento dos mais variados eventos, especialmente em resolução de crimes, auxiliando na localização dos culpados, bem como na identificação de testemunhas e de vítimas. Somente a presença das câmeras de forma ostensiva, sem tentativa de camuflá-las, já cumpre um papel importante de tentar dissuadir pessoas mal-intencionadas de invadir algum local.

A evolução das tecnologias e a produção em massa tornaram os custos de implantação de CFTV mais acessíveis e, portanto, é cada vez mais comum ver sistemas complexos com diversas câmeras, até mesmo em residências.

Existem algumas formas de se detectar a intrusão de pessoas, como por exemplo utilizando sistemas de alarmes baseados em sensores de presença ou barreiras infravermelhas, que cumprem bem o seu papel em locais onde apenas pessoas possam transitar. Por outro lado, em portões e garagens este tipo de tecnologia detectaria pessoas e carros, sem conseguir fazer distinção entre eles, gerando alarmes indesejados.

Uma forma alternativa é trabalhar na prevenção, construindo eclusas que são constituídas por dois portões que confinem o veículo entre eles e que sejam acionados eletronicamente de forma intertravada, ou seja, o segundo portão abre somente após o primeiro estar totalmente fechado, evitando a entrada do invasor carona. O invasor carona pode ser tanto outro veículo quanto uma pessoa a pé, aguardando a oportunidade para entrar junto com o morador [GODOY 2014]. Entretanto, nem todos os condomínios e residências tem espaço físico suficiente para implantação deste tipo de solução, pois são necessárias eclusas separadas para entrada e saída de veículos.

Algumas questões principais motivaram o desenvolvimento deste trabalho focado na detecção de intrusão de pedestres em locais de acesso exclusivo para veículos utilizando visão computacional. Primeiramente, houve o aumento da criminalidade, principalmente o crescimento elevado de assaltos a residências, e, segundo Godoy (2014), o portão de entrada de veículos é um dos pontos mais vulneráveis para permitir a invasão, haja vista a facilidade de um grupo de criminosos vencer a primeira camada de proteção (muros) caminhando a pé e sem serem percebidos pelos condutores dos veículos.

Outra questão é a dificuldade de um operador de CFTV em acompanhar todos os eventos e ocorrências, dado o número crescente de câmeras em uma central de operações e monitoramento. Isto traz um

problema imediato. Segundo Wallace et al. (1998), estimar quantas câmeras efetivamente um operador de CFTV pode monitorar é algo bastante complexo, pois depende de diversos fatores tais como do tipo de cenário sendo observado, ou do tipo de atividade sendo filmada, da frequência da incidência de eventos e sua duração temporal, além de outras tarefas inerentes à central de operação e demais responsabilidades do operador. Embora não seja totalmente conclusivo dizer quantas câmeras um operador pode acompanhar, acredita-se que 16 seja o limite máximo. Existe também o tempo de observação, que idealmente situa-se entre 30 minutos a 2 horas, com intervalos para descanso [WALLACE et al. 1998].

Finalmente, mas não menos importante, é a possibilidade de se monitorar as imagens remotamente, serviço este que pode inclusive ser terceirizado. Assim, condomínios que não dispõem de portaria física com vigilância humana 24 horas por dia, podem controlar melhor o fluxo de pessoas ao condomínio a um custo bem acessível. Do ponto de vista da empresa de monitoramento, esbarra-se em um problema ainda maior do que o comentado na consideração anterior, pois um operador terá que lidar com imagens de vários condomínios.

Dito isto, é importante utilizar um sistema automatizado para análise de vídeos, que auxilie o operador a focar sua atenção em determinados eventos potencialmente mais perigosos.

Partindo-se do conceito de vigilância por CFTV automatizada [THIEL 2000], mais especificamente sobre as cercas virtuais por processamento de imagem, o objetivo primário é aperfeiçoar a filtragem dos objetos presentes na cena de forma mais seletiva, analisando apenas regiões de interesse bem delimitadas. Em outras palavras, dada uma entrada exclusiva de veículos (portões), deseja-se identificar e gerar um alarme para o operador de CFTV toda vez que alguma pessoa tentar entrar a pé e invadir o imóvel no momento em que o portão é aberto para que os automóveis entrem. As cercas virtuais são linhas imaginárias posicionadas em regiões estratégicas do vídeo (sob portões ou muros) e funcionam como barreiras, detectando objetos que a ultrapassam. Movimentações de pessoas do portão para fora podem ser ignoradas, ou, alternativamente, gerar apenas alertas de potencial risco de invasão. Desta forma, o objetivo pode ser alcançado aplicando-se técnicas que identifiquem pessoas em imagens de vigilância.

Também é desejável que o sistema automatizado consiga identificar 100% dos casos de tentativa de invasão por parte dos pedestres (evitando situações de falso-negativos), com o mínimo de geração de alarmes falso-positivos.

1.1 OBJETIVOS

Nas subseções a seguir são detalhados o objetivo geral e os objetivos específicos.

1.1.1 Objetivo Geral

O objetivo geral é desenvolver um sistema de detecção de intrusão de pessoas de forma automática e eficiente, auxiliando o trabalho do operador de CFTV. Para tanto, é necessário melhorar o desempenho do classificador baseado em HOG para detecção de pessoas adicionando em cascata um classificador treinado a partir de informações extraídas do campo vetorial de Fluxo Óptico.

1.1.2 Objetivos Específicos

Neste trabalho experimental são executados testes para encontrar respostas às seguintes situações:

1. Por que HOG não tem o desempenho esperado encontrado na literatura em vídeos adquiridos em situações reais?
2. Qual o melhor tamanho da grade (ou abertura) para se obter um fluxo óptico estável e que possa ser usado como informação para se diferenciar pessoas de carros?
3. O fluxo óptico pode ser usado como filtro de movimentação antes de se executar a varredura da janela HOG?
4. Adicionar a informação do fluxo óptico através de um classificador separado em cascata com o classificador HOG traz alguma melhora de resultados?
5. Qual o efeito de se criar um tensor de fluxo óptico de quadros de imagens consecutivos (matriz tridimensional) como entrada da SVM no desempenho da classificação de pessoas e carros?
6. Como estabelecer uma estratégia de parametrização para geração de alarme na ocorrência do evento de intrusão da área delimitada por cerca virtual em uma sequência de vídeo?
7. O desempenho entre optar primeiramente pela classificação HOG seguida por FO é diferente de optar primeiramente pela classificação FO seguida por HOG (onde são avaliados a

qualidade do reconhecimento das intrusões e o tempo de resposta)?

8. E finalmente, há diferença em se optar entre rede neural MLP ou SVM como ferramenta classificadora no reconhecimento de padrões na identificação de pessoas?

1.2 ESCOPO

O escopo limita-se ao processamento gráfico e aprendizado de máquina, com o objetivo de se identificar a intrusão de pessoas em áreas delimitadas por cercas virtuais. Os experimentos foram realizados em uma base de dados de imagens criada pelo próprio autor e disponíveis em http://www2.joinville.udesc.br/~larva/portal/Projetos.php/vis/115 do *Laboratory for Research on Visual Applications* (LARVA), UDESC, sob o nome de Projeto de Detecção de Intrusão Utilizando Visão Computacional. Outras imagens foram baixadas da Internet e serviram para compor o grupo de testes finais.

A aquisição das imagens é feita por uma câmera (monocular), portanto a análise das imagens limita-se ao plano bidimensional.

As pessoas devem estar na posição ereta e a busca é por corpos inteiros. Via de regra, pessoas entrando de moto, bicicleta ou carro conversível não são tratados como invasores. Para geração de alarmes, não é necessário identificar cada pessoa individualmente e rastreá-la, bastando, portanto, que pelo menos uma pessoa esteja dentro da região delimitada por um intervalo de tempo específico para que esteja caracterizado o evento intrusão. Grupos de pessoas transitando dentro da região delimitada também geram alarme. As oclusões não serão tratadas.

A ideia inicial é que o sistema não necessite ser treinado para cada novo ambiente. Isto inviabilizaria a implantação no cliente final.

Algumas considerações quanto às restrições/requisitos, que foram estabelecidos para que o sistema seja confiável e possa ser usado em tempo real:

- o sistema deve ser capaz de processar imagens de baixa resolução (em torno de 320x240 pixels);
- as pessoas devem ocupar um tamanho de pelo menos 80 pixels de altura, por aproximadamente 33 pixels de largura [DÓLLAR et al. 2012];
- o desempenho do reconhecimento das pessoas e geração de alarmes deve ser em tempo real. Entenda-se o termo tempo real

neste trabalho como sendo um tempo de no máximo 67 milésimos de segundo, porque as imagens são capturadas a uma taxa aproximada de 15 fotos por segundo (fps);

- o sistema será treinado com banco de dados de imagens provenientes de câmeras de vigilância monoculares instaladas em diversos ambientes, com ajustes diferentes.

1.3 RESULTADOS ESPERADOS

O esperado é que a introdução de características extraídas do fluxo óptico, que é uma informação obtida pelo movimento em quadros consecutivos, possa aumentar o índice de acerto dos verdadeiro-positivos no caso da classificação entre pessoas e não-pessoas por uma máquina de vetores suporte. Supõe-se que a melhor resposta seja obtida pela classificação por características HOG conjuntamente com classificação em cascata de características estatísticas extraídas do campo FO. Espera-se também que a classificação conjunta HOG→FO resulte em menor quantidade de falso-positivos. Sendo a movimentação expressa por uma matriz de movimentação de tamanho reduzido, uma fração de NxN em função do tamanho da abertura escolhida para cálculo do fluxo óptico, espera-se também que a classificação por FO seja efetuada mais rapidamente que a classificação por HOG. A métrica para determinar o melhor desempenho é através de curvas ROC.

1.4 ESTRUTURA DO TEXTO

A fim de apresentar a pesquisa realizada, o texto foi dividido da seguinte forma: o segundo capítulo apresenta a fundamentação teórica de conteúdos envolvidos com a pesquisa abordando o reconhecimento de padrões com ênfase na classificação de pessoas, bem como as tecnologias correlatas com o tema e que são de valia no auxílio de localização de pessoas em imagens computacionais. O terceiro capítulo introduz os trabalhos correlatos, enquanto o quarto capítulo traz uma proposta de arquitetura e modelo de classificação em cascata. No quinto capítulo são discutidos os testes e os resultados obtidos a partir dos experimentos projetados. Finalmente, no sexto capítulo são apresentadas as contribuições e considerações finais deste trabalho, seguidas pelas referências e apêndices.

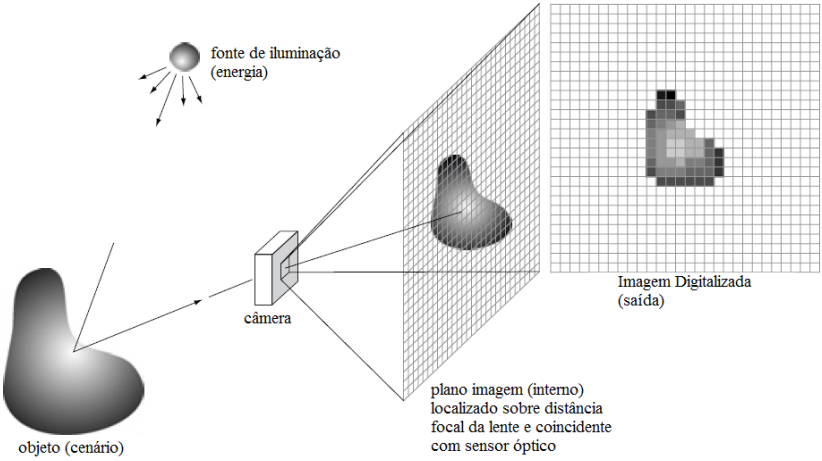
2 FUNDAMENTAÇÃO

Nesta seção serão apresentados alguns conceitos básicos e também técnicas que podem ser utilizadas na detecção de pessoas em vídeos e, portanto, serem úteis na consecução dos objetivos do trabalho.

2.1 CÂMERAS DIGITAIS

A maioria das câmeras digitais possui em seu interior um sensor óptico, composto de vários sensores individuais dispostos na forma de uma matriz de duas dimensões [GONZALEZ et al. 2002]. Assim, uma câmera de 2 Mega pixels (ou células), possui um sensor de 1920 colunas por 1080 linhas. A resposta de cada uma das células é proporcional à integral da energia luminosa projetada sobre ela. A Figura 1 mostra como a energia da fonte luminosa refletida por um objeto é convertida na imagem digitalizada. A primeira tarefa desempenhada pelo sistema é coletar a energia luminosa e focá-la ao plano da imagem. As lentes das câmeras projetam as imagens do cenário na região do seu plano focal. O sensor óptico, que é coincidente com o plano focal da lente, produz saídas proporcionais à integral da energia luminosa recebida por cada uma das milhares de células. Finalmente, circuitos analógicos e digitais fazem as transduções necessárias para se ter um sinal de vídeo analógico único, que posteriormente é digitalizado em outra etapa do processo.

Figura 1 – Sistema de digitalização de imagem

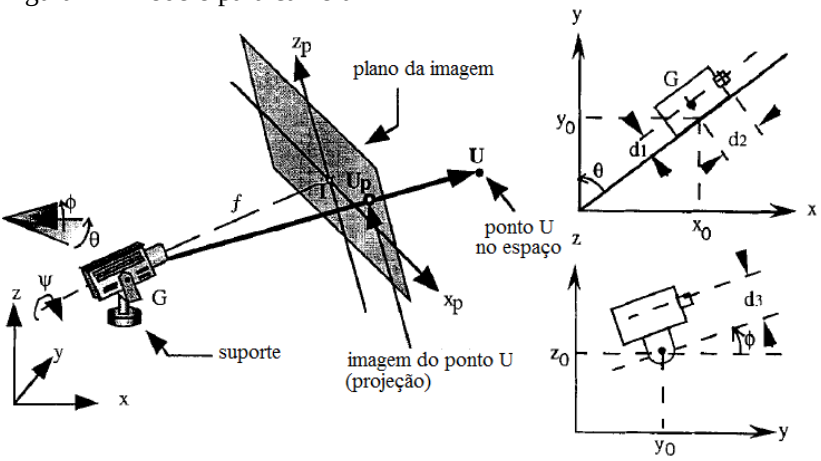


Fonte: Gonzalez et al., 2002.

As imagens digitais são obtidas pela projeção do cenário observado sobre a grade do sensor óptico. Em sistemas calibrados, onde se tem conhecimento dos parâmetros da câmera, tais como distância focal e posicionamento e orientação em relação ao ambiente sendo filmado, é possível ter um modelo geométrico da perspectiva direta (Figura 2), que podem ser expressa como sendo [PARRISH et al. 1977 apud PERRIN et al. 1996]:

$$U_P = SxU \quad U_P = (X_P, Y_P, Z_P, 1)^t \quad U = (x, y, z, 1)^t \qquad (1)$$

Figura 2 – Modelo para câmera



Fonte: Perrin et al., 1996.

A matriz de transformação S é obtida pelo produto das matrizes individuais:

$$S = P \times G \times R \times T \qquad (2)$$

onde:

- P: matriz da imagem, que transforma um ponto do objeto em um ponto da imagem considerando-se apenas a distância focal f da lente;
- G: transformação linear que move o plano de imagem para o centro do suporte da câmera;
- T: matriz da translação, que move o centro do suporte para o quadro de referência do mundo real;
- R: transformadas lineares que alinham o plano de imagem com quadro de referência do cenário real:$R = R\psi \times R\phi \times R\theta$.

Figura 3 – Matrizes de transformação

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1/f & 0 & 1 \end{bmatrix} \qquad G = \begin{bmatrix} 1 & 0 & 0 & -d_1 \\ 0 & 1 & 0 & -(d_2+f) \\ 0 & 0 & 1 & -d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T = \begin{bmatrix} 1 & 0 & 0 & -x_0 \\ 0 & 1 & 0 & -y_0 \\ 0 & 0 & 1 & -z_0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad R_S = \begin{bmatrix} \cos\psi & 0 & \sin\psi & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\psi & 0 & \cos\psi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\phi & \sin\phi & 0 \\ 0 & -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad R_P = \begin{bmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Fonte: Perrin et al., 1996.

Além disto, caso se conheça o fator de escala interno da câmera (elementos/mm do sensor óptico), então pode-se estimar as dimensões dos objetos reais. Conhecer o modelo da câmera, no sentido do tratamento matemático das relações entre a imagem real e a imagem projetada no plano de captação das imagens do conversor óptico da câmera, é fundamental para conseguir ajustar as pessoas eretas de forma que fiquem verticalmente dispostas para correta utilização da classificação por HOG (ver Figura 35). Este ajuste pode ser obtido pela calibração das câmeras, método este descrito na seção seguinte.

2.2 CALIBRAÇÃO DE CÂMERAS

A calibração de câmeras é um passo indispensável para se recuperar métricas 3D a partir de imagens 2D.

Métodos tradicionais de calibração envolvem a utilização de um objeto padrão especial, colocado no campo de visão da câmera. O formato deste objeto é bem conhecido, ou seja, as coordenadas em 3D de qualquer ponto sobre este objeto de referência são perfeitamente conhecidas em relação a um sistema de coordenadas atrelado ao objeto. Geralmente o objeto de calibração é uma chapa plana impressa com um padrão regular. O padrão de impressão é escolhido de forma que as coordenadas da projeção dos pontos de referência (geralmente quinas) possam ser medidas com precisão. Usando um número considerável de

pontos, a matriz de transformação de perspectiva M pode ser calculada (Equação 3).

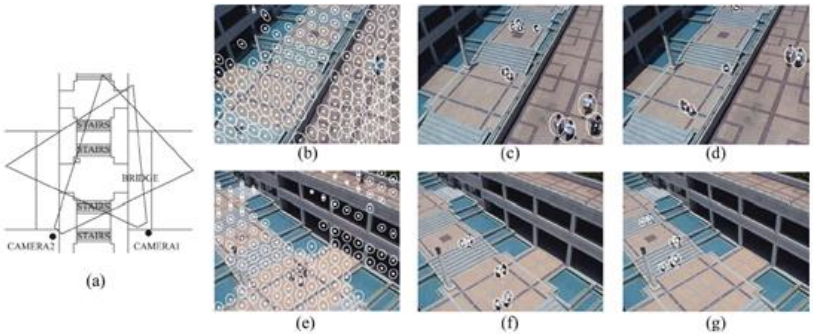
$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \mathbf{A} \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{bmatrix} \mathbf{G} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \equiv \mathbf{M} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (3)$$

Na Equação 3, [su, sv, s] são as coordenadas da imagem, onde s é um fator de escala diferente de zero, [X Y Z 1] são as coordenadas do mundo real, A é uma matriz de transformação 3x3 que diz respeito às características ópticas e de amostragem da câmera, G é a matriz com parâmetros de posicionamento e orientação da câmera e finalmente M é a matriz de transformação de perspectiva, que faz a relação entre as coordenas 3D do cenário real com as coordenadas 2D da imagem projeta e captada pela câmera [FAUGERAS et al. 1992].

Métodos alternativos fazem a calibração das câmeras utilizando as equações de Kruppa, onde são necessários três quadros de imagens diferentes para se obter informação suficiente a fim de se resolver o sistema matricial [FAUGERAS et al. 1992].

Outra abordagem é a segmentação de pedestres em imagens de vídeo descalibradas através de mapas de projeção, onde não é necessário se conhecer os parâmetros da câmera e nem do plano sendo filmado. O tamanho e orientação da projeção de pedestres são estimadas em cada ponto da imagem e registradas em um mapa de projeção de pessoas, durante a etapa de aprendizagem [JO et al. 2009]. Na Figura 4, a passagem consiste de um piso multinível (alternância entre regiões planas e escadas) e um mezanino (*bridge*). As imagens (b), (c) e (d) correspondem à câmera 1, enquanto as imagens (e), (f) e (g) correspondem à câmera 2. Em (b) e (e) é possível ver os mapas de projeção computados, que são de fato as posições e tamanhos esperados das pessoas em cada posição específica da imagem, representados por elipses de tamanhos e inclinações variados. Finalmente em (c), (d), (f) e (g) tem-se o resultado da segmentação de pedestres.

Figura 4 – Segmentação por mapa de projeção



Fonte: Jo et al., 2009.

2.3 FLUXO ÓPTICO

Para se detectar a movimentação de partes ou do todo de pessoas, carros e outros objetos que possam compor uma cena, existem técnicas como o fluxo óptico. Originalmente a ideia de fluxo óptico aborda a questão da percepção de movimento no campo visual e a possibilidade de tais estímulos ópticos poderem estar relacionados com problemas comportamentais e psicológicos [GIBSON 1954].

O fluxo óptico é a distribuição das velocidades aparentes do movimento do padrão de brilho em uma imagem. O fluxo óptico pode surgir do movimento relativo entre objetos e o observador [HORN et al. 1981 apud TOMPKIN 2008]. Outra descrição trata o fluxo óptico como uma aproximação ou estimativa do movimento da imagem, definido como a projeção das velocidades dos pontos de uma superfície 3D sobre o plano de imagem de um sensor visual [BEAUCHEMIN et al. 1995 apud TOMPKIN 2008].

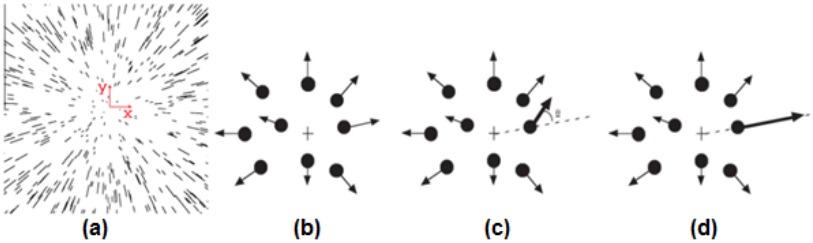
Do ponto de vista computacional, obter-se o fluxo óptico refere-se à tarefa de estimar o campo de vetores de deslocamentos em uma sequência de imagens, ou seja, em um vídeo por exemplo. Em outras palavras, considerando-se uma imagem digital, o fluxo óptico pode ser entendido como uma tentativa de estimar para onde e a que velocidade os pixels estão se movendo. Este é um dos problemas fundamentais em processamento de imagens, obter a velocidade da imagem em determinadas regiões.

Quando se fala em calcular o fluxo óptico, o objetivo é computar uma aproximação ao campo de movimentação 2D – que é

uma projeção das velocidades 3D dos pontos da superfície do objeto no plano da imagem – a partir de padrões espaço-temporal da intensidade da imagem [HORN 1986; VERRI et al. 1987 apud BARRON et al. 1994].

A Figura 5 (a) mostra o campo de fluxo óptico resultante para uma nuvem de pontos 3D, sendo que o observador está transladando diretamente para frente ao longo do eixo Z. Cada linha representa a velocidade da imagem para o ponto específico na cena. Ainda com relação à mesma Figura 5, tem-se em (b) objetos movimentando-se conforme um fluxo óptico radial normal, similar ao caso (a), onde os objetos estão parados e o observador em movimento. Já em (c) o objeto em movimento pode ser percebido pela variação angular. Com exceção do objeto destacado com um vetor mais grosso, todos os demais objetos estão se movendo com fluxo radial. Aquele ponto marcado e destacado com ângulo θ destoa dos demais apenas em relação ao seu ângulo de movimentação, inconsistente com a angulação que teria caso fizesse parte da cena estática. Finalmente, em (d) o objeto em movimento pode ser percebido pela variação de velocidade. Novamente o objeto em movimento é destacado por um vetor mais grosso e como seu comprimento é maior que os demais, ele está muito mais rápido do que deveria estar se movendo caso fosse parte integrante do conjunto de objetos estacionários [ROYDEN et al. 2012].

Figura 5 – Campos de fluxo óptico



Fonte: Royden et al., 2012.

Existem várias maneiras de se computar o fluxo óptico, seja através do uso de cálculo diferencial e integral, minimização de erros, casamento de regiões, ou ainda métodos baseados em energia e em fase. A Tabela 1 mostra um comparativo de desempenho dos diferentes métodos e, entre os que geram campos vetoriais densos para 100% dos pontos de uma imagem, o método modificado de Horn e Schunck teve melhor acurácia [BARRON et al. 1995].

Analisando ainda a Tabela 1, na última coluna chamada densidade (quantidade de pontos para os quais o fluxo óptico pode ser calculado dividido pela quantidade total de pontos da imagem), percebe-se que os métodos para estimar o campo de fluxo óptico podem ser separados em duas categorias: esparsos e densos, sendo que os densos levantam o campo de fluxo óptico para a totalidade dos pixels de uma imagem enquanto os esparsos calculam o fluxo óptico apenas para alguns pontos. Quando se tem uma sequência de imagens ordenadas, que é o caso de imagens de vídeos, é possível através da análise de quadros adjacentes, estimar a velocidade instantânea e/ou deslocamentos discretos (no sentido de infinitesimal). Alguns métodos de cálculo do fluxo óptico tentam medir justamente esta movimentação entre dois quadros entre os instantes t e t+Δt em cada sub-região. Tais métodos usam a aproximação do sinal da imagem por série de Taylor, ou seja, usam derivadas parciais em relação às variáveis espaciais e temporais.

Tabela 1 – Desempenho dos diferentes métodos de Cálculo do FO

Técnica/Método	Erro Médio	Desvio Padrão	Densidade
Horn and Schunck (original)	47,21º	14,60º	100%
Horn and Schunck (original) $\|\nabla I\| \geq 1.0$	27,61º	9,86º	18,9%
Horn and Schunck (modificado)	**32,81º**	**13,67º**	**100%**
Horn and Schunck (modificado) $\|\nabla I\| \geq 1.0$	26,46º	10,86º	42,9%
Lucas and Kanade ($\lambda_2 \geq 1.0$)	0,21º	0,16º	7,9%
Lucas and Kanade ($\lambda_2 \geq 5.0$)	0,14º	0,10º	4,6%
Uras et al. (det(H) > 1.0)	0,15º	0,10º	26,1%
Nagel	34,57º	14,38º	100%
Nagel $\|\nabla I\|_2 \geq 1.0$	26,67º	11,84º	44,0%
Anandan (sem gatilho)	31,46º	18,31º	100%
Anandam ($c_{min} \geq 0.25$	10,46º	5,36º	0,6%
Singh (passo = 1, n = 2, w = 2)	49,03º	21,38º	100%
Singh (passo = 1, n = 2, w = 2, $\lambda_1 \leq 5.0$)	9,85º	21,09º	4,2%
Singh (passo = 1, n = 2, w = 2, $\lambda_1 \leq 3.0$)	2,02º	2,36º	1,6%
Singh (passo = 2, n = 2, w = 2)	45,16º	21,10º	100%
Singh (passo = 2, n = 2, w = 2, $\lambda_1 \leq 0.1$)	46,12º	18,64º	81,9%
Heeger	6,16º	4,02º	29,3%
Waxman et al. $\sigma_f = 1.5$	8,78º	4,71º	1,1%
Fleet and Jepson $\tau = 1.25$	0,07º	0,02º	2.2%
Fleet and Jepson $\tau = 2.50$	0,18º	0,13º	12,6%

Fonte: Barron et al., 1995.

No caso do espaço bidimensional (2D) e acrescentando a variável de tempo t, uma célula localizada em (x, y, t) com intensidade I(x, y, t) terá se deslocado Δx, Δy e Δt entre dois quadros consecutivos, e assumindo que estes deslocamentos sejam pequenos, e impondo a primeira restrição de que as intensidades serão rigorosamente iguais nestes dois instantes, tem-se a Equação 4:

$$I(x, y, t) = I(x+\Delta x, y+\Delta y, t+\Delta t) \quad (4)$$

Fazendo-se a aproximação pela Série de Taylor:

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \frac{1}{1!} * (\frac{\delta I\, \Delta x}{\delta x} + \frac{\delta I\, \Delta y}{\delta y} + \frac{\delta I\, \Delta t}{\delta t}) + \ldots \quad (5)$$

A expansão poderia continuar até termos de maior ordem (derivadas segunda, terceira e assim sucessivamente), mas sendo a análise restrita a dois quadros fica impossível obter-se as derivadas segundas (aceleração) e de ordem superior.

Da restrição imposta de que a intensidade não varia muito entre os instantes t e t+Δt, tem-se:

$$\frac{\delta I\, \Delta x}{\delta x} + \frac{\delta I\, \Delta y}{\delta y} + \frac{\delta I\, \Delta t}{\delta t} = 0 \quad (6)$$

Ou ainda, dividindo-se tudo por Δt:

$$\frac{\delta I\, \Delta x}{\delta x \Delta t} + \frac{\delta I\, \Delta y}{\delta y \Delta t} + \frac{\delta I\, \Delta t}{\delta t \Delta t} = 0 \quad (7)$$

Reescrevendo as parciais $\frac{\Delta x}{\Delta t}$ como Vx e $\frac{\Delta y}{\Delta t}$ como Vy, ou seja, Vx e Vy representando as componentes de velocidade nas direções x e y respectivamente:

$$\frac{\delta I\, Vx}{\delta x} + \frac{\delta I\, Vy}{\delta y} + \frac{\delta I}{\delta t} = 0 \quad (8)$$

Finalmente, de uma forma mais conveniente:

$$\nabla I^{T} * \mathbf{V} = -\frac{\delta I}{\delta t} \quad (9)$$

Este impasse no cálculo do fluxo óptico, onde se tem uma única equação (Equação 9) com duas incógnitas é conhecido como problema da abertura. Então cada pesquisador aborda o problema com métodos, particularidades e restrições diferentes para medir o fluxo óptico.

Por exemplo, o método proposto por Lucas-Kanade, assume que o fluxo óptico é essencialmente constante em torno da vizinhança do ponto sob análise e resolve as equações para todos os pixels nesta

vizinhança usando o critério do menor erro quadrático. Além do mais este método pode ser enquadrado na categoria dos esparsos, haja vista que esta análise não é feita em todos os pontos da imagem, mas tão somente em alguns pontos de interesse pré-selecionados [LUCAS-KANADE 1981].

Por outro lado, o método de Farneback (2000) para estimar o campo vetorial enquadra-se na categoria dos capazes de gerar fluxos ópticos densos (praticamente para todos os pontos da imagem sob análise). Ao invés de tentar encontrar regiões com movimentação coerente tentando mensurar a forma como a intensidade dos pixels está se deslocando em quadros adjacentes, para calcular campos de fluxo óptico densos, este método parte do conceito de minimização da distância ou erro para um modelo de movimentação em torno de cada pixel, levando-se em conta sua vizinhança de forma que haja convergência entre o modelo de movimentação do pixel analisado e o modelo de movimento de seus vizinhos.

2.4 EXTRAÇÃO DE CARACTERÍSTICAS

Reconhecimento de padrões, segundo Duda et al. (1997), é o ato de assimilar e entender dados brutos e tomar alguma decisão baseada na categoria do padrão observado. O ser humano faz isto de certa forma tão automática que até parece ser uma tarefa simples, quando de fato atuam sofisticados sistemas neurais e cognitivos. Por exemplo, desenvolver um algoritmo para separar documentos contendo imagens de pessoas por gênero masculino ou feminino é ainda um desafio.

Em sistemas computacionais programados para reconhecer padrões existem as seguintes subtarefas: (1) captura da imagem, seguida de (2) pré-processamento para simplificação sem perda significativa de informação. Nesta fase a segmentação desempenha papel primordial, na separação dos objetos de interesse entre si e do fundo. Na sequência, (3) extração de características que é o processo de se mensurar determinadas propriedades, cujo objetivo é a redução de dados e transformá-los em grandezas numéricas ou processáveis por uma máquina e finalmente (4) classificação, onde os dados característicos obtidos são avaliados e a decisão sobre o tipo de padrão é definida. As máquinas de vetores suportes (vide seção 2.7.1), do inglês *Support Vector Machine* (SVM) são uma das formas de se obter a classificação, ou seja, dado um conjunto de informações na entrada, determinar o tipo

de padrão na saída. Nesta seção será abordado o uso de SIFT e SURF para localização de objetos.

2.4.1 SIFT

Em 2004, David Lowe propôs um método para transformar dados de imagens em coordenadas invariantes à escala e relacionadas com características locais. O método procura extrair características invariantes e distintas das imagens, características estas que possam ser usadas para conseguir combinação confiável de objetos ou cenas gravados por diferentes pontos de observação (perspectiva). Por ser invariante à escala, o método foi chamado de Transformada de Características Invariantes à Escala, do inglês *Scale-Invariante Features Transform* (SIFT). Entretanto, a transformada SIFT mostra-se invariante também à rotação e com desempenho robusto para transformações afins, mudança do ponto de observação 3D, adição de ruídos e mudanças na iluminação. Outra propriedade marcante é que as características são altamente distintas entre si, ou seja, tem alto poder de discriminação, no sentido de que uma característica tem alta probabilidade de ser corretamente localizada em um banco de dados de características de várias imagens diferentes. A transformada SIFT proporcionou um passo importante na representação de objetos através de suas partes, gerando um avanço em áreas como reconhecimento de objetos e rastreamento.

A obtenção de descritores SIFT segue as seguintes etapas:

1. **Detecção de extremos no domínio espaço-escala**: busca por máximos e mínimos relativos em várias escalas e diversos locais da imagem. O espaço-escala é definido como uma função L(x,y,σ), que resulta da convolução de uma função Gaussiana G(x,y,σ) com a imagem I(x,y) conforme Equação 10:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \quad (10)$$

onde $G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(\frac{x^2+y^2}{2\sigma^2})}$ (11). A detecção estável dos extremos [LOWE 1999] se dá através da busca no espaço-escala D(x,y,σ) resultante da função Diferença de Gaussianas (DoG), entre duas escalas próximas e separadas por um fator escalar k e convoluídas com a imagem (Equação 12):

$$D(x,y,\sigma) = [G(x,y,k\sigma) - G(x,y,\sigma)] * I(x,y) \quad (12)$$

$$\text{Ou ainda, } D(x,y,\sigma) = L(x,y,k\sigma) - L(x,y,\sigma) \quad (13)$$

A determinação dos máximos e mínimos relativos é feita através da análise de 26 vizinhos, sendo 8 da escala atual (região 3x3 pixels) e os 18 das duas escalas adjacentes, sendo 9 de uma escala (fator de amplitude k) acima e 9 de uma escala abaixo.

2. **Localização dos pontos chaves:** em cada região candidata, um modelo detalhado é ajustado para determinar localização e escala. Os pontos chaves são escolhidos de acordo com sua estabilidade, ou seja, são escolhidos pontos que aparecem repetidamente e que representam a mesma característica em diferentes escalas k da imagem. Nesta fase são usadas técnicas de interpolação para determinar com precisão a localização dos pontos de interesse. Outros fatores, tais como pontos de baixo contraste e influência das bordas são minimizados neste processo. Por exemplo, é desejável que na determinação das características de uma porta, somente dados internos a ela sejam relevantes, pois se o caixilho for considerado como parte da porta, pode-se ter dificuldade em obter correspondências entre portas abertas ou fechadas.

3. **Definição da Orientação:** uma ou mais orientações são atribuídas para cada ponto chave com base nas direções dos gradientes locais da imagem. Todas as operações, a partir desse momento, passam a ser executadas em dados da imagem transformados em relação a orientação, escala e localização de cada característica, providenciando invariância a tais transformações.

4. **Descritor de pontos chave:** Nesta etapa é feita a construção dos descritores ao se medir gradientes locais em uma região vizinha a cada ponto de interesse. Estas medidas são então transformadas para uma representação que permite níveis significativos de distorção e mudança na iluminação. Para cada ponto de interesse, são definidas N x N regiões, com k x k pixels cada, ao redor da localização do ponto chave. Geralmente n = k = 4. Para cada região, é feito um

histograma em 8 direções. Este histograma é feito com as magnitudes dos pixels pertencentes a cada região. O descritor é então representado pelos histogramas das regiões.

2.4.2 SURF

O método SURF é utilizado para se obter um conjunto de vetores de características distintas e invariantes às transformadas afins como rotação, translação e escala [JUAN et al. 2009].

Bay et al. (2008) propuseram uma versão relaxada do operador DoG na qual *wavelets* de Haar são usadas para calcular uma aproximação das derivadas de segunda ordem do núcleo gaussiano. Essa aproximação foi usada pelos autores para a construção do método SURF (*Speeded Up Robust Features*). De fato, a forma dessas derivadas é muito similar às usadas no trabalho de Viola e Jones (2001).

A detecção de pontos-chaves do método SURF explora o uso de imagens integrais para calcular uma aproximação do operador DoG em diferentes escalas, o que lhe confere um desempenho de 3 a 7 vezes melhor do que o apresentado no SIFT [BAY et al. 2008]. Como o operador DoG apresenta fortes respostas nos cantos e junções, o número de pontos-chaves detectados pelo SURF geralmente é bem menor do que os reportados pelos operadores Laplacian of Gaussian (LoG) ou DoG. Apesar disso, segundo Bay et al. (2008) o método SURF reporta pontos-chaves tão estáveis quanto aqueles encontrados pelo SIFT. Tanto SIFT quanto SURF são largamente utilizados para detecção e/ou localização de objetos, podendo também serem úteis no rastreamento de pessoas como será visto na próxima seção.

2.5 RASTREAMENTO DE OBJETOS POR CARACTERÍSTICAS

Quando o objetivo, além de se detectar pessoas, é reconhecer individualmente cada uma das pessoas detectadas através de um rótulo e segui-las durante toda sua trajetória, então técnicas de identificação de características como SIFT, SURF e Partícula de Vídeo são necessárias.

É possível ainda rastrear-se objetos fazendo-se uso de SURF, conforme Miao et al. (2011). A ideia central desta abordagem é utilizar pontos característicos extraídos utilizando-se SURF para identificar os objetos, e de forma dinâmica e adaptativa continuar atualizando os classificadores responsáveis por identificar as correspondências das

características SURF entre quadros consecutivos. Aquelas características negativas (que não pertencem ao objeto sendo rastreado, ou seja, que pertencem ao conjunto de imagens negativas usadas durante o treinamento), próximas ao hiperplano de separação da SVM, recebem um peso mais forte com o objetivo de aumentar o poder discriminativo. A Figura 6 mostra que mesmo após a oclusão quase que total do objeto de interesse, sua rastreabilidade foi mantida.

Figura 6 – Comparativo entre métodos de rastreamento

(a) Resultados do Rastreamento utilizando-se apenas SURF

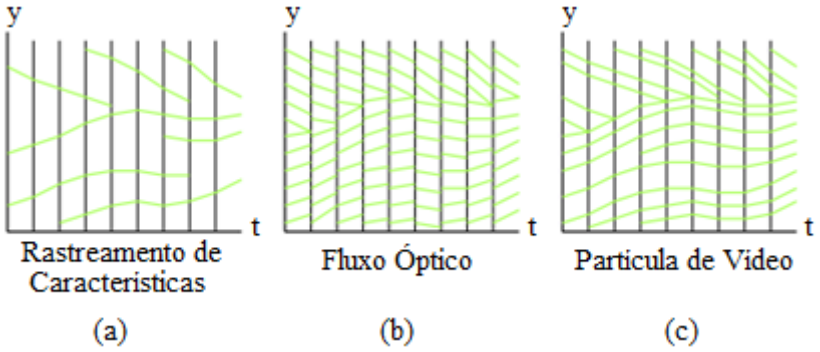
(b) Resultados do Rastreamento pelo Método de Miao et al. [2011]

Fonte: Miao et al., 2011.

Outra forma de estimativa é a proposta por Sand et al. (2008). Trata-se da partícula de vídeo. Esta técnica combina as duas vantagens inerentes ao rastreamento de características (como o SURF, por exemplo) e o rastreamento do fluxo óptico. As estimativas de movimento usando o rastreamento de características costumam ser de longo alcance e esparsas, enquanto as estimativas usando o fluxo óptico são densas e de curto alcance (Figura 7). No método da partícula de vídeo, o objetivo é saber para qualquer ponto da imagem, onde ele aparece em um quadro subsequente, caso o ponto esteja dentro do campo de visão e não ocluso, pretendendo ser densa e de longo alcance. A Figura 7 destaca as propriedades de cada um dos três métodos. Por exemplo, na Figura 7 (a) o rastreamento de características mostra poucos pontos sendo seguidos, mas é possível identificá-los de forma única vários quadros após a primeira detecção, e, portanto, esta técnica pode ser classificada como tendo longo alcance, mas esparsa (apenas alguns pontos são rastreados). Já em (b), o fluxo óptico consegue rastrear uma quantidade bem maior de pontos, mas sem identificá-los de forma inequívoca além de dois quadros consecutivos, podendo, portanto, ser classificado como denso, mas de curto alcance. Finalmente

em (c), o método da partícula de vídeo une as duas características desejáveis dos dois métodos anteriores de rastreamento, sendo ao mesmo tempo densa e de longo alcance.

Figura 7 – Densidade e alcance entre sequências hipotéticas



Fonte: Sand et al., 2008.

2.6 DETECÇÃO DE ELIPSES E CABEÇAS

Um dos problemas na detecção de pessoas são as oclusões. Elas ocorrem quando parte ou a totalidade da pessoa fica sobreposta por algum outro objeto presente na cena. No caso das intrusões via portões de garagens, as pessoas poderiam tentar esconder-se atrás dos carros, por exemplo. Segundo Dóllar et al. (2012), as detecções degradam bastante na presença de oclusões, ou seja, a taxa de acertos é afetada quando parte do objeto procurado é sobreposta por outro objeto e quanto maior a oclusão, pior é o desempenho da classificação.

Uma alternativa para se lidar com oclusões é tentar detectar pessoas a partir da localização de suas cabeças ou busto. Utilizando-se mais de uma técnica para detectar a presença de pessoas em uma cena, torna a detecção mais confiável.

Tomando-se uma determinada imagem representada, por exemplo, por uma reta. Para encontrar o conjunto de pontos que pertençam a esta curva, no caso uma reta, poderia-se encontrar todos os segmentos de retas formados entre cada par de pontos e obter o subconjunto de pontos que estejam próximos dos segmentos de retas. Este procedimento seria altamente inviável, quase impossível para imagens complexas, devido ao custo computacional envolvido.

Para contornar este problema, utiliza-se a Transformada de Hough. Em 1962, Paul Hough desenvolveu e patenteou uma técnica matemática para a detecção de formas geométricas em imagens [ROUGH 1962]. A técnica foi estendida posteriormente para círculos e elipses por Duda e Hart fazendo uso de coordenadas polares [DUDA et al. 1972]. Posteriormente, em 1981, Ballard generalizou a transformada de Hough para detecção de formas arbitrárias [BALLARD 1981].
Geralmente utiliza-se a Transformada de Hough após a imagem ser pré-processada, normalmente pela detecção de bordas, limiarização ou outra técnica. Após esta etapa preparatória, aplica-se na imagem uma transformação de parâmetros, de forma que todos os pontos pertencentes à imagem no domínio espacial sejam mapeados num único ponto no novo espaço, conhecido como domínio de Hough. Veja na Figura 8 o esquema das etapas para aplicação da Transformada de Hough.

Figura 8 – Esquema das etapas para aplicação da transformada de Hough

Fonte: adaptado de Duarte, 2003

Uma vantagem singular da Transformada de Hough é o fato de ser praticamente imune a ruídos e até mesmo oclusões.

Hough desenvolveu esta técnica trabalhando com parametrização da equação de reta convencional (Figura 9 a) na forma *slope-intercept*, ou seja, inclinação-intersecção (Figura 9 b):

$$y = ax + b \tag{14}$$

Onde os parâmetros: a corresponde ao coeficiente angular da reta e y é o valor que a reta cruza no eixo das ordenadas.

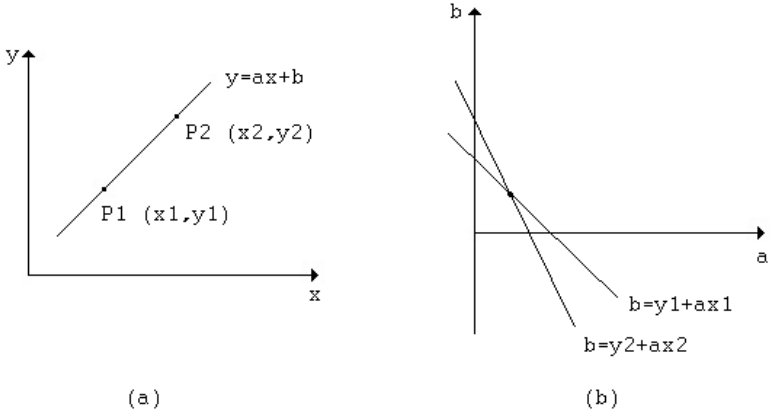
Segundo um dos postulados sobre retas: por um ponto passam infinitas retas. Assim, para um determinado ponto 1 localizado no plano cartesiano xy em P1 (x_1,y_1) passam infinitas retas. Elas obedecem à relação $y_1 = ax_1 + b$, para a e b diversos. Pode-se raciocinar da mesma forma para um ponto 2 no plano xy, descrito por P2 (x_2,y_2), que obedece à $y_2 = ax_2 + b$, também para a e b diversos.

Reescrevendo a Equação 14 isolando o termo b, tem-se:

$$b = y - ax \tag{15}$$

Então pode-se traçar os gráficos no domínio ab (chamado de domínio dos parâmetros, ou ainda, espaço de Hough), para os pontos 1 e 2. Cada ponto originará uma reta neste novo domínio. Onde elas forem concorrentes, tem-se os valores dos parâmetros a e b que, potencialmente, descrevem a equação de reta original que se desejava encontrar (Figura 9).

Figura 9 – Transformada de Hough (a) Reta no domínio xy; (b) Espaço de Hough



Fonte: adaptado de Pedrini et al., 2008

Porém a abordagem original de Hough (forma inclinação-intersecção) não funciona para retas muito verticais, onde a inclinação tende ao infinito. Duda e Hart (DUDA & HART, 1972) eliminaram este problema, utilizando a Transformada de Hough a partir da parametrização da reta por coordenadas polares. Esta técnica derivada atende a qualquer tipo de reta.

Seja a equação de reta em coordenadas polares:

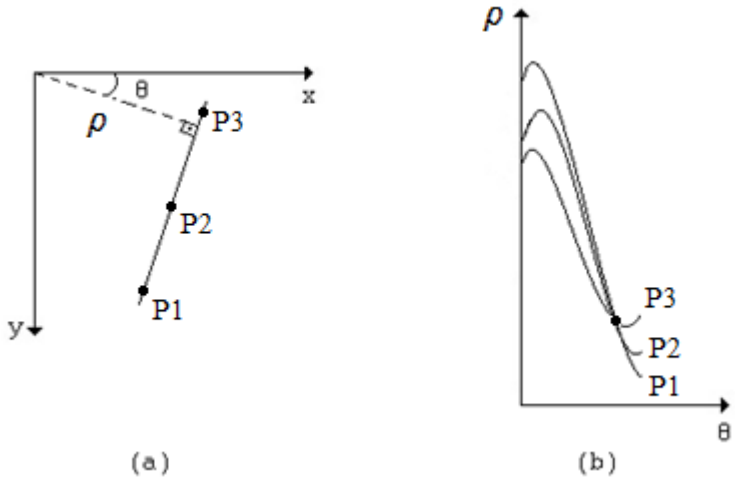
$$\rho = x\cos\theta + y\text{sen}\theta \qquad (16)$$

onde ρ é a distância normal da origem à reta e θ o ângulo formado entre ρ e o eixo das abscissas.

Os novos parâmetros agora são ρ e θ, que formam o novo espaço de Hough (ρ,θ). Os pontos no domínio (x,y) são representados no espaço de Hough por senoides, enquanto que no tratamento anterior era representado por retas. Se o tamanho da imagem é RxS pixels, o

tamanho máximo de ρ será $\sqrt{R^2 + S^2}$, a medida da diagonal deste quadro. Já θ varia de 0 a 180°. O ponto onde essas senoides se cruzam traduzem, potencialmente, valores de ρ e θ da reta original em (x,y). Cada cruzamento traz uma possibilidade de reta. (Figura 10)

Figura 10 – Transformada de Hough (a) Reta em parametrização polar; (b) Espaço de Hough



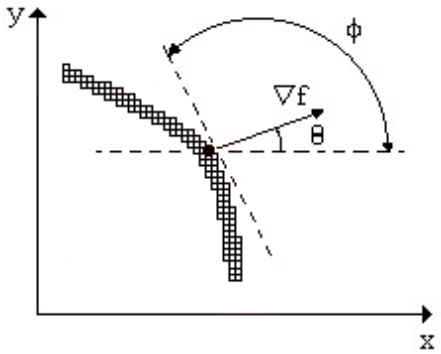
Fonte: adaptado de Pedrini et al., 2008

Discretiza-se o espaço de Hough em células, chamadas células de acumulação, inicialmente zeradas. Para cada ponto em (x,y) calculam-se os valores de ρ e θ que satisfazem a equação da reta e incrementa-se em 1 o acumulador. Após a determinação dos parâmetros de todos os pontos do domínio, os pontos de maior acúmulo no espaço discretizado, indicam forte evidência de retas na imagem. Conforme existam vários picos (máximos dos acumuladores), com estes determinam-se as retas correspondentes e esses picos são eliminados. Uma nova análise é feita e os novos picos restantes escrevem as novas retas, e assim também são removidos. Desta forma a imagem se desenha.

Uma das vantagens da Transformada de Hough é que ela funciona bem até mesmo para regiões obstruídas ou com ruídos. Regiões obstruídas terão como consequência apenas a diminuição do número de acumulações. Com relação aos ruídos, a transformada não é afetada, pois essas deformações não serão mapeadas em uma mesma célula de acumulação.

Como a Transformada de Hough depende do número de dimensões (parâmetros), o processo computacional pode ser dificultado quando há curvas mais complexas. Devido a esse alto custo computacional, lança-se mão do uso do gradiente. A ideia de gradiente já é utilizada para detecção de bordas e no caso da Transformada de Hough a direção do gradiente coincide com θ. Ao invés de calcular os parâmetros para todos os pontos da imagem, apenas são calculados os parâmetros para os pontos das bordas, economizando assim, processo computacional, conforme pode ser visto para o *pixel* de borda destacado na Figura 11, cujo valor do gradiente ultrapassa um valor de gatilho predefinido.

Figura 11 – Relação entre a direção θ do gradiente e a direção da tangente φ em uma curva



Fonte: adaptado de Pedrini et al., 2008

Para detectar círculos, caso seja utilizada a fórmula da circunferência em coordenadas cartesianas (Equação 17), onde (a,b) é o centro da circunferência de raio r, a estrutura de acumulação de parâmetros passa a ser tridimensional, dificultando o processamento envolvido.

$$(x - a)^2 + (y - b)^2 = r^2 \quad (17)$$

Reescrevendo a equação da circunferência parametrizada em coordenadas polares, tem-se:

$$x = a + r \cos \theta \quad (18)$$

$$y = b + r \operatorname{sen} \theta \quad (19)$$

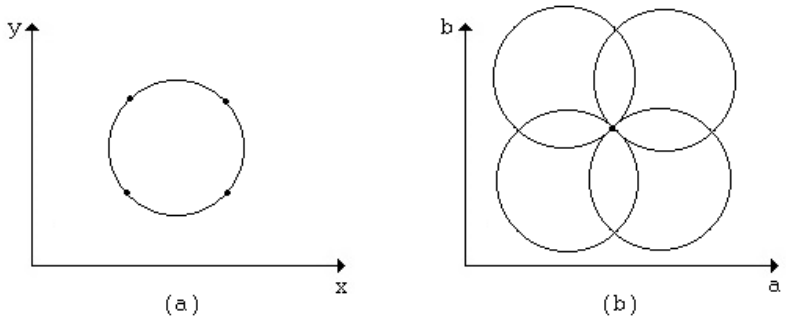
Onde (a,b) é o centro da circunferência de raio r. Considerando-se que se conheçam os raios r dos círculos que se desejam, pode-se agrupar as Equações 18 e 19 desta forma:

$$b = a\,tg\theta - x\,tg\theta + y \quad (20)$$

Usando a ideia do gradiente, calcula-se a direção do ângulo θ. A partir disto, pode-se votar nas células de acumulação de acordo com a Equação 20. Os picos de votação nas células do espaço dos parâmetros indicam que, muito provavelmente, haja um círculo cujo centro no domínio da imagem possua estes parâmetros. (Figura 12)

Assim como pontos das retas no domínio (x,y) criam senoides no espaço de Hough, de forma paralela, pontos dos círculos no domínio (x,y) geram círculos no espaço de Hough. Para cada ponto no domínio (x,y), tem-se um círculo no espaço de parâmetros, onde a intersecção deles indica a posição do círculo da imagem.

Figura 12 – Transformada de Hough (a) Domínio xy; (b) Espaço de Hough

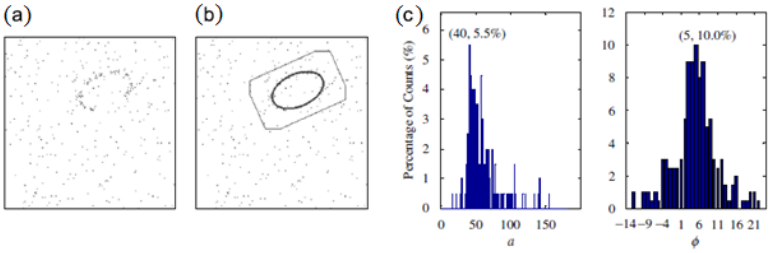


Fonte: Pedrini et al., 2008

A detecção de elipses incompletas pode ser uma ferramenta importante no contexto deste trabalho. Lu et al. (2008) criaram um algoritmo para detecção de elipses incompletas em imagens ruidosas utilizando iterativa e randomicamente a transformada de Hough (IRHT). O método IRHT traz uma melhoria, que é justamente apresentar um processo iterativo onde os pontos de ruído são gradualmente excluídos da região de solução e interesse. Na Figura 13 é mostrada a aplicação da IRHT na detecção de uma elipse borrada. Em (a) tem-se a elipse incompleta composta por 94 *pixels* corrompida por ruído gaussiano e

submersa em um fundo composto de 282 *pixels* de ruído impulsivo, enquanto em (b) é apresentado o resultado da IRHT para figura (a). Finalmente em (c) tem-se os histogramas resultantes após finalização da IRHT.

Figura 13 – IRHT Detectando traços de elipse imersa em ruído

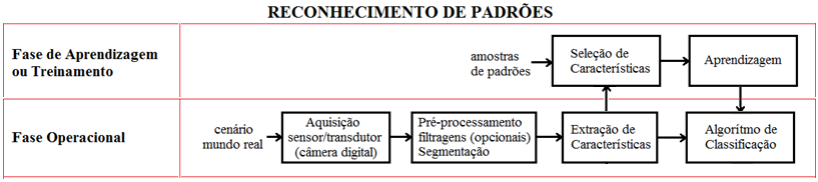


Fonte: Lu et al., 2007.

2.7 CLASSIFICADORES

O reconhecimento de padrões abrange os seguintes processos principais: (a) aquisição da imagem; (b) pré-processamento, onde filtragens são opcionais para simplificar a informação. A segmentação é o principal processo desta fase; (c) extração das características e (d) aplicação de algoritmo de classificação (Figura 14).

Figura 14 – Fases do reconhecimento de padrões



Fonte: produção do próprio autor

Nesta seção será abordado o princípio de funcionamento das Máquinas de Vetores Suportes (SVM), bem como do classificador baseado no método ingênuo de Bayes (do inglês *Naive Bayes*) e uma breve explicação das redes neurais artificiais Multilayer Perceptrons.

2.7.1 Máquinas de Vetores Suporte

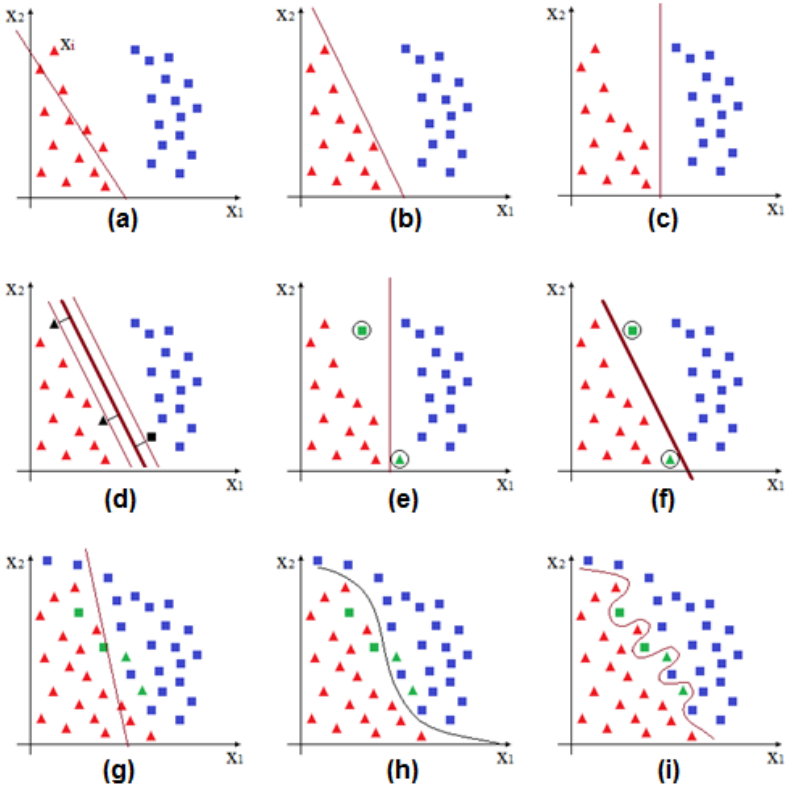
No processo de reconhecimento de padrões, as características extraídas e selecionadas podem ser usadas para classificação. As máquinas de vetores suportes, do inglês *Support Vector Machine* (SVM), podem ser definidas como modelos de algoritmos de aprendizagem supervisionada capazes de analisar dados e reconhecer padrões.

A principal motivação que levou Vladmir Vapnik em 1979 a desenvolver a teoria da aprendizagem estatística foi melhorar os problemas relacionados com os super e sub ajustes: seja um conjunto de dados usados para treinar uma SVM. Diz-se que a capacidade da SVM está super ajustada (do inglês *overfitting*) quando ela praticamente memoriza rigorosamente a forma de separar os grupos com base nos dados usados para treino. Por exemplo, um chaveiro que foi treinado para reconhecer chaves usando como informação a quantidade de ranhuras entre 3 a 8 e que memorizou isto de forma rígida, ao ser apresentado a uma chave com 9 dentes não a reconhecerá como tal. Trata-se do caso de super aprendizagem. Supondo que este chaveiro tenha um aprendiz, que aprendeu por si só que as chaves são metálicas. Para o aprendiz, qualquer metal é uma chave. Neste caso trata-se da sub aprendizagem. O ideal é um equilíbrio entre a acurácia baseada nos dados de treino e a capacidade de generalização [BURGERS 1998].

Em princípio, as SVM são usadas para problemas de classificação binária, ou seja, que envolvam apenas dois grupos diferentes, mas há formas de adaptá-las para classificar mais de duas categorias diferentes.

Na Figura 15 (a) é possível perceber que a linha divisora não é capaz de separar corretamente as classes triângulos e quadrados. Já nas Figuras 15 (b) e (c) tem-se a separação correta, mas então surge a pergunta: qual delas separa melhor as classes? Assim como estes dois últimos exemplos, poderia-se ter uma quantidade enorme de retas para separar linearmente as classes. Em (d) é apresentado o caso ideal de separação, envolvendo o conceito dos vetores suportes e margens máximas de separação. Em (e) e (f) pode-se notar a nítida vantagem quando as margens de separação são maximizadas em relação aos vetores suportes, pois em (e) os valores atípicos (do inglês *outliers* e que aparecem circulados) não são classificados corretamente enquanto em (f), o são. Finalmente nota-se em (g) o caso de sub capacidade ou sub generalização, em (h) o caso de generalização ideal e em (i) o caso de super generalização ou memorização.

Figura 15 – Separação entre classes e capacidade de generalização



Fonte: produção do próprio autor

Dado um conjunto de dados usados para treinamento D, composto de n pontos na forma:

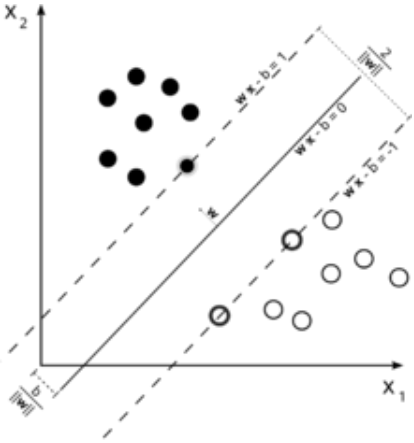
$$D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n}$$

Onde os pontos y_i assumem valor -1 ou +1 e indicam a classe a qual os pontos x_i pertençam (problema de separação binário). Cada x_i é de fato um vetor real p-dimensional. O objetivo é encontrar um hiperplano que separe os pontos pertencentes à classe y=-1 dos da classe y=+1 com a maior margem possível.

Supondo que os dados de treinamento D sejam linearmente separáveis, pode-se escolher dois hiperplanos de maneira que eles separem as duas classes e que não haja nenhum ponto x_i entre eles e

então maximizar esta distância entre os dois hiperplanos limítrofes (Figura 16). A região delimitada pelos hiperplanos também é chamada de margem.

Figura 16 – Maximização das distâncias entre hiperplanos



Fonte: Cortes e Vapnik, 1995

Os hiperplanos podem ser descritos pelas Equações 21:

$$w \bullet x - b = -1 \qquad e \qquad w \bullet x - b = 1 \tag{21}$$

A distância entre os dois hiperplanos é $\frac{2}{||w||}$ e como o propósito é a maximização da distância entre eles, o problema torna-se minimizar ||w||. Como é necessário evitar que pontos x_i comecem a avançar na região de margem, as seguintes restrições precisam ser preservadas para todo e qualquer i:

$$w \bullet x - b \geq 1 \text{ para } x_i \text{ que pertençam à } 1^a \text{ classe } (y_i=1) \tag{22}$$

ou

$$w \bullet x - b \leq -1 \text{ para } x_i \text{ que pertençam à } 2^a \text{ classe } (y_i=-1) \tag{23}$$

As Equações 22 e 23 podem ser reunidas da seguinte forma:

$$y_i(w \bullet x - b) \geq 1 \text{ para } 1 \leq i \leq n \tag{24}$$

Resumindo, tem-se um problema de minimização de ||w||, sendo w e b variáveis e restritas à Equação 24.

Em 1995, Cortes e Vapnik sugeriram uma modificação cuja ideia é permitir que amostras atípicas fossem acomodadas por um hiperplano fronteiriço que se adapte a elas. Assim, embora algumas classes que não sejam idealmente linearmente separáveis, mas que tendam a separação linear, possam ser classificadas por SVM através do ajuste do parâmetro C (Ver Equações 25 e 26)

$$w \bullet x - b \geq 1 - \xi_i \quad \text{para } y_i=1 \tag{25}$$

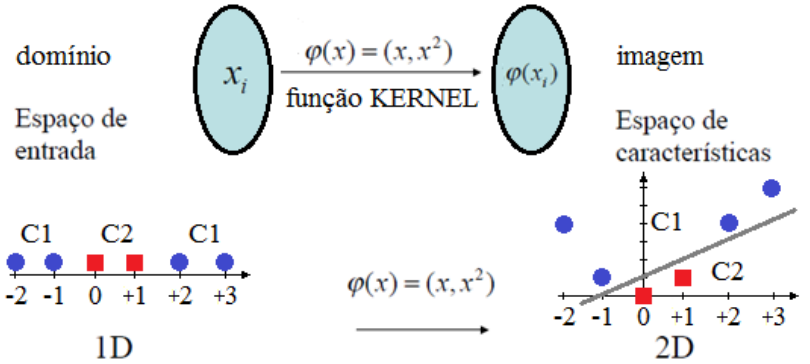
ou

$$w \bullet x - b \leq -1 - \xi_i \quad \text{para } y_i=-1 \tag{26}$$

Resolvendo para as restrições impostas pelas Equações 25 e 26, a nova função de minimização torna-se $\frac{||w||^2}{2} + C * \sum(\xi i)^k$ na qual C é um parâmetro a ser definido pelo usuário e quanto maior o valor de C, maior a penalidade para erros. Em outras palavras, quanto maior o valor de C, mais vetores serão tratados como vetores suportes.

Uma grande parte dos problemas reais não são possíveis de serem separados pelos classificadores lineares. A solução é projetar os dados originais em um novo espaço onde os dados sejam linearmente separáveis. Quando se conhece bem os dados de entrada e o domínio é unidimensional, é relativamente fácil achar uma função de mapeamento. Mas quando se está diante de espaços multidimensionais, esta transformação (conforme proposta na Figura 17) não é tão imediata. Este mecanismo de mapear um espaço de entrada em um espaço de características através de uma função φ() é chamada de truque de Kernel.

Figura 17 – Mapeamento de funções – Kernel Trick



Fonte: produção do próprio autor

Os primeiros Kernels a serem testados foram:

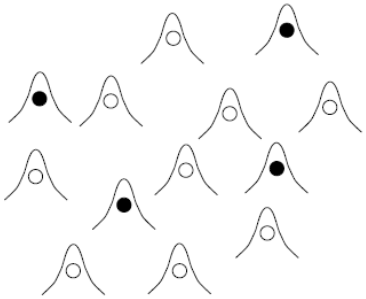
$K(x,y) = (x \bullet y + 1)^p$ polinomial

$K(x,y) = e^{-\frac{||x-y||^2}{2\sigma^2}}$ gaussiano

$K(x,y) = \tanh(k*x \bullet y - \delta)$ tangente hiperbólico

Para o caso do kernel gaussiano, o mapeamento é capaz de reproduzir um espaço n-dimensional em um espaço de infinitas dimensões. Considerando-se SVM de comprimento suficientemente pequeno, se comparadas às distâncias entre os pontos de treinamento, então tem-se o caso limite em que cada ponto é considerado um vetor suporte (Figura 18).

Figura 18 – Kernel gaussiano



Fonte: Cortes e Vapnik, 1995

Embora por definição as SVM são classificadores binários, elas podem ser adaptadas para separar multiclasses. Seja N o número de classes a serem separadas. Uma combinação simples de N treinamentos um-contra-todos (do tipo "classe um" representando os positivos e as "classes restantes", os negativos) é capaz de resolver o impasse da limitação de separações binárias. Após o treinamento, uma amostra de teste pode ser classificada segundo o critério de votação, qual seja, a saída positiva de maior valor (aquela mais próxima de 1) é a categoria na qual a mostra se enquadra.

2.7.2 Teorema de Bayes e Classificador Naive Bayes

A abordagem proposta neste trabalho consiste em se medir a persistência com que pessoas são detectadas em sucessivos quadros de

vídeo para gerar um alarme de intrusão. O teorema de Bayes serve de embasamento para justificar a escolha do número de detecções e ausência de detecções necessárias para caracterizar a ocorrência do evento de intrusão.

Segundo a teoria de Bayes, que deriva da teoria das probabilidades, uma hipótese inicial para um evento é uma probabilidade *a priori*, que é afetada por novas condições, gerando uma probabilidade *a posteriori* [REISSWITZ 2009]. Em termos de equação, o teorema de Bayes é descrito assim:

$$P(A|B) = \frac{P(A) \bullet P(B|A)}{P(B)} \qquad (27)$$

onde: P(A) e P(B) são as probabilidades *a priori* dos eventos A e B acontecer, respectivamente; P(A|B) é a probabilidade *a posteriori* do evento A ocorrer, sabendo que o evento B ocorreu; P(B|A) é a probabilidade *a posteriori* do evento B ocorrer, sabendo que o evento A ocorreu. Ou seja, os resultados servem como realimentação no processo de probabilidades.

Partindo do teorema de Bayes é possível projetar algoritmos de aprendizagem, ou seja, dado um problema de aprendizagem supervisionada, deseja-se reduzi-lo a uma função de aproximação objetiva desconhecida na forma f : X → Y, que é o mesmo que aproximá-lo à P(Y|X). Supondo que Y é uma variável booleana (lógica) que pode assumir os valores verdadeiro ou falso e X, um vetor contendo n atributos booleanos. Em outras palavras, $X = \langle X_1, X_2, X_3, ..., X_n \rangle$ onde X_i corresponde ao i-ésimo atributo (valor booleano) de X. Aplicando a regra ou teorema bayesiano, a probabilidade $P(Y=y_i|X)$ pode ser representada, segundo Mitchell (2015), por:

$$P(Y=y_i|X=x_k) = \frac{P(X=xk|Y=yi)*P(Y=yi)}{\sum_j P(X=xk|Y=yj)*P(Y=yj)} \qquad (28)$$

onde y_j denota o j-ésimo valor da variável aleatória Y, x_k corresponde ao valor da k-ésima posição do vetor X e finalmente o somatório do denominador é aplicado sobre todos os valores possíveis de Y, em outras palavras, sobre todo conjunto Y. Uma forma de se aprender P(Y|X) é usando o conjunto de dados de treinamento para estimar P(X|Y) e P(Y) e então aplicar as regras da Equação 28 para determinar $P(Y|X=x_k)$ para qualquer nova instância de x_k. Para se estimar corretamente P(X|Y) assumindo que Y possa assumir 2 valores apenas,

verdadeiro ou falso e X, por outro lado é um vetor com n posições, as quais podem assumir cada uma delas os valores verdadeiro ou falso, então é necessário se avaliar $2*(2^n-1)$ combinações para construir um classificador baseado nestas condições. A título de exemplo, se o vetor X é formado por 30 características booleanas, será necessário avaliar mais do que 30 bilhões de parâmetros para se projetar corretamente $P(Y|X=x_k)$ [MITCHEL 2015]. Ainda segundo Mitchell (2015), para que o treinamento de um classificador bayesiano seja possível na maioria dos casos, é necessário reduzir esta complexidade. O classificador bayesiano ingênuo reduz a complexidade assumindo a independência condicional (Equação 29), reduzindo desta forma a quantidade de parâmetros ou termos a serem estimados quando se modela $P(X|Y)$ dos originais $2*(2^n-1)$ para apenas $2*n$. A independência condicional pode ser definida como: dadas as variáveis aleatórias X, Y e Z, diz-se que X é condicionalmente independente de Y dado Z, se e somente se a distribuição de probabilidade que governa X é independente do valor de Y dado Z; ou seja:

$$(\forall i, j, k) P(X=x_i|Y=y_j, Z=z_k) = P(X=x_i|Z=z_k) \quad (29)$$

O algoritmo *Naive Bayes* é um classificador baseado no teorema de Bayes que assume que todos os atributos $X_1...X_n$ são condicionalmente independentes em si, dado Y. No caso de, por exemplo, $X=<X_1, X_2>$ e considerando a restrição imposta pela Equação 29, tem-se:

$$P(X|Y) = P(X_1, X_2|Y) \quad (30)$$
$$= P(X_1|X_2,Y)*P(X_2|Y) \quad (31)$$
$$= P(X_1|Y)*P(X_2|Y) \quad (32)$$

A Equação 31 deriva a partir da Equação 30 devido uma propriedade geral das probabilidades, enquanto a Equação 32 é resultado direto da própria definição da independência condicional. Finalmente, de forma genérica, quando X contém n atributos que sejam condicionalmente independentes uns dos outros dado Y, tem-se:

$$P(X_1...X_n|Y) = \prod_{i=1}^{n} P(Xi|Y) \quad (33)$$

2.7.3 Multilayer Perceptrons

Assim como as SVM, as redes neurais artificiais formadas por Multilayer Perceptrons (MLP) , segundo Osowski (2004), também são

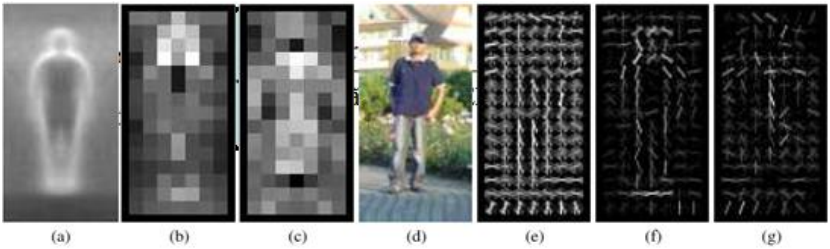
modelos computacionais abstratos para simular o comportamento do cérebro humano, compostas de neurônios artificiais e uma massiva rede de interligação entre eles. Estes neurônios artificiais recebem o nome de perceptron e têm internamente uma função de ativação geralmente do tipo sigmoidal. Nas MLP os perceptrons são dispostos em camadas: camada de entrada formada pelos nós de entrada, seguida pelas camadas escondidas ou internas e finalizadas pela camada de saída. As interconexões são permitidas apenas entre camadas vizinhas. Os sinais são propagados a partir das entradas, percorrendo as camadas intermediárias até gerar os sinais de saída. A rede aprende por um processo de minimização de erros dos pesos das funções de ativação [OSOWSKI 2004].

2.8 HISTOGRAMA DE GRADIENTES ORIENTADOS (HOG)

HOG, associado ao classificador SVM, é um dos algoritmos mais bem-sucedidos na detecção de humanos [PANG et al. 2011]. O algoritmo HOG+SVM, comumente chamado simplesmente por HOG, se concentra no contraste do contorno da silhueta das pessoas com o plano de fundo.

Pessoas diferentes podem ter aparências e roupas diferentes, mas seu contorno será similar. Portanto, o contorno é representativo para distinguir entre pessoas e não-pessoas. É importante notar que os contornos não são diretamente detectados. Os vetores normais dos hiperplanos separadores obtidos pela SVM tem um peso preponderante em determinar as características HOG ao longo do contorno humano. O detector HOG apoia-se principalmente no contorno da silhueta (em especial na cabeça, ombros e pés). Seguidos dos blocos dos contornos estão os blocos que representam o fundo da imagem, sendo mais ativos que os blocos internos às pessoas. Na Figura 19 tem-se em (a) o gradiente médio das imagens usadas para treinamento, em (b) para cada *pixel*, é associado um peso máximo para SVM das imagens positivas (que contêm pessoas). Já em (c) tem-se o equivalente a (b), mas para a SVM das imagens negativas (que não contêm pessoas). Ainda na Figura 19 (d) é apresentada uma imagem de teste, e seu respectivo descritor HOG computado em (e). Após a aplicação dos pesos das SVM positiva (b) e negativa (c) tem-se respectivamente os descritores HOG (f) e (g).

Figura 19 – Influência dos pesos atribuídos pela SVM



Fonte: Dalal et al., 2005.

O método HOG pode ser resumido da seguinte maneira [PANG et al. 2011]:

Entrada: A imagem a ser classificada. O tamanho da janela de detecção deslizante é de 64x128 pixels. O passo de deslocamento d (d=8 por exemplo). Outros tamanhos podem ser usados para a janela deslizante, desde que se mantenha a proporção exata entre os tamanhos das células, dos blocos (conjunto de 4 células) e da própria janela deslizante (conjunto de blocos). Por exemplo, partindo-se de um tamanho de célula de 3x3, obrigatoriamente os blocos teriam que ter dimensão 6x6 e a mínima dimensão para a janela deslizante nesta condição específica seria de 6x6. Qualquer dimensão de janela deslizante tem que acomodar números inteiros de blocos. Além do mais, este tamanho padrão de 64x128 é apropriado para a maioria dos casos e também é ideal para este trabalho. A maioria das bibliotecas, incluindo OpenCV[1], foram treinadas utilizando-se este tamanho de janela deslizante.

Saída: A localização das subimagens de tamanho 64x128 que alegam conter humanos.

Passo 1: Para todo pixel da imagem de entrada, calcular a magnitude $|\nabla f(x,y)|$ e a orientação $\theta(x,y)$ do gradiente $\nabla f(x,y)$. O operador ∇ (nabla) é usado na matemática para denominar o operador diferencial no cálculo vetorial. O resultado da operação $\nabla f(x,y)$, sendo $f(x,y)$ a imagem digital, será um campo vetorial associado a cada ponto da imagem.

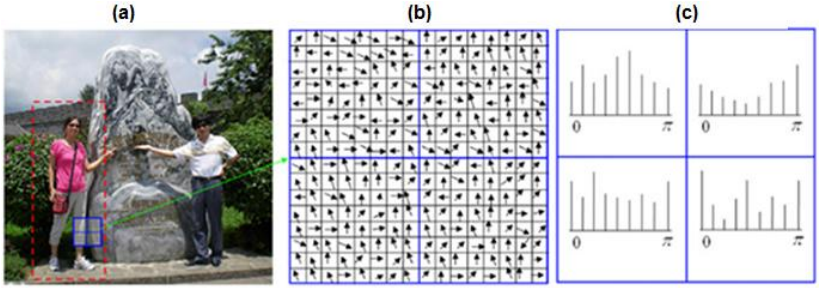
Passo 2: A partir do topo até embaixo e da esquerda para direita, varrer toda imagem com a janela de tamanho 64x128. Extrair as características HOG da subimagem coberta pela janela de detecção e

[1] OpenCV, disponível em http://opencv.org (em 10/Junho/2015)

submetê-las ao classificador SVM para selecioná-la como pessoa ou não-pessoa. A Figura 20 mostra em mais detalhes como o histograma é computado, onde (a) apresenta uma janela de detecção de 64x128 (retângulo maior tracejado que envolve completamente a mulher à esquerda). Em (b) uma ampliação do bloco de dimensão 16x16 composto de 4 células e finalmente em (c) o histograma de gradientes orientados correspondente às 4 células.

Considerando-se uma janela de 64x128 pixels, dividida em células de 8x8 e blocos de 16x16 (conjunto de 4 células) e levando-se em conta uma sobreposição de 50% entre as células, no total teria-se 105 blocos. A cada um destes blocos seriam atribuídos quatro histogramas com 9 subdivisões (de 40 em 40 graus) formados a partir da orientação e intensidade dos gradientes globais, o que resultaria em um vetor de características de 3780 posições.

Figura 20 – Histograma de gradientes orientados densidade e alcance entre sequências hipotéticas



Fonte: Pang et al., 2011.

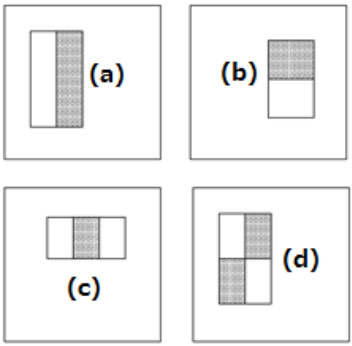
2.9 VIOLA E JONES

O método de detecção de objetos Viola-Jones foi um dos primeiros *frameworks* de detecção de objetos a fornecer taxas de detecção competitivas em tempo real, reconhecendo faces a uma taxa de 15 fotos por segundo, desempenho 15 vezes mais rápido do que o sistema construído em 1998 por Rowley [Viola et al. 2004]. Embora possa ser treinado para detectar uma variedade de classes de objeto, foi motivada principalmente pelo problema de detecção de faces. O recurso empregado pelo *framework* de detecção utiliza as somas dos pixels da imagem com áreas retangulares. Embora tenham alguma semelhança

com transformada de Haar, os recursos utilizados por Viola e Jones contam com mais de uma área retangular. A Figura 21 mostra quatro diferentes tipos de características utilizados no quadro. O valor das características é a soma dos pixels dentro de retângulos claros subtraídos da soma dos pixels dentro de retângulos sombreados. Como é de se esperar, essas características retangulares são bastante primitivas se comparado a alternativas mais modernas. Apesar de serem sensíveis às características verticais e horizontais, o retorno é consideravelmente mais grosseiro. No entanto, com o uso de uma representação global da imagem elas podem ser calculadas em tempo real, o que lhes confere uma vantagem considerável em velocidade diante dos demais.

Como cada área retangular em uma característica é sempre do lado de pelo menos outra, qualquer característica de dois retângulos pode ser calculada em seis referências na matriz, qualquer característica de três retângulos em oito e qualquer característica de quatro retângulos em nove.

Figura 21 – Exemplos de características retangulares



Fonte: Viola-Jones, 2001.

A velocidade com que as características podem ser calculadas podem não compensar seu número. Por exemplo, em uma subjanela de 24x24 pixels, há um total de 162.336 características. Seria proibitivamente caro, do ponto de vista computacional, avaliar todos eles. Assim, o *framework* de detecção de objetos proposto por Viola e Jones utiliza uma variante do algoritmo de aprendizagem *AdaBoost* [FREUND et al. 1997] tanto para selecionar as melhores características quanto para treinar classificadores que utilizam este método. O algoritmo *AdaBoost* refina sucessivamente um classificador fraco adotando técnicas de reforço através de realimentação para impulsionar

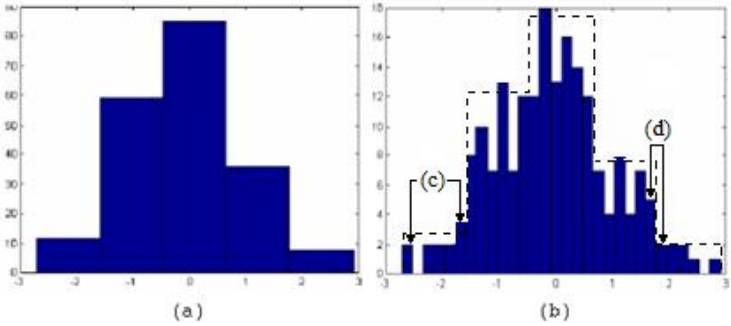
ou fortalecer o classificador. O algoritmo *AdaBoost* é, portanto, composto por vários classificadores em série, sendo que os primeiros estágios se encarregam das classificações mais triviais e apresentam taxas de verdadeiro-positivos bem próximas de 100%, mas também com elevado número de falso-positivos. As imagens falso-positivas do primeiro estágio passam a compor a base de dados negativa para o treinamento do próximo estágio e assim sucessivamente. Uma imagem para ser considerada positiva, precisa passar por toda a cadeia em série dos classificadores. Por outro lado, basta a imagem ser rejeitada por qualquer uma das etapas dos classificadores em série para ser considerada negativa.

2.10 JANELAS DE PARZEN

No presente trabalho, o interesse é em se obter medidas estatísticas que possam ser aplicadas em características observadas no comportamento do fluxo óptico e que possam ser convertidas em entradas para treinamento e utilização de classificadores. As Janelas de Parzen cumprem este papel, obtendo o grau de aderência do alinhamento dos vetores do campo fluxo óptico da região das pernas em relação ao alinhamento dos vetores das regiões superiores, como o tronco. Em pessoas, a tendência é que não haja uma grande correlação entre tais alinhamentos, enquanto para carros, por exemplo, o campo fluxo óptico tende a estar alinhado em todas as regiões do automóvel.

Segundo Pedrini (2008), um método mais antigo e mais simples é o histograma. Nele, o domínio é dividido em intervalos iguais e cada ocorrência é associada ao intervalo que a contém. Porém a largura escolhida para os intervalos e sua origem no espaço definem onde a amostra será classificada, e assim, podem gerar sensíveis diferenças na distribuição de probabilidades. Por exemplo, duas amostras vizinhas podem ser classificadas em diferentes intervalos como destacado na Figura 22 (d), enquanto amostras mais distantes podem ser enquadradas juntas, conforme destacado na Figura 22 (c), quando se considera o tamanho de largura dos intervalos da imagem em (a).

Figura 22 – Histograma (a) Intervalo maior; (b) Intervalo menor



Fonte: adaptado de Pedrini et al., 2008

Mesmo com intervalos mais estreitos, pode-se inviabilizar matematicamente as análises, uma vez que o número de dimensões do espaço afeta exponencialmente o número de intervalos. A descontinuidade gerada pelos histogramas também contribui para a escolha de outro método.

Segundo Duda e Hart (1972), a estimativa da densidade de uma função desconhecida é dada por:

$$p(x) = \frac{(k/n)}{V} \tag{34}$$

onde V é o volume da região analisada, k o número de pontos que caem nesta região R e n, o número total de pontos.

Duas abordagens são possíveis a partir da Equação 34:

- escolhido um valor fixo para k, pode-se determinar V. Este método é conhecido como "Regra do Vizinho Mais Próximo (kNN)";
- escolhido um valor fixo para V, pode-se determinar k. Este é o método das Janelas de Parzen.

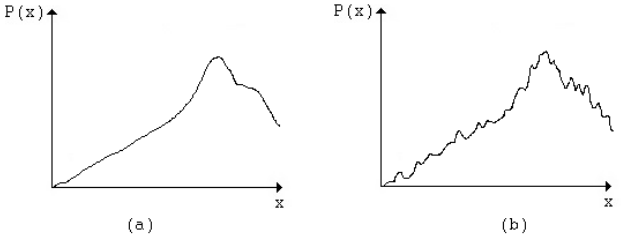
Ao centralizar o volume pré-definido V na região R, define-se uma função K(x) para estimar a quantidade k de pontos x que pertençam à R, sendo os elementos x_i pertencentes ao conjunto extraído. Esta função é conhecida como função Kernel (função núcleo). Assim, tem-se para a probabilidade P(x):

$$P(x) = \left(\frac{1}{n}\right) \sum_{i=1}^{k} K(x - xi) \quad (35)$$

A função ainda preserva a característica de hipercubo e, portanto, há descontinuidade. Para suavizá-la, costuma-se usar a função Gaussiana como Kernel (ver Equação 36). E para evitar o alto custo computacional, é possível usar a variância de 3σ.

$$K(x) = [1/(\sigma\sqrt{2\pi})] \exp[-x^2/(2\sigma^2)] \quad (36)$$

Figura 23 – Distribuições de Probabilidade: (a) Janelas de Parzen Gaussiana mais suave (com variância alta); (b) Janelas de Parzen Gaussiana menos suave (com variância baixa)



Fonte: Pedrini et al., 2008

As Janelas de Parzen são uma alternativa viável ante o histograma, pois elas carregam a influência dos elementos da vizinhança no cálculo da probabilidade, dando relevância para elementos próximos, enquanto que no histograma, se vizinhos ficassem classificados em intervalos diferentes, a relevância seria desconsiderada.

2.11 CURVAS DE CARACTERÍSTICAS OPERACIONAIS DO RECEPTOR

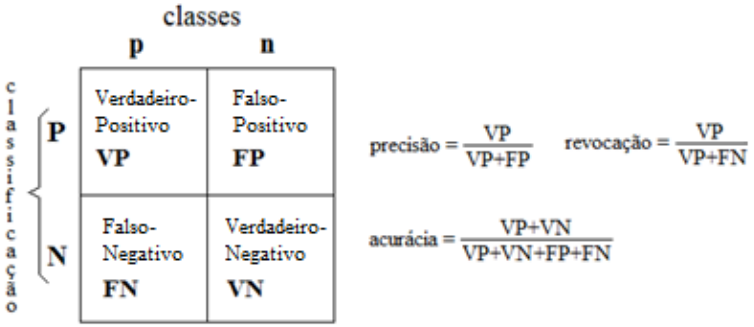
As curvas características operacionais do receptor (do inglês *Receiver Operating Characteristics* ROC) tiveram sua origem na teoria de detecção de sinais para analisar a relação custo benefício entre as taxas de acerto (verdadeiro-positivos) e alarmes falsos (falso-positivos) de classificadores [FAWCETT 2006]. Elas constituem uma forma de visualização importante para se medir o desempenho de um classificador e, portanto, são usadas neste trabalho para avaliar o desempenho dos métodos testados.

Partindo do problema da classificação de duas classes distintas do conjunto {p, n}. Cada instância ou elemento de um conjunto a ser testado pode ser mapeado como amostra positiva p ou negativa n. Um classificador, como as SVM por exemplo, são modelos que mapeiam instâncias em classes preditas, podendo gerar respostas do tipo {P,N}, onde o P indica que a instância pertence à classe p e o N, que não pertence à classe p e consequentemente, pertence à classe n.

Assim, dada uma instância e um classificador, pode-se ter uma das quatro situações seguintes: 1) a instância é positiva {p} e classificada como positiva {P}. Trata-se de um verdadeiro-positivo; 2) se a mesma instância em questão fosse classificada como {N}, teria-se a situação de falso-negativo; 3) Agora, uma instância negativa {n} classificada como {N} gera o verdadeiro-negativo; 4) Finalmente, se esta instância negativa fosse classificada como {P} teria-se o falso-positivo

Dado um classificador e um conjunto de instâncias (conjunto de teste) pode-se montar uma matriz 2x2 chamada de matriz de confusão e também conhecida como matriz de contingência. A diagonal principal traz os acertos decisórios do classificador, enquanto a outra diagonal, os erros ou confusão entre as classes. Esta matriz fornece as bases para o estabelecimento de várias métricas (Figura 24).

Figura 24 – Matriz de Confusão e métricas



Fonte: Fawcett, 2006.

As Curvas ROC são gráficos bidimensionais, onde os verdadeiro-positivos (VP) são traçados no eixo Y, enquanto os falso-positivo (FP) são traçados no eixo X e representam o custo benefício entre ambos.

3 TRABALHOS RELACIONADOS

Os trabalhos encontrados na literatura estão basicamente relacionados com a contagem de pessoas e abordam a detecção de pessoas das mais variadas formas. Este capítulo apresenta os trabalhos que mais se assemelham com os desafios de se detectar a intrusão de uma pessoa em região delimitada por cerca virtual e serão apresentados primeiramente aqueles trabalhos voltados para a localização de pessoas em imagens, seja utilizando apenas um tipo de informação, ou combinando mais de uma informação, como por exemplo, contorno, histograma de gradientes orientados, localização de partes do corpo ou ainda, que utilizem qualquer outro de tipo de informação adicional como movimento, textura ou cor para detectar e rastrear pessoas. Na sequência, são apresentados trabalhos que realizam a contagem de pessoas presentes na imagem inteira ou apenas de pessoas que passam por regiões específicas, como por catracas, por exemplo. Finalmente, são apresentados alguns trabalhos que identificam comportamentos das pessoas executando algum tipo de atividade, que tem relação com a detecção de eventos. Terminadas as apresentações dos trabalhos, é feita uma breve discussão relacionando alguns destes com o presente trabalho.

3.1 APRESENTAÇÃO DOS TRABALHOS

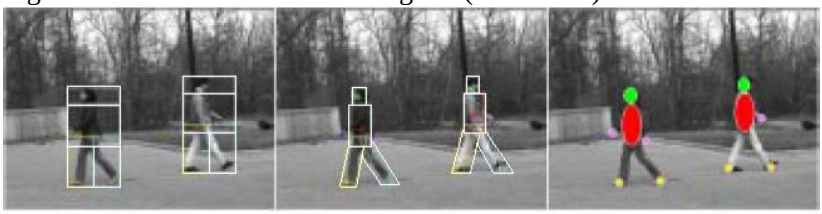
Ainda que o invasor esteja totalmente visível, sem oclusões parciais ou totais, dependendo do posicionamento da câmera, do carro e do invasor, o carro pode fazer parte do fundo móvel da imagem, tornando a segmentação e a classificação do pedestre desafiadora. Fora isto, há outras questões como a possibilidade de um número elevado de pessoas tentarem entrar ao mesmo tempo.

Algumas abordagens para identificação de pessoas se baseiam no princípio de janela deslizante. Viola-Jones (2001) usam o conceito de padrões de imagens integrais para acelerar a extração de características e uma arquitetura de detecção em cascata (classificadores fortes e fracos), além de processar de forma automática as características mais significativas.

Haritaoglu et al. (1998) modelaram um cartão retangular (ver Figura 25), onde dividiram o retângulo envolvente em torno de uma pessoa em sub-regiões de tamanho adequado e com a localização mais provável de se encontrar partes do corpo humano, como cabeça, tórax,

braços e pernas, bem como o provável tamanho destas partes. A Figura 25 (a) traz as caixas envolventes iniciais nas imagens em primeiro plano (*foreground*), e após análise do cartão retangular (*cardboard*) a cabeça, tronco e membros são destacados em (b). Em (c) é ilustrado a localização das partes do corpo por meio de elipses.

Figura 25 – Modelo de cartão retangular (*cardboard*)



Fonte: Haritaoglu et al., 1998.

Dalat et al. (2005) criaram o método de análise de Histogramas de Gradientes Orientados (HOG) onde as características são computadas de forma muito similar ao SIFT e igualmente utiliza o conceito de janela deslizante. Do ponto de vista de ferramentas isoladas, é um dos mais eficazes métodos na detecção de humanos, e há inúmeros trabalhos associando HOG com outras características, tais como textura, cor, movimento, padrões locais como SIFT, forma ou contorno e padrões binários locais (LBP).

Schwartz (2011) obteve uma representação melhor para detecção de humanos combinando HOG com informações de cor e textura, o que resultou em um espaço de características com muitas dimensões e, para processá-las, usou computação paralela envolvendo CPUs com mais de um núcleo e múltiplas unidades de processamento gráfico (GPU).

Por outro lado, Benenson et al. (2012) perceberam que se, ao invés de se selecionar as características predominantes em várias escalas, fizessem um treinamento isolado para cada escala individualmente, então a detecção seria cerca de 20 vezes mais rápida. Enquanto nos métodos tradicionais envolvendo HOG sucessivos redimensionamentos da imagem original são necessários, esta nova abordagem transfere parte do esforço computacional para a fase de treinamento, que é realizado uma única vez. Depois de treinadas as SVM para as diversas escalas, janelas de deslizamento percorrem a imagem original sem necessidade de redução de escala, garantindo taxas de processamento de 100 quadros por segundo.

Outras técnicas, envolvem o treinamento de um classificador SVM para detectar contornos da região da cabeça [SUBBURAMAN et al. 2012]. Para acelerar o processo, regiões de interesse são estimadas através da orientação de gradientes que sejam similares ao topo da cabeça nas imagens em tons de cinza. A seleção e integração das melhores características são feitas automaticamente.

Wei et al. (2013) fazem uso de dois classificadores em cascata, sendo o primeiro um classificador fraco baseado nas características do tipo Haar [VIOLA et al. 2005], seguido de um classificador mais refinado baseado em HOG e no algoritmo *AdaBoost*.

O movimento é uma informação importante para percepção de pessoas. Novamente Dalal et al. modelaram um sistema de estatísticas de movimento de diferenças internas em um campo de fluxo óptico [DALAL et al. 2006].

Liu et al. (2009) ampliaram a representação unidimensional do contorno das pessoas em uma abordagem tridimensional, através da transformada de distância entre quadros. Assim, trabalhando nas dimensões espaço-temporal, incluíram a variável movimento fundida às características de contorno. Quadros sucessivos de uma mesma pessoa formam uma imagem volumétrica, das quais são extraídas características Haar 3D. Técnicas de *boosting* são empregadas para selecionar as características mais marcantes e para construção do classificador em cascata. Eles batizaram este conjunto que incorpora em uma mesma variável as informações de aparência e movimento com o nome de Características de Contorno-Movimento (CMF, do inglês *Contour-Motion Feature*).

Já Conte et al. (2010) fazem uso da extração de características SURF e as analisam em movimento. Através de um processo de agrupamento de tais vetores de características, tentam associar quais características estão se movendo juntas, ou seja, quais características pertencem a cada grupo de pessoas sendo rastreado.

Para Martin et al. (2012), o movimento é também um fator importante na detecção de humanos e um sistema completo deve utilizar as informações providas de forma integrada pelos sistemas de localização de pessoas (aparência) e de movimentação (rastreamento). Os resultados de cada sistema são realimentados para o outro sistema com o objetivo de aumentar o desempenho global.

As pessoas apresentam padrões de movimentação distintos. Considerando a pessoa como um objeto envolto por retângulo imaginário é possível detectar-se uma pessoa analisando dois comportamentos deste retângulo envolvente: (1) o comportamento

cíclico em sua trajetória e (2) uma correlação entre a alternância de posicionamento e tamanho das dimensões do retângulo envolvente [BORGES 2013]. Ainda segundo Borges (2013), o procedimento padrão na extração de características HOG e treinamento da SVM para posterior utilização na classificação envolve a análise das melhores características que se repetem em diferentes escalas, seja ampliando ou reduzindo a imagem original. Concluído o treinamento, a janela padrão deslizante faz a varredura de toda a imagem procurando por pessoas. A imagem original é sucessivamente reduzida e o processo de varredura é reiniciado com a finalidade de se localizar pessoas nas mais variadas escalas de tamanho.

Existem trabalhos que visam a contagem de pessoas que serão aqui discutidos e apresentam técnicas de detecção e segmentação de pessoas para enfrentar ambientes complexos e desafiadores.

A proposta de Sim et al. (2012) usa apenas informações espaciais, portanto, sem se preocupar com a questão temporal e o rastreamento dos objetos. Primeiramente um detector local treinado do tipo Viola-Jones é usado para localizar indivíduos em imagens altamente povoadas (multidões). Esta abordagem apresenta altos índices de falso-positivos. Os falso-positivos são minimizados em passos posteriores, seja usando segmentação por cor ou através de um modelo fraco da perspectiva de uma câmera não calibrada, auxiliando em determinar se o tamanho e posição da região candidata podem confirmá-la como sendo uma cabeça.

Garcia et al. (2013) criaram um método para contar pessoas usando uma única câmera fixa. A contagem requer duas fases: a detecção propriamente dita e o rastreamento das pessoas, onde a detecção é baseada na localização das cabeças dos indivíduos através da correlação das imagens com alguns padrões circulares, enquanto o rastreamento é feito pela aplicação de um filtro de Kalman [KALMAN 1960] para determinar a trajetória dos candidatos. A subtração de fundo desempenha um papel importante na segmentação.

A simples segmentação de indivíduos pertencentes a pequenos grupos ou multidões, não resolve o problema da contagem [LIU et al. 2005]. Entretanto, para o escopo do presente trabalho, ela é suficiente, pois o objetivo é alarmar intrusões, independentemente de quantas pessoas estão participando dela, ou seja, se dentro de um grupo de pessoas pelo menos uma for detectada, já é o suficiente para configurar o evento de intrusão. Uma abordagem possível é a definição de um portão virtual. O conceito de portão ou cerca virtual é útil para realizar a

contagem de pessoas, bem como determinar o sentido de deslocamento, se a pessoas está entrando ou saindo de uma região de interesse.

Vishwakarma et al. (2013) fizeram uma pesquisa sobre reconhecimento de atividades e comportamentos de humanos em vídeos de vigilância. As pessoas presentes nas imagens passam a ter seus gestos, atividades, interações e atividades em grupo estudadas.

3.2 DISCUSSÃO DOS TRABALHOS

Os trabalhos que fazem uso da técnica Viola-Jones para localizar cabeças ou outras partes do corpo humano, uma vez treinados, não lidam com transformações afins como rotação e variação de escala tão bem quanto HOG é capaz de lidar. Além disto, HOG também supera Viola-Jones no quesito variação de luminosidade de imagem.

O conceito de *cardboard* utilizado por Haritaoglu et al. (1998) funciona quando há apenas pessoas presentes na cena. Com a introdução de outros objetos, como por exemplo carros, em primeiro plano (*foreground*) juntamente com as pessoas, então tentar ajustar os cartões às pessoas fica impraticável. Mesmo que não haja carro presente na cena, grupos nos quais pessoas estejam caminhando próximas umas às outras ou nos quais pessoas estejam sobrepondo partes de outras pessoas também impossibilitam o uso desta técnica como foi concebida. Por outro lado, se for possível primeiramente localizar a cabeça e partir de sua posição ajustar o cartão para encontrar as demais partes do corpo, então a abordagem do cartão pode ser uma boa estratégia. No presente trabalho pretende-se localizar as cabeças utilizando-se a transformada de Hough.

Pesquisadores como Dalal et al. (2006) e Liu et al. (2009) entenderam que o movimento é uma informação importante na localização de pessoas. Os primeiros usaram informações extraídas do fluxo óptico enquanto Liu e outros criaram um vetor de características que aglutina o contorno e movimento das pessoas. No presente trabalho, o fluxo óptico tem relevante importância no processo de detecção de pessoas, pois ele não só é capaz de identificar as regiões de movimentação, como também pode servir de estimativa para se avaliar com qual intensidade e para onde o deslocamento dos objetos ocorre (rastreamento de curto alcance). A diferença da presente abordagem é que as características extraídas do campo fluxo óptico são grandezas estatísticas, sofrendo, portanto, pouca interferência do posicionamento relativo das câmeras. O objetivo é que uma vez treinado para um

conjunto de imagens obtidas de cenários diferentes, com ajustes de câmeras diferentes e calibradas, o sistema possa ser usado em novos ambientes sem necessidade de treinamento, dependendo apenas do ajuste de alguns parâmetros.

Os trabalhos que se baseiam em técnicas de rastreamento, como os de Conte et al. (2010) e Borges (2013) geralmente precisam de uma sequência longa de quadros do vídeo e demandam processamento computacional. No presente trabalho, o foco é na rapidez do sistema em lidar com os vídeos em tempo real (16 fps) e gerar alarmes assim que caracterizada a intrusão. Para isto, o rastreamento das pessoas será descartado.

Martin et al. (2012) se baseia no uso conjunto de dois classificadores trabalhando de forma paralela, um focado na aparência (silhueta) e outro focado no rastreamento (movimento) das pessoas. Além disto, os classificadores trocam informações entre eles (realimentação) para aprimorar o resultado da classificação final. Por outro lado, Sim et al. (2012) não lidam com o rastreamento e usam classificadores em série (ou cascata), sendo o primeiro uma seleção mais grosseira, com níveis de falso-positivos elevados e que são passados para um próximo filtro mais seletivo similar ao mapa de projeções proposto por Jo et al. (2009). Seguindo esta linha de pesquisa, no presente trabalho o comportamento do campo de vetores de fluxo óptico é estudado para se extrair informação que possa melhorar o resultado da classificação entre pessoas e não-pessoas. Como o rastreamento não é prioridade, será projetado algo similar ao proposto por Sim et al. (2009) e Wei et al. (2013). Ou seja, o presente trabalho aborda o uso de dois classificadores em série que não trocam informações entre si e são treinados separadamente.

Trabalhos como o de Vishwakarma et al. (2013) ajudam a entender como detectar eventos, ao invés de apenas se detectar pessoas.

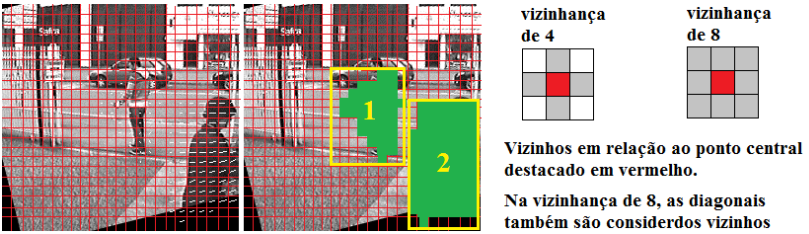
Resumindo, o presente trabalho utiliza técnicas diferentes de detecção de humanos, buscando obter o mais próximo possível de 100% de identificação de pessoas, mas mantendo baixos os níveis de falso-positivos. Independentemente do método a ser utilizado, partindo-se do pressuposto de que os vetores de deslocamentos dos automóveis e das pessoas serão diferentes, é provável que a partir desta premissa seja possível identificar pessoas tentando invadir o estabelecimento pelo portão aberto. Lembrando ainda que o processamento tem que ser em tempo real, na casa dos milissegundos. A invasão será configurada pela ultrapassagem de uma cerca virtual. A maioria dos trabalhos relacionados concentra-se na contagem de pessoas, em locais onde se

espera a presença exclusiva de pessoas. O presente trabalho tem por finalidade gerar um alarme para o operador de CFTV toda vez que uma pessoa entrar pelo portão exclusivo para carros, mas ignorar e tratar como normal a situação onde somente carros ultrapassem o portão virtual. As situações críticas ocorrem justamente quando: (1) o invasor tenta ocultar-se atrás do carro e (2) o algoritmo não consegue segmentar as pessoas quando atrás delas aparecem os carros em movimento. Os detalhes da arquitetura proposta são vistos no próximo capítulo.

4 MÉTODO FOGI PARA DETECÇÃO DE INTRUSÃO

Para este trabalho os vídeos são capturados por câmeras de resolução 320x240 pixels. A seguir, é calculado o campo vetorial de fluxo óptico. Um limiar ou gatilho de movimentação é estabelecido para se determinar as regiões de interesse. A granulosidade, ou melhor, o tamanho da matriz de movimentação é definido em função de tamanho das subdivisões N x N medidas em pixels. Quanto maior a granulosidade, menor a sensibilidade de detecção de movimento e também será menor a matriz de movimentação. Na Figura 26 está representada uma matriz de movimentação obtida a partir de uma célula de tamanho 10x10 pixels. A referida célula é usada no cálculo do fluxo óptico denso pelo método proposto por Farneback (2000).

Figura 26 – Segmentação por limiar de movimentação do fluxo óptico

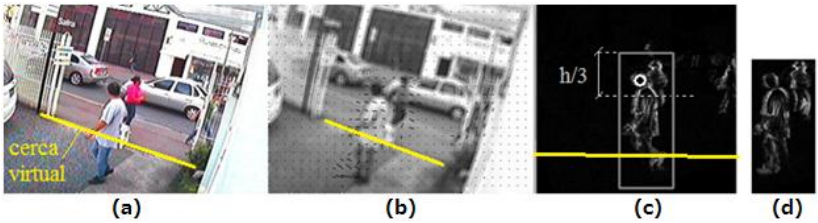


Fonte: produção do próprio autor

O cálculo do fluxo óptico serve a dois propósitos. Primeiramente ele será usado para segmentação. Após a aplicação de classificador HOG nas imagens de interesse, o fluxo óptico destas regiões será analisado para auxiliar na decisão final de se classificar o objeto como pessoa ou não-pessoa.

Na Figura 27, por exemplo, tem-se em (a) a imagem original, em (b) a conversão em tons de cinza, onde aplicou-se a média gaussiana [PAULUS et al. 1995], seguida do cálculo do fluxo óptico, em (c) pessoas com a cabeça localizada por transformada de Hough e em (d) a figura candidata selecionada pela transformada de HOG.

Figura 27 – Exemplo de detecção bem-sucedida de pessoa

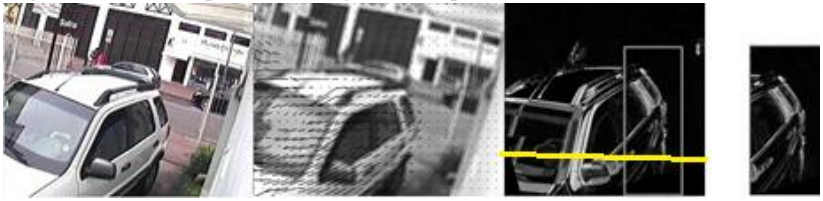


Fonte: produção do próprio autor

Assim, para minimizar o tempo de processamento, a janela deslizante da transformada de HOG é aplicada apenas nas regiões onde há movimentação medida pela intensidade do vetor de fluxo óptico. Uma vez delimitada uma região candidata, a confirmação de ser ou não uma pessoa poderia ser feita pela localização da cabeça em regiões de interesse específicas da região candidata, como na terça parte superior do retângulo na Figura 27 (c). A linha imaginária que representa a cerca virtual pode ser observada nas Figuras 27 (a), (b) e (c), sendo que em (c) a imagem teve sua inclinação corrigida (processo de calibração da câmera).

Já a Figura 28 traz um exemplo de falso-positivo, que poderia ser descartado pela ausência da detecção da cabeça na região candidata.

Figura 28 – Exemplo de detecção falso-positiva



Fonte: produção do próprio autor

A quantidade de objetos segmentados para análise é determinada por um processo de definição de sementes e crescimento de região em todas as direções. À medida que objetos diferentes tenham contato entre si considerando-se uma vizinhança de 8 elementos, eles são agrupados como um objeto único. A Figura 26 ilustra também o processo de segmentação a partir do campo fluxo óptico, com duas regiões (em verde) que cresceram separadamente e não se conectaram.

Na Figura 29 (b) os pontos classificados como pertencentes a região de borda, levando-se em consideração vizinhança de 8 elementos,

estão destacados por círculos pretos, enquanto os pontos classificados pertencentes ao interior do objeto, foram destacados por círculos brancos.

Figura 29 – Imagem destacando pontos de borda x interior

Fonte: produção do próprio autor

A subtração de quadros consecutivos foi usada com o objetivo de se destacar as bordas dos objetos em movimento, descartando os gradientes de imagens estáticas. Este método apesar de simples, responde instantaneamente às variações de luminosidade o que é desejável. Por outro lado, esta alta sensibilidade pode trazer ruídos caso a câmera esteja sujeita a oscilações, como às causadas por ação do vento.

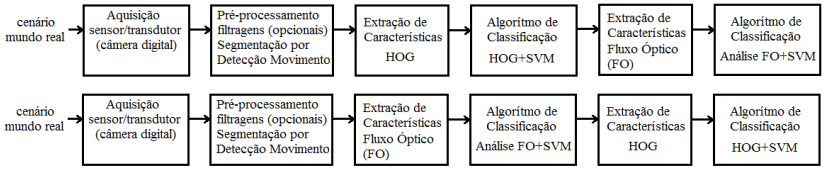
A binarização é o processo de se transformar uma imagem composta de dois níveis distintos, verdadeiro ou falso, ou em termos de intensidade luminosa, branco ou preto. Uma imagem composta de 256 tons de cinza, por exemplo, pode ser binarizada a partir de um valor limiar, também chamado de gatilho. Os tons de cinza da imagem variam de 0 (preto) ao 255 (branco). No decorrer da binarização, todo pixel abaixo do valor de gatilho será considerado 0, enquanto os níveis acima dele serão considerados 255.

Uma vez selecionadas as regiões de interesse em que houve movimentação, HOG é aplicado sobre elas, bem como a análise do fluxo óptico para se determinar com maior nível de segurança se a região candidata é ou não uma pessoa através de técnicas de reconhecimento de padrões.

O objetivo é garantir um maior nível de confiança na detecção de intrusos ao se associar técnicas como HOG e a análise de fluxo óptico para reconhecer padrões, sendo que os classificadores serão independentes e dispostos em cascata. A partir desta premissa, após a

segmentação por movimento há a possibilidade de se classificar usando HOG seguido da análise do comportamento do fluxo óptico, ou inversamente, primeiramente classificar usando a análise do comportamento do fluxo óptico e depois, HOG (esquema da Figura 30).

Figura 30 – Classificadores independentes em cascata



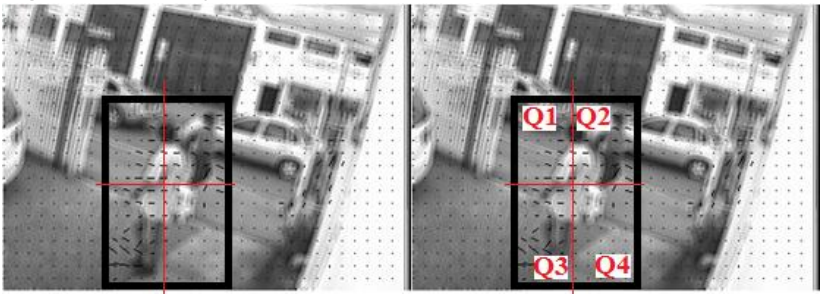
Fonte: produção do próprio autor

Na Figura 31, o retângulo maior é o resultado da detecção de pessoas utilizando-se HOG. Para se analisar o comportamento dos vetores do campo FO, a região é subdividida em outras quatro regiões denominadas quadrantes Q1, Q2, Q3 e Q4. As seguintes características são selecionadas:

a) Histograma das direções do fluxo óptico. Contagem das ocorrências das angulações do fluxo óptico, levando-se em conta intervalos fixos (partições). A região de 360º pode ser particionada, por exemplo, em 24 fatias de 15º em 15º e assim o histograma das direções é formado. Existe apenas um histograma global para o objeto em análise, não sendo feito o cálculo por quadrantes. Para que estes dados estejam normalizados, será escolhida a frequência de maior ocorrência e suas 4 vizinhas imediatas a esquerda e direita, denominadas de freq1, freq2, freq3, freq4, freq5 (a de maior ocorrência ou modal), freq6, freq7, freq8 e freq9;
b) Densidade de pontos que sofreram movimentação dentro do retângulo maior. Uma vez selecionada a região candidata, é feita a contagem de sub-regiões de tamanho N x N que sofreram movimentação e dividida pela quantidade total de sub-regiões que compõem o objeto. Também se trata de uma característica global, e não é avaliada por quadrantes;
c) Utilizando o conceito de janelas de Parzen, medição da aderência das frequências em cada um dos quadrantes Q1 a Q4 em relação à frequência dominante da região global;
d) Desvio padrão das direções do fluxo óptico em relação à média em diversas regiões: global (Q1 a Q4), quadrantes

superiores (Q1 e Q2), quadrantes inferiores (Q3 e Q4) e em cada um dos quadrantes isoladamente (Q1, Q2, Q3 e Q4).

Figura 31 – Extração de características



Fonte: produção do próprio autor

Ao se comparar as partes inferiores (Q3 e Q4) da Figura 31 (pessoa) com a região correspondente à detecção falso-positiva na Figura 28 (carro), nota-se que nos carros os vetores do campo de fluxo óptico são mais alinhados como um todo, inclusive nas regiões dos quadrantes Q1 e Q2.

Deve-se levar em conta, que provavelmente a utilização de qualquer uma destas técnicas isoladas para detecção de pessoas, sem usar outras informações adicionais, dificilmente atingirá o objetivo de detectar 100% das pessoas presentes em diversos vídeos diferentes com o mínimo de falso-positivos, sendo que as câmeras terão ajustes diferentes, principalmente em relação ao posicionamento. Inclusive há o trabalho de Cheng et al. (2013) que corrobora esta teoria, ou seja, o uso de características adicionais combinadas traz melhores resultados do que a utilização de HOG isoladamente, aumentando o nível de confiança da detecção.

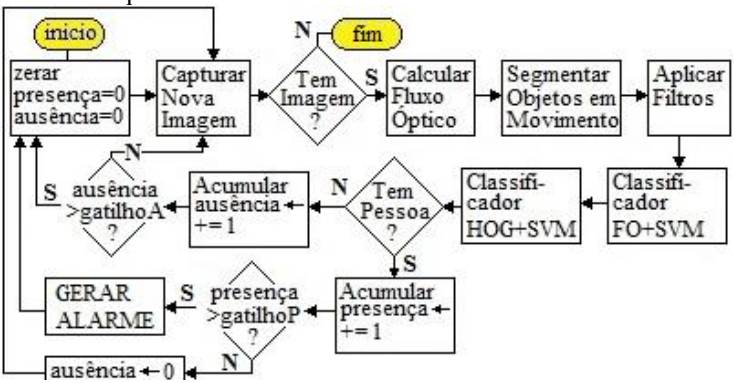
Uma vez feita a detecção das pessoas, é necessário estabelecer uma estratégia para que o evento intrusão seja determinado de forma estável. Classificar e contar pessoas em um quadro (imagem estática) é uma tarefa fácil, ainda que possam haver casos de falso-positivos. Por outro lado, medir eventos já é uma tarefa um pouco mais complexa, pois envolve a análise de uma certa quantidade de quadros ao longo de um intervalo de tempo. Para se medir eventos, o principal critério é a contagem de objetos corretamente classificados restritos a um intervalo temporal. Além disto, a definição do evento pode tornar a tarefa mais fácil ou difícil. Por exemplo, é mais fácil detectar o evento de intrusão de um objeto qualquer em uma zona de interesse, do que a detecção de

uma pessoa invadindo a mesma zona de interesse. Por sua vez, determinar o evento “pessoas correndo”, ou “pessoas lutando” está em um nível de complexidade maior do que os dois exemplos anteriores [NGHIEM et al. 2007].

O sistema não pode basear-se na detecção de apenas uma pessoa em um único quadro para gerar o alarme. É necessário considerar o conceito de persistência do evento. Assim, o primeiro passo é limitar a região de interesse, a partir de uma linha imaginária traçada ao longo do portão de entrada, que será chamada de cerca virtual. Toda vez que a base do retângulo envolvente (RE) em torno das pessoas ultrapassar esta cerca virtual, então se está diante de um potencial evento de intrusão. Diz-se potencial evento, porque a persistência do evento precisa ser confirmada em uma sequência de quadros posteriores, para eliminar oscilações e falsos alarmes desnecessários. Em outras palavras, o evento precisa ser consistente por um intervalo de tempo mínimo antes de gerar um alarme para o operador de CFTV.

Como o sistema é baseado em um classificador treinado com informações extraídas do campo de FO no qual os vetores lembram uma neblina ou névoa, cascateado com outro classificador baseado em HOG, resolveu-se chamar o sistema completo de FOGI, aglutinando os dois nomes e acrescentando o termo Intrusão. O sistema FOGI é maior do que apenas os dois classificadores em cascata, abrangendo toda estratégia de parametrização de contadores para configurar o evento intrusão e geração de alarme. O esquema completo da arquitetura pode ser visto na Figura 32.

Figura 32 – Arquitetura FOGI



Fonte: produção do próprio autor

A Figura 33 apresenta alguns exemplos de filtragem realizados exclusivamente pelo classificador FO. As imagens resultantes são direcionadas para o classificador HOG.

Figura 33 – Resultado do classificador FO

Fonte: produção do próprio autor

O capítulo 5 apresenta alguns testes utilizados para se determinar qual das opções discutidas ao longo deste capítulo 4 apresenta o melhor desempenho, além dos testes necessários para responder as perguntas levantadas nos objetivos gerais.

5 EXPERIMENTOS E RESULTADOS

Na literatura, os trabalhos relacionados eram muito vagos em relação a métricas para medição de eventos, tratando muito rapidamente sobre o tema, citando apenas que métricas de eventos são estabelecidas a partir de ocorrências de algo durante um certo intervalo de tempo. Na ausência de uma base de dados específica sobre intrusão, foi criado um banco de dados com imagens reais e específicas de sistemas de vigilância em ambientes externos com fluxo concomitante de pessoas e carros. Este capítulo concentrou-se em experimentos para localização de pessoas pela escolha dos melhores parâmetros para cálculo do fluxo óptico e ajuste fino dos classificadores HOG e FO, culminando em testes que auxiliassem na definição de uma estratégia prática para medição de eventos de intrusão. Para implementar os algoritmos e avaliar as métricas, foram usadas as ferramentas Python v2.72[2], OpenCV v2.43, Weka v3.6.83[3] instaladas em um computador PC com processador AMD Turion II P520 Dual Core 2.3 GHz, 64 bits, 4 GB de memória e com sistema operacional Windows 7. Alguns experimentos foram projetados de forma a responder as questões de pesquisa levantadas na seção objetivos gerais do capítulo 1.

Os testes são divididos em 9 experimentos. Assim, o experimento 1 concentra-se em avaliar por que o classificador HOG da biblioteca OpenCV treinado para detectar corpos inteiros das pessoas não tem o mesmo desempenho relatado na literatura. O experimento 2 aborda o uso de variações de configurações dos classificadores FO e HOG em cascata e a variação de parâmetros do classificador FO, além de responder se o FO pode ser usado como detector de regiões de movimento. O experimento 3 estuda a possibilidade do uso de filtros nas imagens antes da etapa de classificação, enquanto o experimento 4 estuda a viabilidade de se usar a Transformada de Hough para localizar a cabeça das pessoas. O experimento 5 tem o propósito de medir o tamanho de grade necessário para se obter um campo de fluxo óptico, a partir do qual se consiga extrair características estatísticas que possam diferenciar pessoas de carros. O experimento 6 verifica se a adição do classificador FO (uso da informação de movimento presente na sequencia de imagens) apresenta melhor desempenho do que o uso isolado do classificador HOG, além de testar se a inversão de ordem de

[2] Python, disponível em https://www.python.org (em 10/Junho/2015)
[3] Weka, disponível em http://www.cs.waikato.ac.nz/ml/weka (em 10/Junho/2015)

aplicação dos classificadores influi na resposta final. O experimento 7 mede o efeito de se utilizar campos de FO acumulados por mais do que apenas dois quadros consecutivos no treinamento da SVM. O experimento 8 testa o uso das redes neurais MLPs como alternativa para se reconhecer o padrão de movimentação das pessoas através da análise do FO. Finalmente, a estratégia utilizada para se detectar o evento intrusão e gerar alarme para operador de CFTV também é descrita de forma mais detalhada no experimento 9.

Ao final do capítulo, simulações demonstram os resultados superiores obtidos pela classificação FO→HOG para geração de alarmes de intrusão.

5.1 BANCO DE DADOS CONTENDO PESSOAS E CARROS

As imagens foram obtidas através de uma placa de captura Geovision modelo GV-800, instalada em uma plataforma Windows XP. A microcâmera é colorida e analógica, com resolução de 320x240 pixels. O formato do arquivo gravado é o padrão AVI. O sistema está instalado em um ambiente real, de um estacionamento de veículos comercial e, portanto, com elevado fluxo de automóveis e pessoas. Outras imagens foram obtidas a partir da Internet para compor o grupo de testes.

Os vídeos obtidos pela câmera deste estacionamento foram separados em dois grupos: treinamento e testes. No grupo de treinamento, há 42 pessoas marcadas através de 791 retângulos e 22 carros marcados em 277 retângulos, totalizando 780 quadros. Já o grupo de teste foi separado em dois subgrupos. O primeiro subgrupo de teste é composto de 27 pessoas, marcadas por 597 retângulos e três carros não marcados, cujas imagens foram obtidas a partir da mesma câmera usada para treinamento, tratando-se, portanto, de um grupo de teste para controle. Já o segundo subgrupo de teste é composto de 1478 marcações de pessoas e 1322 marcações de carros, provenientes de diversas câmeras em ambientes e regulagens diferentes. Considerando as imagens em que há movimentação, e consequentemente com potencial risco de invasão, 52,90% delas contêm pelo menos uma pessoa (podendo ter carros também ao fundo ou em primeiro plano), enquanto 47,10% possuem apenas carros em movimento.

O objetivo principal deste trabalho é detectar o evento intrusão de pessoas, por um classificador que supere o uso da classificação por apenas HOG. Portanto, as regiões de interesse foram marcadas

exatamente correspondendo às regiões classificadas pelo HOG como pessoas. Para tanto, o deslocamento do hiperplano da SVM do classificador HOG foi variado da posição original zero para posição -1,0 que é a região com índice de acerto na faixa de 90%, mas com elevados números de falso-positivos. Deslocar o hiperplano da SVM no sentido negativo aumenta a possibilidade de se detectar uma pessoa (verdadeiro-positivos), enquanto o deslocamento no sentido positivo causa o efeito contrário, qual seja, o aumento dos falso-negativos e também redução dos falso-positivos.

Apesar do objetivo primário ser a detecção de pessoas com o mínimo de ocorrências falso-negativas, caso contrário intrusos poderiam invadir o local sem serem notados, deve-se minimizar também as ocorrências falso-positivas, senão muitos alarmes seriam gerados sem que de fato ocorressem intrusões. Para minimização de falso-positivos, o foco do estudo é justamente nas áreas marcadas como falso-positivas pelo classificador HOG e avaliar como o fluxo óptico se comporta nestas regiões comparativamente com o fluxo óptico das regiões classificadas corretamente como verdadeiro-positivas.

Assim, todas as regiões obtidas acima foram marcadas apropriadamente, como pessoas e não pessoas através de um aplicativo desenvolvido pelo autor. O escopo destas marcações abrangeu apenas as imagens em movimento. Em última análise, pessoas estáticas não são detectadas, e na eventualidade de serem marcadas por detecção HOG, serão descartadas pela análise das características do Fluxo Óptico.

O padrão referencial ou ideal (do inglês *ground truth*) das marcações foi definido manualmente por um operador humano e todos os retângulos tem o tamanho variável na faixa de 64x128 pixels até 96x192 pixels. No total são 1680 quadros, envolvendo 25 carros (277 marcações), 69 pessoas (1388 marcações).

5.2 EXPERIMENTO 1

Nas fases iniciais dos experimentos, foi avaliada a eficácia da implementação da detecção de humanos utilizando apenas HOG. Uma das premissas mais importantes deste trabalho é o reconhecimento de humanos independentemente de treinamento do processo classificatório com a base de imagens de cada usuário final e, portanto, foram utilizadas as bibliotecas do OpenCV com a máquina SVM já treinada para detectar humanos a partir das características HOG.

O treinamento leva em conta que as pessoas estejam na posição ereta. A oclusão é outro fator que causa a degradação no sistema de reconhecimento.

Um dos primeiros problemas percebidos foi que a classificação por HOG para a detecção de humanos utilizando o treinamento padrão do OpenCV é relativamente boa para avaliar pessoas eretas (máquina SVM treinada para detectar corpos inteiros, da cabeça aos pés), mas nos casos em que as câmeras estejam inclinadas e sem calibração, as pessoas podem não ser detectadas. A Tabela 2 e a Figura 35 demonstram esta situação. Para testar o quanto a inclinação da câmera degrada os resultados de detecção das pessoas, foram utilizados os quadros de 3 a 10, sequência parcial esta que pode ser vista na Figura 34 e que contém apenas uma pessoa. Até 10 graus de inclinação não afetam os resultados, mas a partir de 15 graus o desempenho cai drasticamente e em 20 graus ou mais já não é possível detectar a pessoa. Um processo de rotação do vídeo originalmente capturado resolveu a questão abordada neste primeiro teste. Pessoas entrando curvadas também podem não ser detectadas.

Tabela 2 – Influência da inclinação das câmeras na detecção HOG

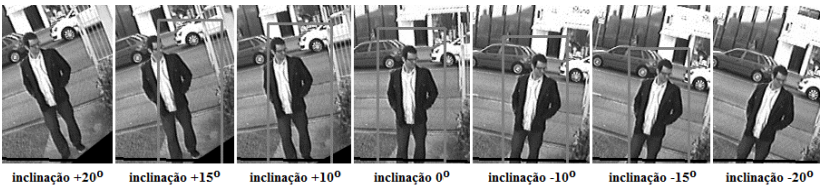
Inclinação do vídeo	Quantidade de Verdadeiro-Positivos	Quantidade de Falso-Positivos
-20^{o}	0	0
-15^{o}	3	0
-10^{o}	7	0
0^{o}	7	0
$+10^{o}$	7	2
$+15^{o}$	3	0
$+20^{o}$	0	0

Fonte: produção do próprio autor

Figura 34 – Sequência parcial de aquisição original sem tratamentos

Fonte: produção do próprio autor

Figura 35 – Influência da inclinação da câmera na detecção HOG



Fonte: produção do próprio autor

5.3 EXPERIMENTO 2

Foi constatado que a utilização da segmentação por fluxo óptico reduz a região a ser testada. Nas imagens de testes, quando haviam apenas pessoas, 42,77% da área total foi segmentada e repassada para classificação por HOG. As regiões em que não ocorreram movimentações foram descartadas (57,23% restantes da área total do quadro). Nas imagens usadas para treinamento, com a presença de pessoas e carros, a área processada representava 53,16% do total.

As máquinas SVM podem ser treinadas com pesos diferentes para as classes. Assim, como o foco é na detecção das pessoas, foram feitos testes considerando-se relações diferentes de pesos para as classes pessoas e carros. Também deseja-se verificar se a aplicação do classificador FO em série com classificador HOG traz resultados diferentes de se usar apenas HOG.

A Tabela 3 apresenta uma compilação dos resultados obtidos com diferentes abordagens de análise e configurações de parâmetros,

aplicado nos vídeos contendo 601 pessoas marcadas. Quando se utiliza o classificador FO, a quantidade de falso-positivos diminui e isto é bom, pois evita a geração de alarmes incorretos.

Tabela 3 – Influência dos Parâmetros no Desempenho de Classificação

Processo	Parâmetros	Verdadeiro-Positivos	Falso-Negativos	Falso-Positivos
(1) HOG sob imagem original	Não se aplica	175	426	564
(1) HOG sob imagem diferencial	Gatilho=40 binarização	314	287	228
(1) Segmentação FO → (2) HOG	Gatilho=40	**356**	**245**	107
(1) Segmentação FO → (2) Análise FO SMV linear → (3) HOG	passoLin=1 passoCol=1	343	258	89
(1) Segmentação FO → (2) Análise FO SMV linear → (3) HOG	passoLin=3 passoCol=3	290	311	69
(1) Segmentação FO → (2) Análise FO SMV grau 3 → (3) HOG	passoLin=1 passoCol=1	345	256	91
(1) Segmentação FO → (2) Análise FO SMV grau 3 → (3) HOG	passoLin=3 passoCol=3	295	306	**65**

Fonte: produção do próprio autor

5.4 EXPERIMENTO 3

Também foram feitas algumas experiências para tentar melhorar o desempenho das detecções através da aplicação de filtros. A primeira ideia testada foi suavizar as imagens com um filtro gaussiano [PAULUS et al. 1995], porque este filtro suaviza a imagem, simplificando-a e alterando os gradientes. Outra questão que mereceu atenção foi o fato de HOG dar maior peso aos contornos. Então, além da filtragem das imagens pela média (filtro gaussiano), foram testadas as transformadas de Sobel [PAULUS et al. 1995] e a subtração de quadros consecutivos para destacar as bordas, sendo que a subtração de quadros teve um desempenho superior comparativamente ao uso direto das imagens originais (conforme pode ser visto na Tabela 4), considerando-se que em nenhum momento a máquina SVM foi treinada com as imagens adquiridas para o propósito deste trabalho.

Muitas vezes, detecções de falso-positivos ocorriam em regiões estáticas (ausência de objetos em movimento). Uma das maiores recorrências pode ser vista na Figura 36. O destaque em vermelho na imagem foi inserido manualmente para destacar a região que o método HOG acabou classificando como uma pessoa.

Figura 36 – Exemplo de detecção falso-positiva de objeto estático

Fonte: produção do próprio autor

Uma vez que o foco do trabalho é em pessoas se movimentando, uma forma simples de eliminar estes falso-positivos foi associar a procura de pessoas apenas nas áreas onde houve variação do fluxo óptico, ou seja, onde ocorreu alguma movimentação. Isto, não só diminui a incidência de falso-positivos detectados por HOG, como também torna a localização de pessoas mais rápida, descartando a análise nas regiões estáticas.

A Tabela 4 apresenta alguns resultados obtidos nesta etapa inicial, comparando o desempenho com e sem a utilização de pré-processamento das imagens. Nota-se pelos dados da Tabela 4, que a utilização de HOG sob a imagem filtrada (média gaussiana) para localizar pessoas melhorou o desempenho comparativamente ao uso de HOG sob a imagem original, tornando a detecção mais sensível a ponto de aumentar a quantidade de indivíduos reconhecidos (exceção arquivo 12), mas ao mesmo tempo ocasionou o aumento dos falso-positivos. A aplicação de filtros para simplificar a imagem, suavizando-a, causa a redução dos gradientes, que é justamente a informação utilizada pelo classificador HOG na detecção de pessoas. Por fim, a coluna final mostra que os falso-positivos podem ser consideravelmente reduzidos, utilizando-se as operações apropriadas. Em termos de desempenho, o critério adotado foi considerar melhor o pré-processamento ou filtro que levou o classificador a encontrar o maior número de pessoas. Estes verdadeiro-positivos foram destacados em negrito. Em caso de empate, o critério de desempate foram os falso-negativos e falso-positivos, nesta ordem e estas situações aparecem sublinhadas na tabela. De forma

resumida, o arquivo original teve melhor desempenho em apenas um dos vinte casos analisado. A filtragem por média gaussiana ajudou a superar pelo menos um dos outros dois métodos concorrentes em seis casos. Finalmente, a aplicação da filtragem pela média associada à subtração de quadros consecutivos, fez com que a classificação superasse pelo menos uma das demais em 7 ocasiões.

Tabela 4 – Comparativo da detecção utilizando HOG

		FILTROS								
		ARQUIVO ORIGINAL			MÉDIA GAUSSIANA			Média+Subtração Quadros		
Vídeo	Conteúdo	VP	FP	FN	VP	FP	FN	VP	FP	FN
P01	6 pessoas	5	0	1	5	0	1	5	0	1
P02	2 pessoas	2	1	0	2	1	0	2	0	0
P03	2 pessoas	2	0	0	2	1	0	2	0	0
P04	2 pessoas	2	0	0	2	0	0	2	0	0
P05	5 pessoas	4	0	1	**5**	2	0	**5**	0	0
P06	2 pessoas	2	0	0	2	1	0	2	0	0
P07	1 pessoa	1	0	0	1	0	0	1	0	0
P08	1 pessoa	0	0	1	**1**	0	0	**1**	0	0
P09	3 pessoas	3	0	0	3	0	0	3	0	0
P10	5 pessoas	5	0	0	5	0	0	5	0	0
P11	4 pessoas	1	0	3	**2**	0	2	**2**	0	2
P12	3 pessoas	**3**	0	0	2	0	1	**3**	0	0
P13	1 pessoa	1	1	0	1	1	0	1	0	0
P14	2 pessoas	2	0	0	2	1	0	2	0	0
P15	4 pessoas	4	1	0	4	1	0	4	0	0
P16	2 pessoas	2	1	0	2	1	0	2	0	0
P17	4 pessoas	3	0	1	**4**	0	0	3	0	1
P18	5 pessoas	3	0	2	**5**	0	0	4	0	1
P19	3 pessoas	3	0	0	3	1	0	3	0	0
P20	2 pessoas	1	0	0	**2**	2	0	1	0	1

Fonte: produção do próprio autor

Se por um lado a filtragem isolada teve um desempenho superior, pois o classificador conseguiu identificar a maioria das pessoas, por outro lado ela introduziu o problema do aumento dos falso-positivos. Já a média gaussiana, seguida da subtração de quadros foi a

que teve melhor desempenho considerando-se o objetivo de minimizar os falso-positivos, praticamente eliminando-os. Portanto, há espaço para a busca de uma forma para equilibrar melhor os resultados e desempenhos dos filtros média e média+subtração de quadros, extraindo a melhor característica de cada um deles.

Aplicou-se o mesmo processamento para arquivos que continham apenas carros ou motos. São em torno de 45 imagens de automóveis. Neste caso, foi contado o número de falso-positivos e chegou-se aos resultados apresentados na Tabela 5, onde é possível perceber que o filtro Média+Subtração Quadros teve menor quantidade de falso-positivos.

Tabela 5 – Análise de falso-positivo em imagens contendo somente carros

Conteúdo	ARQUIVO ORIGINAL	MÉDIA GAUSSIANA	Média+Subtração Quadros
Automóveis – arquivos 1 ao 25	Não foram testados	100% deles tiveram pelo menos 1 Falso-Positivo	**76% deles tiveram pelo menos 1 Falso-Positivo**

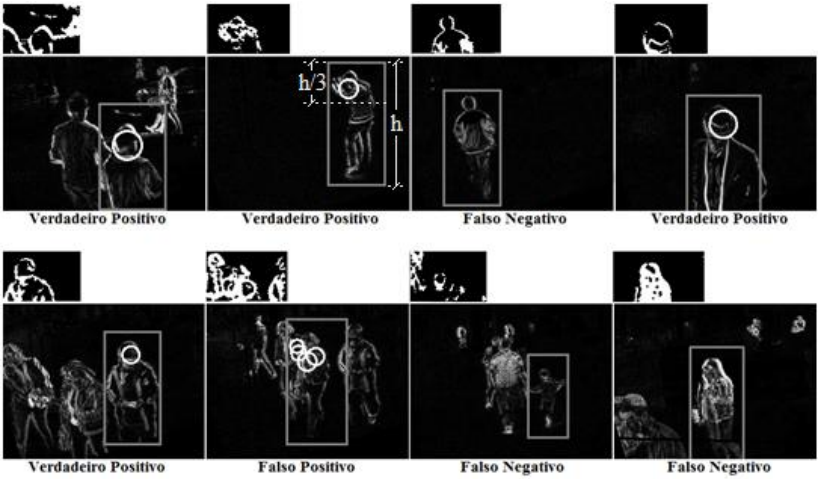
Fonte: produção do próprio autor

5.5 EXPERIMENTO 4

Outro experimento realizado, diz respeito ao uso da transformada de Hough com a finalidade de se localizar as cabeças das pessoas. Esta busca é feita através do uso da transformada de Hough em uma região de interesse restrita à terça parte superior dos retângulos que envolvem as pessoas (ver Figura 37). Na Figura 37 são apresentados alguns exemplos de resultados. Acima de cada imagem capturada, em tamanho menor, aparece a região de interesse obtida seguindo o critério da busca por cabeças na terça parte superior do retângulo, porque em todas as imagens do banco de dados utilizadas para treinamento, subdividir a janela encontrada pelo método HOG em três partes foi suficiente para garantir que as cabeças estivessem circunscritas na terça parte superior do retângulo envolvente.

Estas regiões de interesse, foram binarizadas para destacar ainda mais as bordas, antes de se aplicar a transformada de Hough.

Figura 37 – Tentativas de detecção de cabeças pela transformada de Hough



Fonte: produção do próprio autor

Observando-se o campo de fluxo óptico é possível constatar que a direção dos vetores no caso de objetos maiores como os carros, e que não tem possibilidade de movimentação relativa de suas partes constitutivas, apresentam um alinhamento harmônico de quase todos os vetores. Já no caso das pessoas, por terem movimentação relativa de braços e pernas principalmente, o campo do fluxo óptico tende a ter um comportamento mais disperso, onde os vetores podem ter direções bem diferentes para a totalidade do objeto pessoa. De fato, muitos trabalhos relacionados, como o proposto por Dalal et. al. (2006) na área de reconhecimento de comportamento usam a informação do campo vetorial do fluxo óptico para extrair características capazes de detectar não somente pessoas, mas também classificar algumas atividades humanas, por exemplo. O diferencial da abordagem proposta no presente trabalho é explorar o uso de descritores do ponto de vista estatístico, utilizando medidas de dispersão e de densidade (Figura 31), como por exemplo o desvio padrão e as janelas de Parzen respectivamente, de forma que independentemente do ajuste de instalação da câmera (inclinação/perspectiva), seja possível utilizar o mesmo sistema classificatório para se detectar as pessoas, sem que haja necessidade de treinamento para cada situação específica.

5.6 EXPERIMENTO 5

Com base na premissa estabelecida no experimento 4 de que o campo de fluxo óptico tende a ter os vetores mais alinhados comparativamente ao campo gerado por pessoas caminhando, projetou-se o teste para se determinar qual a melhor relação N x N em pixels para se extrair informação útil da análise do campo de fluxo óptico a partir das medições estatísticas listadas no capítulo 4. Como métrica foi usada a matriz de confusão pessoas e não-pessoas (carros). O *ground truth* foi determinado manualmente, pela marcação de retângulos envolventes em torno das pessoas e não pessoas (carros) nos arquivos capturados. Utilizou-se o Weka para processar os dados, através do processo de validação cruzada (do inglês *cross-validation*) com 10 divisões e os resultados estão listados na Tabela 6.

Para determinar se as regiões demarcadas pelo classificador estão de acordo com o *ground truth* estabelecido, é empregado o critério de medição usado tanto por Everingham et al. (2010) quanto por Dóllar et al. (2012), onde a sobreposição das áreas dos retângulos envolventes (RE) resultantes das detecções, com os RE das marcações deve exceder 50% (Equação 37).

$$ao = \frac{area(REdet \cap REmarc)}{area(REdet \cup REmarc)} > 0.5 \qquad (37)$$

Tabela 6 – Desempenho da classificação em função da subdivisão NxN

		VP	FP
6x6 pixels	Pessoas	0,965	0,159
	não-pessoas	0,841	0,035
10x10 pixels	Pessoas	0,968	0,144
	não-pessoas	0,856	0,032
12x12 pixels	Pessoas	0,965	0,123
	não-pessoas	0,877	0,035
16x16 pixels	Pessoas	0,959	0,116
	não-pessoas	0,884	0,041
20x20 pixels	Pessoas	0,952	0,166
	não-pessoas	0,834	0,048

Fonte: produção do próprio autor

Pelos dados obtidos com o segundo experimento, é possível verificar o potencial de se utilizar informações estatísticas do comportamento do fluxo óptico para se classificar e separar pessoas de

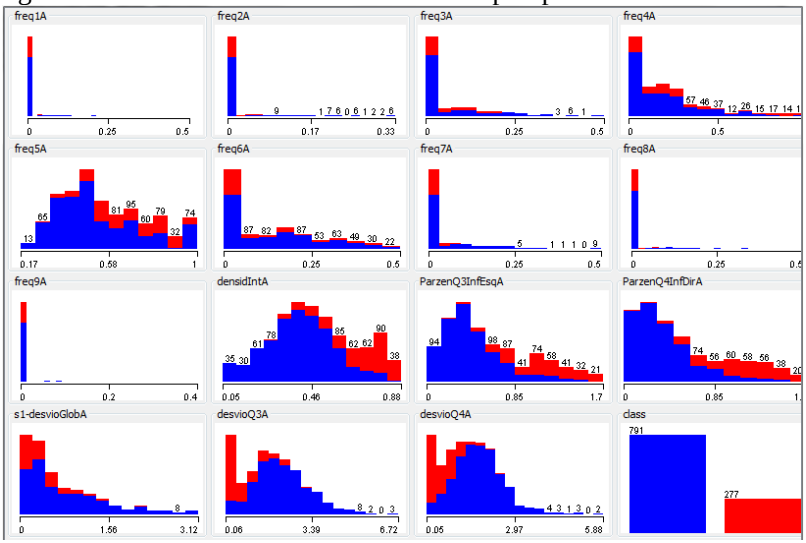
não-pessoas (no caso bem específico de carros, que é justamente o escopo deste trabalho).

5.7 EXPERIMENTO 6

Como as marcações são feitas em função das regiões candidatas selecionadas por HOG, com a finalidade de aperfeiçoar a classificação utilizando esta informação como realimentação para um novo treinamento, mas a partir de uma metodologia diferente, qual seja, explorando descritores obtidos do campo de fluxo óptico, é de se esperar que o melhor desempenho seja alcançado quando primeiramente se faz a seleção por HOG e depois é aplicada a análise do fluxo óptico como filtro adicional e independente em cascata. Independentemente disto, foram testadas as duas situações, conforme Figura 30. Novamente as métricas utilizadas são os verdadeiro-positivos e falso-positivos, dispostos na forma de matriz de confusão para se determinar qual proposta tem o melhor desempenho.

Na Figura 38 estão relacionadas as características extraídas a partir da informação campo vetorial de fluxo óptico conforme já descrito no capítulo 4 sobre o modelo proposto (em azul, tem-se a classe das pessoas, enquanto a classe dos carros é representada pela cor vermelha).

Figura 38 – Características extraídas do campo óptico

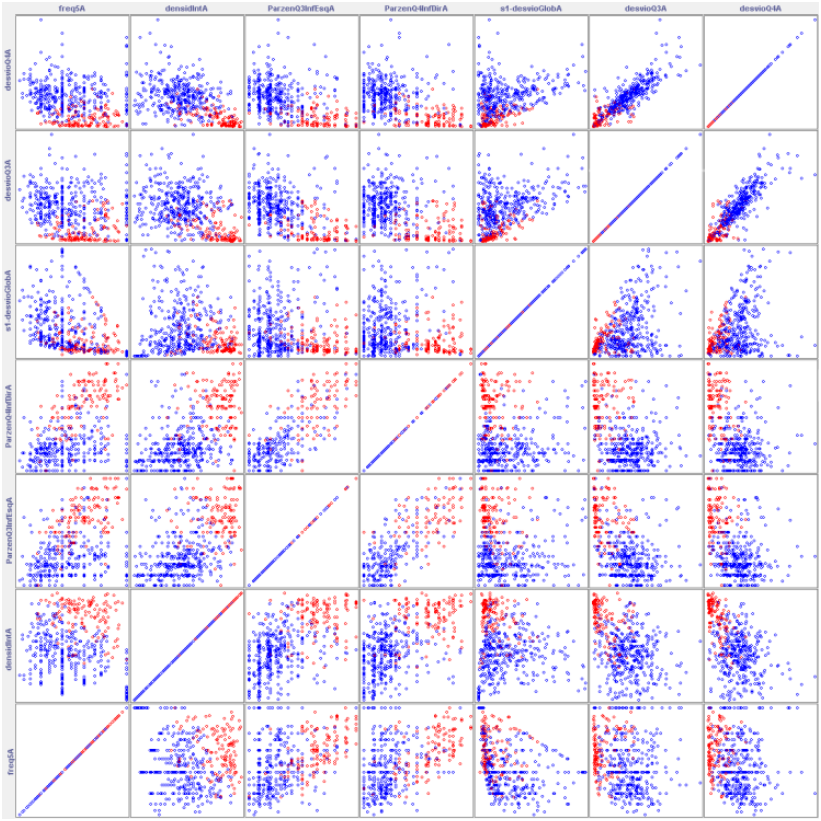


Fonte: produção do próprio autor

A dispersão das características tomadas duas a duas pode ser visualizada na Figura 39. Algumas associações separam melhor os dados do que outras, e é a partir destas informações que a máquina SVM é treinada para aprender a separar as classes pessoas de não-pessoas (ou carros). Pela Figura 39 é possível ver que algumas associações separam melhor as classes pessoas (em azul) dos carros (vermelho). É justamente aí que as SVM atuam, fazendo a separação linear utilizando alguns vetores suportes na fronteira entre os dois grupos.

Ao treinar a SVM com polinômio de grau 1 (linear) a partir dos atributos selecionados do campo fluxo óptico, atribuindo-se peso 10 às pessoas e peso 1 aos carros e validando os resultados nas imagens de testes, obtém-se um reconhecimento das pessoas da ordem de 95,83%.

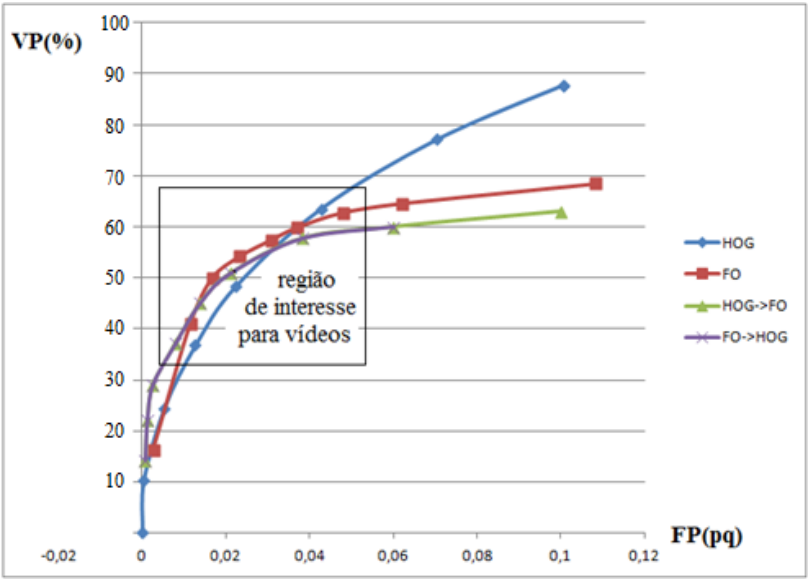
Figura 39 – Dispersão das características 2 a 2



Fonte: produção do próprio autor

Adicionar a informação de fluxo óptico traz melhorias ao reconhecimento ou classificação, conforme pode ser visto no gráfico ROC (Figura 40). No eixo das ordenadas, tem-se o percentual de acerto dos verdadeiro-positivos (VP) e no eixo das abscissas tem-se o valor relativo dos falso-positivos em relação à quantidade total de quadros analisados, ou seja, falso-positivos por quadro FP(pq). Nota-se que o gráfico em vermelho (classificação por FO) tem melhor desempenho que o gráfico em azul (somente HOG) dentro da região de interesse para vídeos, pois nesta região destacada por um retângulo a ocorrência de falso-positivos é baixa. Nesta região a curva FO (quadrados em vermelho) atinge maiores ocorrências de verdadeiro-positivos comparativamente com HOG (losangos em azul).

Figura 40 – Comparativo curva ROC dos classificadores



Fonte: produção do próprio autor

5.8 EXPERIMENTO 7

E se, ao invés de se explorar a informação do movimento restrito a apenas dois quadros consecutivos que é justamente o que o fluxo óptico se propõe a registrar, a análise fosse extensiva a mais quadros consecutivos, como por exemplo 3, 4, 5 ou até mais quadros,

através do armazenamento das informações estatísticas numa matriz tridimensional, o desempenho da classificação seria melhor? Com base neste questionamento o presente teste foi projetado para determinar se o tensor de fluxo óptico de mais do que dois quadros consecutivos traz benefícios à separação entre pessoas e não-pessoas (carros). Os resultados obtidos foram sumarizados na Tabela 7. Percebe-se que ao aumentar a quantidade de quadros analisados a quantidade de verdadeiro-positivos aumenta e a quantidade de falso-positivos diminui, o que é desejável.

Tabela 7 – Influência da inclusão de sucessivos quadros

Quadros Sucessivos Acumulados	Verdadeiro-Positivos	Falso-Positivos
2	83,8%	16,2%
3	85,5%	14,5%
4	87,1%	12,9%
5	**89,1%**	**10,9%**

Fonte: produção do próprio autor

5.9 EXPERIMENTO 8

Este experimento concentra-se na avaliação de desempenho MLP x SVM para reconhecer padrões ao se analisar exclusivamente as características extraídas a partir do campo fluxo óptico objetivando separar as classes pessoas e carros. Foram utilizadas 1680 imagens (quadros de vídeos), contendo 25 carros (277 marcações) e 69 pessoas (1388 marcações). Como métricas, são considerados percentual de instâncias corretamente classificadas e taxas de verdadeiro-positivos e falso-positivos. Conforme dados apresentados na Tabela 8, os percentuais de instâncias classificadas corretamente usando MLP e SVM ficaram muito próximos, bem como o percentual de verdadeiro-positivos de pessoas.

Tabela 8 – Comparativo do desempenho da classificação MLP x SVM

Atributos	MLP	SVM
Instâncias Classificadas Corretamente	**88,153%**	87,777%
Verdadeiro-Positivo PESSOAS	94,3%	**96,9%**
Falso-Negativo PESSOAS	5,7%	**3,1%**
Falso-Positivo PESSOAS	**28,6%**	37,3%
Verdadeiro-Positivo CARROS	**71,4%**	62,7%
Falso-Negativo CARROS	**28,6%**	37,3%
Falso-Positivo CARROS	5,7%	**3,1%**
Precisão PESSOAS	**90,0%**	87,7%
Revocação PESSOAS	94,3%	**96,7%**

Fonte: produção do próprio autor

5.10 EXPERIMENTO 9

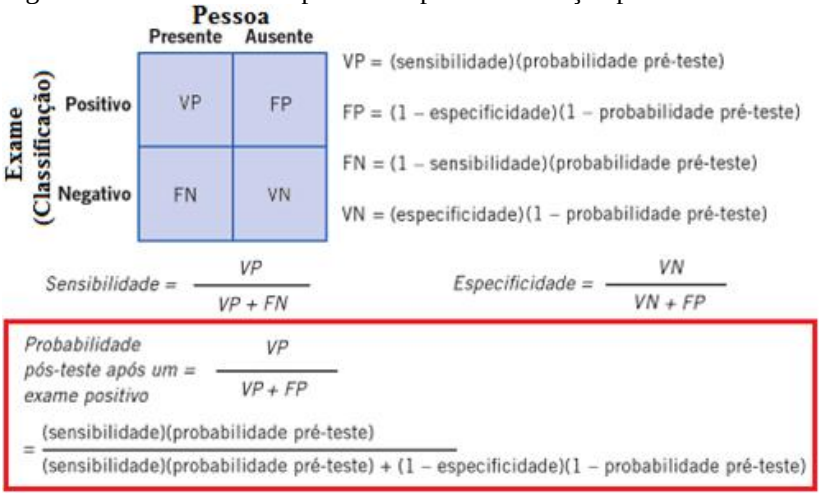
Gerar um alarme toda vez que uma pessoa é localizada em algum quadro isolado não é viável, até porque se a medição fosse realizada desta forma a confiabilidade da resposta seria baixa, além de que todo falso-positivo também geraria um alarme, causando descrédito dos alarmes gerados. O evento intrusão, portanto, deve ser avaliado como a ocorrência persistente da detecção de pessoa por parte do classificador.

O evento intrusão é definido como a existência de pelo menos uma pessoa circulando dentro da região delimitada por cerca virtual durante alguns quadros sucessivos.

Devido a ausência na literatura de trabalhos relacionados com formas de se avaliar a eficácia das detecções do evento intrusão e geração de alarmes, recorreu-se a conceitos empregados no campo da medicina. É comum estudos na medicina para se avaliar as probabilidades de uma pessoa ser diagnosticada positiva ou negativa em determinado exame e realmente estar correto o resultado. Por exemplo, se uma gestante realiza um teste de gravidez várias vezes, com alternâncias de resultados positivos e negativos, qual a probabilidade de ela realmente estar grávida ou não? Ou se um paciente com câncer realiza duas vezes o mesmo teste, com resultados opostos, em qual deve confiar e com qual nível de certeza? Para responder estes questionamentos, a medicina usa os conceitos da teoria bayesiana, partindo de uma probabilidade pré-teste para calcular a probabilidade

pós-teste sucessivamente. As fórmulas são apresentadas na Figura 41 [NICOLL et al. 2014].

Figura 41 – Probabilidade pós-teste após classificação positiva



Fonte: adaptado de Nicoll et al., 2014

Utilizando os dados referentes ao banco de dados usados para os testes finais, e a teoria de Bayes [NICOLL et al. 2014], chega-se à conclusão de que uma boa estratégia de parametrização para a detecção do evento intrusão é, no melhor caso, detectar uma pessoa em três quadros consecutivos e no pior caso, após detectar uma pessoa, ela pode ficar no máximo três quadros sem ser detectada, até que seja detectada novamente e este processo se repita até que três detecções sejam acumuladas e contadas (ver Tabelas 9 a 13). Em todas estas tabelas, os dados de sensibilidade e especificidade são fixos, pois são informações inerentes ao desempenho do sistema de classificação. A probabilidade pré-teste de "haver pessoas" apresentada no cabeçalho das tabelas foi obtida a partir do banco de dados. No banco de dados criado, 52,90% dos quadros possuem pelo menos 1 pessoa, enquanto 47,10% não possuem pessoas presentes na cena. Estas probabilidades de haver ou não pessoas é que vão mudando (pós-teste), dependendo se o classificador localiza ou não uma pessoa. Seja uma sequência de vídeo que contém pelo menos uma pessoa. Sendo de 52,90% a probabilidade inicial de haver pelo menos uma pessoa na cena, e a sensibilidade do classificador ser de 54,15% chega-se aos valores de 28,64% (52,90% x 54,15%) e 24,26% (52,90% - 28,64%) para as probabilidades do

classificador acertar quando diz ter detectado uma pessoa. Já para o caso da probabilidade inicial de 47,10% de não haver pessoas, sendo a especificidade do classificador de 91,76%, chega-se aos valores de 43,22% (47,10% x 91,76%) e 3,88% (47,10% - 43,22%) para as probabilidades do classificador estar certo quando diz não ter detectado uma pessoa. Combinando as duas informações sobre sensibilidade e especificidade da classificação, chega-se ao percentual de 88,07% ou $(\frac{28,64\%}{(28,64\%+3,88\%)})$ de probabilidade pós-teste de haver pessoa quando o classificador de fato detectar pessoa. Esta informação pode ser vista nas Tabelas 9 a 13. Após assumir que o classificador detectou uma pessoa, o resultado de 88,07% passa a ser a nova probabilidade de haver pessoas. Na Tabela 10, por exemplo, como o resultado da segunda classificação foi não ter achado uma pessoa, a probabilidade cai de 88,07% para 78,67% ou $(\frac{40,38\%}{(40,38\%+10,95\%)})$ e assim sucessivamente as tabelas vão sendo atualizadas, dependendo da informação prévia (probabilidade pré-teste) e do resultado de classificação (probabilidade pós-teste). A probabilidade nestes casos é vista como uma informação cumulativa que carrega o histórico de todo um conjunto de detecções anteriores.

Usando esta estratégia de avaliação possibilitou parametrizar os valores dos gatilhos dos contadores de presença e ausência de pessoas.

Tabela 9 – Caso I teoria de Bayes aplicada aos dados testes finais

OffsetFo = 0,25 (deslocamento do hiperplano da SVM FO) Sensibilidade do classificador = 0,5415 Especificidade do classificador = 0,9176 Probab. haver pessoas (número quadros com pessoa/número total quadros) = 0,529029								
CASO I acha 4 vezes seguidas sendo que em todos 4 quadros há pessoa	Qua-dro	probab. haver pessoas	classif acha pessoa	classif não acha pessoa	probab. não haver pessoa	classif não acha pessoa	classif acha pessoa	Nova probab. haver pessoa
	1	52,90	28,64	24,26	47,10	43,22	3,88	0,8807
	2	88,07	47,69	40,38	11,93	10,95	0,98	0,9798
	3	97,98	53,06	44,92	2,02	1,85	0,17	**0,9969**
	4	99,69	53,98	45,71	0,31	0,29	0,03	0,9995

Fonte: produção do próprio autor

Em cada uma das tabelas foi destacada a probabilidade de se estar diante de uma pessoa após a terceira detecção positiva por parte do classificador, sendo que em cada um dos cinco casos foram intercaladas diferentes quantidades de não-detecção entre detecções positivas.

Tabela 10 – Caso II teoria de Bayes aplicada aos dados testes finais

OffsetFo = 0,25 (deslocamento do hiperplano da SVM FO) Sensibilidade do classificador = 0,5415 Especificidade do classificador = 0,9176 Probab. haver pessoas (número quadros com pessoa/número total quadros) = 0,529029								
CASO II acha 1 não acha 1 acha 1 não acha 1 acha 1 não acha 1 acha 1 sendo que em todos 7 quadros há pessoa	Qua-dro	probab. haver pessoas	classif acha pessoa	classif não acha pessoa	probab. não haver pessoa	classif não acha pessoa	classif acha pessoa	Nova probab. haver pessoa
	1	52,90	28,64	24,26	47,10	43,22	3,88	0,8807
	2	88,07	47,69	40,38	11,93	10,95	0,98	0,7867
	3	78,67	42,60	36,07	21,33	19,57	1,76	0,9604
	4	96,04	52,00	44,03	3,96	3,64	0,33	0,9237
	5	92,37	50,02	42,35	7,63	7,00	0,63	**0,9876**
	6	98,76	53,48	45,28	1,24	1,14	0,10	0,9755
	7	97,55	52,82	44,73	2,45	2,25	0,20	0,9962

Fonte: produção do próprio autor

Tabela 11 – Caso III teoria de Bayes aplicada aos dados testes finais

OffsetFo = 0,25 (deslocamento do hiperplano da SVM FO) Sensibilidade do classificador = 0,5415 Especificidade do classificador = 0,9176 Probab. haver pessoas (número quadros com pessoa/número total quadros) = 0,529029								
CASO III acha 1 não acha 2 acha 1 não acha 2 acha 1 não acha 2 acha 1 sendo que em todos 10 quadros há pessoa	Qua-dro	probab. haver pessoas	classif acha pessoa	classif não acha pessoa	probab. não haver pessoa	classif não acha pessoa	classif acha pessoa	Nova probab. haver pessoa
	1	52,90	28,64	24,26	47,10	43,22	3,88	0,8807
	2	88,07	47,69	40,38	11,93	10,95	0,98	0,7867
	3	78,67	42,60	36,07	21,33	19,57	1,76	0,6483
	4	64,83	35,10	29,72	35,17	32,28	2,90	0,9237
	5	92,37	50,02	42,35	7,63	7,00	0,63	0,8582
	6	85,82	46,47	39,35	14,18	13,01	1,17	0,7515
	7	75,15	40,69	34,46	24,85	22,80	2,05	**0,9521**
	8	95,21	51,56	43,65	4,79	4,40	0,39	0,9085
	9	90,85	49,20	41,65	9,15	8,40	0,75	0,8323
	10	83,23	45,07	38,16	16,77	15,39	1,38	0,9702

Fonte: produção do próprio autor

Tabela 12 – Caso IV teoria de Bayes aplicada aos dados testes finais

OffsetFo = 0,25 (deslocamento do hiperplano da SVM FO) Sensibilidade do classificador = 0,5415 Especificidade do classificador = 0,9176 Probab. haver pessoas (número quadros com pessoa/número total quadros) = 0,529029								
	Qua-dro	probab. haver pessoas	classif acha pessoa	classif não acha pessoa	probab. não haver pessoa	classif não acha pessoa	classif acha pessoa	Nova probab. haver pessoa
CASO IV acha 1 não acha 3 acha 1 não acha 3 acha 1 não acha 3 acha 1 sendo que em todos 13 quadros há pessoa	1	52,90	28,64	24,26	47,10	43,22	3,88	0,8807
	2	88,07	47,69	40,38	11,93	10,95	0,98	0,7867
	3	78,67	42,60	36,07	21,33	19,57	1,76	0,6483
	4	64,83	35,10	29,72	35,17	32,28	2,90	0,4794
	5	47,94	25,96	21,98	52,06	47,77	4,29	0,8582
	6	85,82	46,47	39,35	14,18	13,01	1,17	0,7515
	7	75,15	40,69	34,46	24,85	22,80	2,05	0,6017
	8	60,17	32,58	27,59	39,83	36,54	3,28	0,4302
	9	43,02	23,30	19,72	56,98	52,29	4,70	**0,8323**
	10	83,23	45,07	38,16	16,77	15,39	1,38	0,7126
	11	71,26	38,59	32,67	28,74	26,37	2,37	0,5533
	12	55,33	29,96	25,37	44,67	40,99	3,68	0,3823
	13	38,23	20,70	17,53	61,77	56,68	5,09	0,8027

Fonte: produção do próprio autor

Tabela 13 – Caso V teoria de Bayes aplicada aos dados testes finais

OffsetFo = 0,25 (deslocamento do hiperplano da SVM FO) Sensibilidade do classificador = 0,5415 Especificidade do classificador = 0,9176 Probab. haver pessoas (número quadros com pessoa/número total quadros) = 0,529029								
	Qua-dro	probab. haver pessoas	classif acha pessoa	classif não acha pessoa	probab. não haver pessoa	classif não acha pessoa	classif acha pessoa	Nova probab. haver pessoa
CASO V acha 1 não acha 4 acha 1 não acha 4 acha 1 não acha 4 acha 1 sendo que em todos 16 quadros há pessoa	1	52,90	28,64	24,25	47,09	43,21	3,88	0,8807
	2	88,06	47,68	40,37	11,93	10,94	0,98	0,7867
	3	78,67	42,60	36,07	21,32	19,57	1,75	0,6483
	4	64,82	35,10	29,72	35,17	32,27	2,89	0,4794
	5	47,94	25,96	21,98	52,05	47,76	4,28	0,3151
	6	31,51	17,06	14,44	68,48	62,84	5,64	0,7515
	7	75,14	40,69	34,45	24,85	22,80	2,04	0,6017
	8	60,17	32,58	27,59	39,82	36,54	3,28	0,4302
	9	43,01	23,29	19,72	56,98	52,28	4,69	0,2739
	10	27,39	14,83	12,55	72,60	66,62	5,98	0,1586
	11	15,86	8,58	7,27	84,13	77,20	6,93	**0,5533**
	12	55,33	29,96	25,36	44,66	40,98	3,68	0,3823
	13	38,23	20,70	17,52	61,76	56,67	5,08	0,2362
	14	23,62	12,79	10,83	76,37	70,08	6,29	0,1339
	15	13,38	7,24	6,13	86,61	79,47	7,13	0,0717
	16	7,16	3,88	3,28	92,83	85,18	7,64	0,3366

Fonte: produção do próprio autor

Partindo da arquitetura da Figura 32, o evento intrusão é medido através de dois acumuladores: um contador de presença de pessoa e outro contador de ausência de pessoa. Ambos começam em zero. Toda vez que o classificador detecta uma pessoa, o contador de presença é incrementado e o contador de ausência, zerado. Na sequência, a cada quadro subsequente em que nenhuma pessoa é localizada, o contador de ausência é incrementado. Caso o contador de ausência ultrapasse três (quatro quadros sem detectar pessoa), então ambos os contadores são zerados. Este processo se repete até que o contador de presença de pessoa acumule o valor três, situação em que o

alarme de intrusão é disparado. Foram escolhidos o valor 3 para o gatilho de ausência de pessoas (gatilhoA>=3) e 3 para o gatilho de presença de pessoas (gatilhoP>=3), porque o sistema configurado desta forma garante probabilidade mínima de 83,23% de haver pessoa em caso de geração de alarmes, conforme pode ser visto no valor destacado na última coluna da Tabela 12. Caso o valor usado para este o gatilho de ausência fosse 4, esta probabilidade cairia para 55,33%, conforme valor destacado na última coluna da Tabela 13.

Com esta estratégia de parametrização, e em sistemas de aquisição de imagem de 16 quadros por segundo, na melhor situação uma invasão seria detectada em 0.1875s enquanto no pior caso, levaria 0.5626s para ser percebida. O classificador FO foi treinado com imagens capturadas a uma taxa de 16 quadros por segundo (fps).

A partir da primeira detecção de pessoa presente na cena, caso ela se confirme em mais 3 quadros dentro de uma janela de tolerância de não recorrência da detecção (ausência de detecção), então um alerta de intrusão é gerado. A janela de ausência de detecção foi ajustada em 3 quadros. Assim, para que um alerta seja gerado, na melhor das hipóteses são necessários 3 quadros, onde em todos eles consecutivamente o classificador em cascata FO→HOG detectou a presença de pessoas. No caso limite, seriam necessários 9 quadros (correspondentes a aproximadamente meio segundo de tempo real de vídeo adquirido a uma taxa de 16 quadros por segundo) para se gerar um alerta, situação na qual a pessoa é detectada no primeiro quadro, nos 3 quadros seguintes nenhuma ocorrência é detectada até que no quarto quadro a detecção de pessoa ocorre novamente e o ciclo se repete até o 9º quadro, quando ocorrerá a geração efetiva do alerta de intrusão por ultrapassagem da cerca virtual. Sempre que 4 quadros ou mais ocorrem sem nenhuma detecção, ambas as contagens são zeradas e uma nova série de contagem reestabelecida a partir da próxima detecção de pessoa. Mesmo que um quadro apresente mais de uma pessoa, a contagem é feita por quadros que apresentam uma pessoa ou mais. Em outras palavras, mesmo que um quadro tenha 3 pessoas detectadas, a contagem incrementa em um o contador de ocorrências até que se chegue a 3 quadros acumulados, dentro do limite de ausência de 3 quadros.

5.11 SIMULAÇÕES

Por fim, uma simulação foi realizada como teste final para avaliar o desempenho da estratégia proposta para detectar o evento

intrusão. Foram comparados os seguintes casos: (1) classificação HOG sem deslocamento do hiperplano da SVM, (2) classificação HOG→FO também sem deslocamentos dos hiperplanos de ambas SVM, (3) classificação HOG→FO com deslocamento -0,15 para HOG e deslocamento +0,25 para FO, que corresponde ao ponto de melhor desempenho obtido a partir da análise das curvas ROC dentro da região de interesse para vídeos, (4) classificação HOG com deslocamento de hiperplano de +0,31, (5) classificação HOG com deslocamento de hiperplano de +0,50 e finalmente (6) classificação HOG com deslocamento de hiperplano de +0,59. As simulações (4) e (5) foram escolhidas para serem comparadas diretamente com a simulação (3), enquanto a simulação (6) é a comparação direta com a simulação (2). Para que as simulações possam ser consideradas equivalentes, a quantidade de verdadeiro-positivos e/ou a quantidade de falso-positivos são as mesmas para as simulações sob comparação e estes valores foram obtidos a partir da interpolação dos pontos das tabelas das matrizes de confusão (ver Tabelas 17 a 20 em Apêndice A). Por este motivo foram escolhidos os valores de separação de hiperplano +0,31, +0,50 e +0,59.

Após concluídas as 6 simulações, considerando os 54 eventos da Tabela 16 (vídeos testes finais), o melhor resultado obtido utilizando apenas HOG (deslocamento zero do hiperplano da SVM), resultou em 20 verdadeiro-positivos, 15 verdadeiro-negativos, 5 falso-negativos e 15 falso-positivos. Caso se tentasse reduzir os falso-positivos deslocando-se o valor do hiperplano com incrementos positivos, então os falso-negativos aumentariam consideravelmente, conforme pode ser visto na Tabela 18 no Apêndice A.

O classificador FO→HOG gerou 22 verdadeiro-positivos, 29 verdadeiro-negativos, 2 falso-negativos e 1 falso-positivo. Os dois falso-negativos são situações difíceis de detectar. Em um deles, o invasor entra escondido atrás de um carro e de forma curvada. Na outra situação não alarmada, o invasor tinha uma altura em pixels insuficiente, e apesar dele estar na posição ereta, a pessoa parecia inclinada por falta de calibração da câmera. Destaca-se que em nenhuma das simulações foi possível detectar estes dois casos extremos.

O método proposto de detecção de intrusão utilizando classificador FO→HOG tem uma acurácia superior comparado à utilização de classificadores HOG. Este último, além de deixar de alarmar outros 3 eventos, gerou uma grande quantidade de alertas falso-positivos (15 ao todo). O método proposto utilizando conjuntamente FO→HOG, apesar dos verdadeiro-positivos terem caído quase pela metade se comparados os dados das Tabelas 18 e 20 do Apêndice A,

conseguiu obter mais verdadeiro-positivos nos alarmes de detecção de intrusão. O resultado mais expressivo foi em relação aos falso-positivos, chegando a uma acurácia 15 vezes superior ao obtido somente com classificação por HOG. Todas as combinações podem ser vistas na Tabela 21 no Apêndice A.

Por fim, foram levantadas as respostas para outras combinações de valores dos contadores de presença (gatilhoP) e de ausência (gatilhoA) nas redondezas de gatilhoP=3 e gatilhoA=3, com um fator de redução de escala de 1.30 para cada varredura HOG. Os dados da Tabela 22 referem-se ao sistema completo, com a análise do comportamento do FO sendo usado para filtrar as imagens antes de repassá-las ao classificador HOG, enquanto a Tabela 23 apresenta os resultados considerando que o classificador FO esteja desativado. Novamente fica evidente a importância da informação da movimentação para separar pessoas de carros, pois os falso-positivos reduzem significativamente quando se usa o classificador FO. Além disto, os resultados comprovam que a estratégia usada para parametrizar os contadores funciona, pois considerando que o objetivo é a maximização dos verdadeiro-positivos (que é o mesmo que minimizar os falso-negativos), seguido da minimização dos falso-positivos, então o melhor ponto de operação é gatilhoP=3 e gatilhoA=3 conforme estimado (vide Tabela 22 no Apêndice A).

5.12 DISCUSSÃO DOS RESULTADOS

Analisando os gráficos traçados em uma mesma escala (Figura 40), percebe-se que a classificação baseada nas características extraídas do campo de vetores do FO tem um desempenho superior na região que vai do zero ao 60% dos verdadeiro-positivos, ponto a partir do qual sofre degradação e é totalmente superado pelo classificador HOG isoladamente. Dependendo do contexto e das exigências impostas por cada situação, pode-se optar por trabalhar com a classificação por HOG isoladamente, como por exemplo detecção de pessoas em imagens estáticas. Entretanto, para geração de alarmes de eventos de intrusão, isto é inviável e deve-se usar ambos os classificadores em cascata, sendo necessária a confirmação de ambos para garantir a detecção de pessoas em quadros isolados. Por exemplo, no caso específico de análise de vídeos, no âmbito de imagens de sistemas de vigilância, a ocorrência de falso-positivos em excesso causa um descrédito nos eventuais alertas gerados. Portanto, o ideal é se trabalhar na faixa inferior da escala, com

taxas de verdadeiro-positivos entre 40% e 60% e taxas de falso-positivos por quadro de vídeo entre 0 e 0,04.

Embora a taxa entre 40% e 60% possa parecer relativamente baixa, implicando na eventual perda de geração de alerta de pessoas entrando em regiões definidas pelas cercas virtuais, este problema é contornado pelo uso da probabilidade estatística. Cada segundo de vídeo é composto de 16 quadros. Considerando que as classificações individuais quadro a quadro são eventos distintos, e que se esteja trabalhando com taxas de acerto em torno de 50% para os verdadeiro-positivos, tem-se após a identificação potencial de pessoas em 3 casos, ou melhor, em 3 quadros consecutivos que a probabilidade, segundo a teoria de Bayes, de se estar realmente diante de uma pessoa é de 99,69% (ver Tabela 9). Deste ponto de vista, a estratégia para mensurar o evento intrusão foi estabelecida seguindo o critério de consistência ou persistência das detecções e se revelou eficaz na geração de alertas.

Salientando mais uma vez que o classificador FO não é um classificador novo e independente, no sentido estrito, pois ele é assessório ao classificador HOG. Dito isto, para se obter a melhor resposta que é em torno da curva FO até o nível de acerto de 60% dos VP, é possível da seguinte formas: primeiro, fixando o deslocamento do hiperplano do classificador FO em torno de 0,25 e variando-se o valor do hiperplano do classificador HOG na faixa de -2,0 a +2,0 e traçar o gráfico correspondente a partir das matrizes de confusão obtidas ponto a ponto.

Analisando as matrizes de confusão do classificador HOG sem deslocamento do hiperplano da SVM (Tabela 18), com às do classificador FO com deslocamento de hiperplano SVM de +0,25 seguido do classificador HOG (Tabela 20), no primeiro caso para as pessoas tem-se 1012 verdadeiro-positivos contra 137 falso-positivos e no segundo caso, 615 verdadeiro-positivos contra 15 falso-positivos. Ou seja, apesar dos verdadeiro-positivos terem caído quase pela metade (redução de 39,22%), os falso-positivos caíram a uma taxa muito maior, equivalente à 89,05%.

Não é possível realizar uma comparação direta com outros trabalhos relacionados, pois a literatura pesquisada não contempla de forma explícita uma maneira de realizar medições de alarme, e, portanto, este trabalho também propôs uma abordagem para detectar este tipo de evento.

6 CONCLUSÃO

O objetivo deste trabalho é a identificação de praticamente todas as pessoas que venham a ultrapassar um portão virtual (linha ou curva imaginária) definida em determinada região da imagem capturada por um sistema de CFTV. Para tanto, experimentos foram realizados utilizando a técnica HOG, adicionando informações estatísticas a partir de características extraídas do campo de FO e uma SVM.

Utilizou-se o classificador baseado em HOG implementado pela ferramenta OpenCV e o desempenho na detecção de pessoas ficou próximo ao relatado nos trabalhos similares pesquisados. Entretanto, constatou-se que inclinações da câmera maiores do que 15 graus em relação à posição ereta degradaram os resultados e inclinações maiores do que 20 graus tornavam impossível a localização de pessoas pela técnica HOG. O principal motivo para que a utilização de HOG na detecção de humanos não tivesse o mesmo desempenho do relatado na literatura pode ser atribuído ao fato da câmera estar ligeiramente inclinada e consequentemente as pessoas capturadas não estavam necessariamente perpendiculares ao plano horizontal, haja vista que o próprio plano horizontal estava inclinado em relação ao eixo zero grau. Este fenômeno pode ser entendido ao se observar as Figuras 4 e 35, pois nelas, tanto a inclinação da câmera quanto desníveis acentuados no terreno causam projeções ou perspectivas diferentes das pessoas em pé. Mesmo sem a presença de oclusões, poucas pessoas eram detectadas.

A segmentação por FO foi eficiente ao segmentar corretamente todos os quadros que continham pessoas se movimentando.

A utilização da transformada de Hough para localização da cabeça não foi possível segundo os testes realizados. O motivo é que a transformada foi aplicada diretamente na imagem diferencial entre dois quadros consecutivos, o que gerava bordas espessas. Técnicas de afinação ou esqueletização podem amenizar o problema.

As características extraídas do campo de fluxo óptico baseadas em medidas de dispersão e densidade tinham o objetivo de serem menos sensíveis ao ajuste da câmera, principalmente no que diz respeito ao seu posicionamento. Assim, independentemente do local de instalação e campo de visão da câmera, o desempenho da detecção não seria afetado. A classificação através do FO isoladamente, obteve acurácia de 95,83% na detecção das pessoas, com um tamanho de grade de 16x16 pixels. Não é possível garantir esta acurácia para classificação por Fluxo Óptico isoladamente, pois todo o treinamento foi feito apenas em áreas demarcadas previamente pelo classificador HOG. As características

selecionadas foram usadas para treinar uma máquina SVM, mas de uma forma mais imediata, onde a precisão não for um fator crítico no momento da classificação, a combinação de duas características pode substituir a SVM, como por exemplo desvio padrão dos vetores do fluxo óptico do quadrante inferior esquerdo em relação à média global (desvioQ3A, na Figura 39) pelo desvio padrão global destes mesmos vetores (s1-desvioGlobA, também na Figura 39).

Além disto, a combinação em cascata dos classificadores FO conjuntamente com HOG teve um desempenho melhor do que o uso isolado de classificação por características HOG, causando a redução de falso-positivos.

Experimentos foram realizados para avaliar a inclusão indireta de características de movimento na melhoria da classificação. Os testes feitos para 2, 3, 4 e 5 quadros consecutivos classificaram positivamente 83,8%, 85,5%, 87,1% e 89,1% das pessoas, comprovando que o movimento é uma informação útil na detecção de pessoas em vídeos. No presente trabalho não é feita correlação entre classificações sucessivas no sentido de se analisar fluxos ópticos de vários quadros consecutivos e técnicas de rastreamento podem trazer um ganho neste sentido.

No gráfico mostrado na Figura 40, os resultados apontam que a ordem de aplicação dos filtros é irrelevante em função da qualidade da resposta. Ambas as curvas traçadas para os filtros FO→HOG e HOG→FO coincidiram em 100% dos pontos. Já a análise temporal, deixou claro que o a classificação por FO é três vezes mais rápida do que a classificação por HOG (Tabelas 17 e 18).

Como a análise é feita em imagens de vídeos, a perda dos verdadeiro-positivos não é tão crítica e plenamente compensada pela redução significativa dos falso-negativos, ou seja, a inclusão do classificador FO não causou a diminuição de alarmes verdadeiro-positivos e reduziu significativamente a geração de alarmes falsos.

Finalmente, após traçadas as curvas ROC, obteve-se a partir do gráfico o valor do deslocamento apropriado do hiperplano do classificador FO→HOG. O ponto escolhido tem desempenho de 54,15% de acerto dos verdadeiro-positivos, enquanto a utilização de HOG isoladamente tem desempenho de 48,39% para a mesma faixa de FP. Apesar de não parecer significativa do ponto de vista individual, para comprovar a qualidade superior dos resultados obtidos em relação a uma sequência de quadros, como é o caso do evento intrusão, as 6 simulações executadas comprovam que não foi possível obter o mesmo nível de geração de alertas de intrusão usando apenas a classificação por HOG,

independentemente do ponto em que fosse deslocado o hiperplano de sua SVM. Na arquitetura FOGI, o melhor classificador FO→HOG (deslocamento dos hiperplanos HOG=-0,15 e FO=+0,25) atingiu 22 alarmes verdadeiro-positivos, 29 verdadeiro-negativos, 2 falso-negativos e 1 falso-positivo (Tabela 21). Se for comparado com a geração de alarmes pelo uso de classificador HOG isolado na detecção de pessoas, o mais próximo que se chegou foi de 20 alarmes verdadeiro-negativos, 15 verdadeiro-negativos, 5 falso-negativos e 15 falso-positivos. Ao se ignorar a informação do movimento obtida do FO, além do sistema deixar de alarmar 3 eventos a mais (aumento de 150% de FN), também acabava gerando enorme quantidade de alertas falso-positivos (15 ao todo, aumento de 1400%).

6.1 CONTRIBUIÇÕES / RESULTADOS

Os seguintes resultados e contribuições podem ser elencados com a realização deste trabalho:

- Criação de base de dados com finalidade de treinamento e testes contendo vídeos relacionados a intrusão, com a presença de pessoas e carros. O trabalho está disponível em http://www2.joinville.udesc.br/~larva/portal/Projetos.php/vis/115 do Laboratory for Research on Visual Applications (LARVA), UDESC, sob o nome de Projeto de Detecção de Intrusão Utilizando Visão Computacional;
- Implementação e uso de ferramenta para marcação de pessoas, para estabelecer o padrão referencial ou ideal;
- Proposta de uma estratégia ágil e dinâmica para medição de eventos de intrusão com parametrização estabelecida a partir do teorema de Bayes, reduzindo consideravelmente a ocorrência de alarmes falso-positivos (arquitetura FOGI, ver Figura 32) e sem precisar recorrer a técnicas de rastreamento;
- Proposta de um método capaz de reconhecer automaticamente a intrusão de pessoas em qualquer sistema de vigilância, com imagens de resolução 320x240, cujas câmeras descalibradas estejam posicionadas entre -15º e +15º de angulação horizontal.

6.2 CONSIDERAÇÕES FINAIS

Desenvolveu-se um sistema classificatório de pessoas usando conjuntamente HOG e FO em cascata, e a utilização de uma estratégia para detectar eventos de intrusão de pessoas em áreas delimitadas por cerca virtual. Todo o treinamento do classificador FO foi baseado em áreas de verdadeira e falsa classificação de pessoas, bem como o acréscimo de algumas regiões abrangidas por carros. Toda a análise faz sentido apenas quando há movimentação. Regiões estáticas não são avaliadas. O classificador FO tem seu uso atrelado ao uso do classificador HOG em um sistema similar ao de votação, onde ambos precisam concordar com a presença de uma pessoa para que a resposta final efetiva seja positiva para detecção de pessoas.

6.3 TRABALHOS FUTUROS

Existem algumas possibilidades que ainda precisam ser melhor exploradas em trabalhos futuros:

- A subtração de quadros consecutivos gera bordas espessas, impossibilitando o uso direto da transformada de Hough. Técnicas de afinamento e esqueletização poderão ser empregadas para testar a viabilidade de se usar a transformada de Hough na detecção de cabeças e aumentar o nível de confiabilidade da detecção de humanos.
- Como sugestão adicional para trabalhos futuros, as técnicas de rastreamento utilizando extração de características SIFT ou SURF trarão um nível de certeza maior e confiabilidade na contagem de quadros onde pessoas foram apontadas por HOG→FO.
- Outra questão a ser abordada é o uso dos mapas de projeção para tratar melhor os diferentes desvios de inclinação em relação a posição ereta que possam surgir em virtude de diferentes elevações do relevo e também do posicionamento da câmera. Esta é uma forma prática de se calibrar câmeras.
- Além dos mapas de projeção, que trariam uma adaptabilidade maior ao sistema de detecção de intrusão frente às diversas configurações de câmeras, a ampliação das imagens também poderia ser proporcional à posição da pessoa no mapa de

projeção, garantindo uma altura mínima de pixels e consequentemente melhor qualidade da classificação final.

- Estudos podem ainda ser feitos para analisar se o uso de filtros, como por exemplo o de Sobel, na imagem gerada por subtrações de quadros consecutivos tem um desempenho melhor do que usar as imagens originais.
- Treinar um novo classificador, aglutinando as características HOG e FO como entradas e avaliar se há algum ganho em relação à classificação em cascata FO→HOG.
- Apesar de HOG ser um padrão bem consolidado, é necessário estudar o uso de outras formas de detecção baseadas na silhueta das pessoas, independentemente de estarem na posição ereta ou curvada. Teoricamente qualquer classificador que tenha o mesmo desempenho do classificador HOG até a faixa de 60% de acerto dos verdadeiro-positivos poderia substituí-lo sem problemas, mas é uma teoria a ser confirmada por futuras pesquisas.
- Implementar os algoritmos na linguagem C para melhorar o desempenho temporal, para que possa processar imagens a mais de 16 fps.
- Viabilizar a implantação do sistema de forma embarcada em hardware específico, seja em dispositivos DVR/NVR ou nas próprias câmeras.
- Testar a estratégia de detecção de eventos conjuntamente com a classificação FO→HOG em outras bases de dados, expandindo o aprendizado do classificador baseado em FO.
- Expandir os testes para outros ambientes reais para verificar a possibilidade de se ajustar o sistema apenas variando parâmetros, sem retreinar os classificadores para cada caso específico.

REFERÊNCIAS

BAY, H., ESS, A., TUYTELAARS, T., VAN GOOL, L. Speeded Up Robust Features (SURF). **Computer Vision and Image Understanding**, volume 110, n. 3, pp. 346-359, 2008.

BENENSON, R., MATHIAS, M., TIMOFTE, R., VAN GOOL, L. Pedestrian Detection at 100 Frames per Second. **Computer Vision and Pattern Recognition, IEEE Conference on**, Providence, pp. 2903-2910, 2012.

BORGES, P. V. K. Pedestrian Detection Based on Blob Motion Statistics. **IEEE Transactions on Circuits and Systems for Video Technology**, volume 23, n. 2, pp. 224-235, 2013.

CHENG, H., ZENG, Y., LEE, C., HSU, S. Segmentation of Pedestrians with Confidence Level Computation. **Journal of Signal Processing Systems**, volume 72, pp. 87–97, 2013.

CONTE, D., FOGGIA, P., PERCANNELLA, G., TUFANO, F., VENTO, M. A Method for Counting Moving People in Video Surveillance Videos. **EURASIP Journal on Advances in Signal Processing**, volume 2010, 10 páginas, 2010.

DALAL, N., TRIGGS, B. Histogram of Oriented Gradients for Human Detection. **Computer Vision and Pattern Recognition, IEEE Computer Society Conference**, San Diego, volume 1, pp. 886-893, 2005.

DALAL, N., TRIGGS, B., SCHMID, C. Human Detection Using Oriented Histograms of Flow and Appearance. **Proceedings of the European Conference on Computer Vision**, Graz, volume 3952, pp.428-441, 2006.

DENMAN, S., CHANDRAN, V., SRIDHARAN, S. An Adaptative Optical Flow Technique for Person Tracking Systems. **Pattern Recognition Letters**, volume 28, issue 10, pp. 1232-1239, 2007.

DOLLÁR, P., WOJEK, C., SCHIELE, B., PERONA, P. Pedestrian Detection: and Evaluation of the State of the Art. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, volume 34, n. 4, pp. 743-761, 2012.

DUARTE, G. Uso da Transformada de Hough na Detecção de Círculos em Imagens Digitais. **Thema Revista Científica do Centro Federal de Educação Tecnológica de Pelotas**, volume 4, n. 1, pp. 51-58, 2003.

DUDA, R. O., HART, P. E. Use of the Hough Transformation to Detect Lines and Curves in Pictures. **Communications of the ACM**, volume 15, n. 1, pp. 11-15, 1972.

DUDA, R. O., HART, P. E., STORK, D. G. **Pattern Classification.** 2. ed. New York: Wiley-Interscience, 2001.

EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C.K.I., WINN, J., ZISSERMAN, A. The PASCAL Visual Object Classes (VOC) Challenge. **International Journal of Computer Vision**, volume 88, n. 2, pp. 303-338, 2010.

FARNEBACK, G. Fast and Accurate Motion Estimation Using Orientation Tensors and Parametric Motion Models. **Proceedings of the 15th International Conference on Pattern Recognition**, volume 1, pp. 135-139, 2000.

FAWCETT, T. An Introdution to ROC Analysis. **Pattern Recognition Letters**, volume 27, issue 8, pp. 861-874, 2006.

FREUND, Y., SCHAPIRE, R. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences**, volume 55, issue 1, pp. 119-139, 1997.

GARCIA, J., GARDEL, A., BRAVO, I., LÁZARO, J., MARTÍNEZ, M., RODRÍGUEZ, D. Directional People Counter Based on Head Tracking. **IEEE Transactions on Industrial Electronics**, volume 60, n. 9, pp. 3991-4000, 2013.

GODOY, J. E. **Técnicas de Segurança em Condomínios.** 4. ed. rev. atual. São Paulo: Senac, 2014.

HARITAOGLU, I., HARWOOD, D., DAVIS, L. S. W4S: A real-time system detecting and tracking people in 2 ½D. **Proceedings of the 5th European Conference on Computer Vision**, volume 1, pp. 877-892, 1998.

HOUGH, P.V.C. **Method and Means for Recognizing Complex Patterns**, US Patent n. 3069654, 1962.

JUAN, L., GWUN, O. A Comparison of SIFT, PCA-SIFT and SURF. **International Journal of Image Processing**, volume 3, issue 4, pp. 143-152, 2009.

KALMAN, R. E. A New Approach to Linear Filtering and Prediction Problems. **Jornal Basic Engineering**, volume 82, issue 1, pp. 35-45, 1960.

LIU, Y., CHEN, X., YAO, H., CUI, X., LIU, C., GAO, W. Contour-motion feature (CMF): A space–time approach for robust pedestrian detection. **Pattern Recognition Letters**, volume 30, pp. 148–156, 2009.

LIU, X., TU, P., RITTSCHER, J., PERERA, A., KRAHNSTOEVER, N. Detecting and Counting People in Surveillance Applications. **Advanced Video and Signal Based Surveillance, IEEE Conference on**, pp. 306-311, 2005.

LU, W., JINGLU, T. Detection of Incomplete Ellipse in Images with Strong Noise by Iterative Randomized Hough Transform (IHRT). **Pattern Recognition**, volume 41, issue 4, pp. 1268-1279, 2008.

LUCAS, B. D., KANADE, T. An Iterative Image Registration Technique with an Application to Stereo Vision. **Proceedings of the 7th International Joint Conference on Artificial Intelligence**, Vancouver, pp. 674-679, 1981.

MARTIN, A., MARTINEZ, J. On Collaborative People Detection and Tracking in Complex Scenarios. **Image and Vision Computing**, volume 30, pp. 345-354, 2012.

MIAO, Q., WANG, G., SHI, C., LIN, X., RUAN, Z. A New Framework for On-Line Object Tracking Based on SURF. **Pattern Recognition Letters**, volume 32, issue 13, pp. 1564-1571, 2011.

MITCHELL, T. M. Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression. In: _____. **Machine Learning**. 2. ed. New York: McGraw Hill, 2015. Disponível em: <http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf>. Acesso em: 2 out. 2015.

NICOLL, D., LU, C. M., PIGNONE, M., MCPHEE, S. J. **Manual de Exames Diagnósticos.** 6. ed. Porto Alegre: AMGH, 2014.

NGHIEM, A.T., BREMOND, F., THONNAT, M., VALENTIN, V. ETISEO, performance evaluation for video surveillance systems. **Advanced Video and Signal Based Surveillance, AVSS IEEE Conference on**, pp. 476-481, 2007.

OSOWSKI, S., SIWEK, K., MARKIEWICZ, T. MLP and SVM Networks: A Comparative Study. **Proceedings of the 6th Nordic Signal Processing Symposium NORSIG**, Finland, 2004.

PAULUS, D. W. R., HORNEGGER, J. **Pattern Recognition and Image Processing in C++.** New York: Springer Verlag, 1995.

PANG, Y., YUAN, Y., LI, X., PAN, J. Efficient HOG Human Detection. **Signal Processing**, volume 91, issue 4, 2011.

PEDRINI, H., SCHWARTZ, W. **Análise de Imagens Digitais:** Princípios, Algoritmos e Aplicações. São Paulo: Thomson Learning, 2008.

REISSWITZ, F. **Análise de Sistemas V8:** Probabilidade e Estatística. São Paulo: Clube de Autores, 2009.

SAND, P., TELLER, S. Particle Video: Long-Range Motion Estimation Using Point Trajectories. **International Journal of Computer Vision**, volume 80, issue 1, pp. 72-91, 2008.

SCHWARTZ, W. R. Human Detection Based on Large Feature Sets Using Graphics Processing Units. **Informatica – An International Journal of Computing and Informatics**, Brazil, volume 35, n. 4, pp. 473-479, 2011.

SIM, C., RAJMADHAN, E., RANGANATH, S. Detecting People in Dense Crowds. **Machine Vision and Applications**, volume 23, issue 2, pp. 243-253, 2012.

SUBBURAMAN, V., DESCAMP, A., CARINCOTTE, C. Counting People in the Crowd Using a Generic Head Detector. **Advanced Video and Signal-Based Surveillance, IEEE Nineth Conference on**, pp. 470-475, 2012.

THIEL, G. Automatic CCTV Surveillance – Towards the VIRTUAL GUARD. **Aerospace and Electronic Systems Magazine, IEEE**, volume 15, issue 7, pp. 3-9, DOI 10.1109/62.854018, 2000.

TOMPKIN, J. **Optical Flow: An Introduction,** 2008. Disponível em: <http://www.cs.ucl.ac.uk/staff/a.moore/mvpractical2.pdf>. Acesso em: 17 nov. 2013.

VISHWAKARMA, S., AGRAWAL, A. A Survey on Activity Recognition and Behavior Understanding in Video Surveillance. **The Visual Computer**, volume 29, issue 10, pp. 983-1009, 2013.

VIOLA, P., JONES, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. **IEEE Conference on Computer Vision and Pattern Recognition**, volume 1, pp. 511-518, 2001.

VIOLA, P., JONES, M. J. Robust Real-Time Face Detection. **International Journal of Computer Vision**, volume 37, issue 2, pp. 137-154, 2004.

VIOLA, P., JONES, M. J., SNOW, D. Detecting Pedestrians Using Patterns of Motion and Appearance. **International Journal of Computer Vision**, volume 63, n. 2, pp. 153-161, 2005.

WALLACE, E., DIFFLEY, C. **CCTV: Making It Work. CCTV Control Room Ergonomics**, In: Police Scientific Development Branch of the Home Office, Publication n. 14/98, 1998. Disponível em: <http://puls-global.com/Downloads/CCTV-Control-Room-Ergonomics.pdf>. Acesso em: 20 dez. 2013.

WEI, Y., TIAN, Q., GUO, T. An Improved Pedestrian Detection Algorithm Integrating Haar-Like Features and HOG Descriptors. **Advances in Mechanical Engineering**, volume 2013, 8 páginas, 2013.

APÊNDICE A – Tabelas

A.1 Tabelas contendo breve descrição dos eventos

Tabela 14 – Descrição breve dos eventos usados para treinamento

Evento	Arquivo	Quadros	Tempo (s)	Atores	Descrição
01	C01	35	2	1 carro	Um carro entrando pelo portão de um estacionamento.
02	C02	13	1	1 carro	Um carro entrando pelo portão de um estacionamento.
03	C03	13	1	1 carro	Um carro entrando pelo portão de um estacionamento.
04	C04	27	2	1 carro	Um carro entrando pelo portão de um estacionamento.
05	C05	11	1	1 carro	Um carro entrando pelo portão de um estacionamento.
06	C06	5	0,5	2 adultos	Dois homens próximos de um portão.
07	C07	10	1	1 carro	Um carro entrando pelo portão de um estacionamento.
08	C07	9	1	1 adulto	Um homem passando pela calçada.
09	C09	13	1	1 carro	Um carro entrando pelo portão de um estacionamento.
10	C09	17	1	1 carro	Um carro entrando pelo portão de um estacionamento.
11	C10	11	1	1 moto	Uma moto entrando pelo portão de um estacionamento.
12	C11	12	1	1 carro	Um carro entrando pelo portão de um estacionamento.
13	C12	11	1	1 carro	Um carro entrando pelo portão de um estacionamento.
14	C13	6	0,5	1 carro	Um carro entrando pelo portão de um estacionamento.
15	C15	39	2	2 carros	Um carro saindo pelo portão de um estacionamento, seguido de outro carro.
16	C15	33	2	2 adultos	Um homem passando pela calçada e uma mulher saindo pelo portão do estacionamento
17	C16	11	1	1 adulto	Um homem saindo pelo portão de um estacionamento.

Tabela 14 – Continuação

Evento	Arquivo	Quadros	Tempo (s)	Atores	Descrição
18	C17	32	2	1 carro	Um carro saindo pelo portão de um estacionamento.
19	C17	6	0,5	1 carro	Um carro saindo pelo portão de um estacionamento.
20	C18	28	2	1 carro	Um carro saindo pelo portão de um estacionamento.
21	C19	26	2	1 carro	Um carro entrando pelo portão de um estacionamento.
22	C20	13	0,5	1 carro	Um carro saindo pelo portão de um estacionamento.
23	C34	48	3	1 carro	Um carro entrando pelo portão de um estacionamento.
24	C38	45	3	1 carro	Um carro entrando pelo portão de um estacionamento.
25	C39	60	4	1 carro	Um carro saindo pelo portão de um estacionamento.
26	C41	36	2	1 carro	Um carro entrando pelo portão de um estacionamento.
27	P01	18	1	1 adulto	Uma mulher numa bicicleta entrando pelo portão de um estacionamento.
28	P01	43	3	1 adulto	Um homem saindo pelo portão de um estacionamento.
29	P01	16	1	2 adultos	Um homem andando e outro passando pela calçada.
30	P01	7	0,5	1 adulto	Um homem andando pelo pátio
31	P01	34	2	1 adulto	Um homem saindo pelo portão de um estacionamento.
32	P01	26	2	2 adultos	Um homem entrando pelo portão de um estacionamento, seguido de outro homem.
33	P02	35	2	2 adultos	Duas mulheres entrando pelo portão de um estacionamento.
34	P03	53	3	2 adultos	Um casal saindo pelo portão de um estacionamento.
35	P04	60	4	2 adultos	Dois homens saindo pelo portão de um estacionamento.

Tabela 14 – Continuação

Evento	Arquivo	Quadros	Tempo (s)	Atores	Descrição
36	P05	53	3	6 adultos	Um homem e duas mulheres entrando pelo portão de um estacionamento, enquanto dois homens e uma mulher estão próximos à calçada.
37	P05	23	1	2 adultos	Um homem andando e uma mulher passando pela calçada.
38	P06	23	1	2 adultos 1 criança	Duas mulheres entrando pelo portão de um estacionamento com uma criança.
39	P07	12	1	1 adulto	Um homem caminhando.
40	P08	8	0,5	1 adulto	Uma mulher numa bicicleta entrando pelo portão de um estacionamento.
41	P09	31	2	3 adultos	Um homem e duas mulheres entrando pelo portão de um estacionamento.
42	P10	36	2	2 adultos 2 crianças	Um casal com 2 crianças saindo pelo portão de
43	P10	42	3	1 adulto	Um homem saindo pelo portão de um estacionamento.

Tabela 15 – Descrição breve dos eventos usados para testes (controle)

Evento	Arquivo	Quadros	Tempo (s)	Atores	Descrição
01	C01	37	2	1 carro	Um carro entrando pelo portão de um estacionamento.
02	C21	55	3	1 carro	Um carro entrando pelo portão de um estacionamento.
03	C22	70	4	1 carro	Um carro entrando pelo portão de um estacionamento.
04	C32	42	3	1 carro	Um carro entrando pelo portão de um estacionamento.
05	C33	27	2	1 carro	Um carro entrando pelo portão de um estacionamento.
06	C35	33	2	1 carro	Um carro entrando pelo portão de um estacionamento.
07	C36	62	4	1 carro	Um carro saindo pelo portão de um estacionamento.
08	C37	44	3	1 carro	Um carro saindo pelo portão de um estacionamento.
09	C40	15	1	1 carro	Um carro entrando pelo portão de um estacionamento.
10	C42	27	2	1 carro	Um carro entrando pelo portão de um estacionamento.
11	P11	38	2	2 adultos 3 crianças	Um casal e três crianças entrando pelo portão de um estacionamento.
12	P12	74	5	2 adultos 1 criança	Uma mulher e uma criança entrando pelo portão de um estacionamento. Em seguida entra outra mulher.
13	P13	23	1	1 adulto	Um homem entrando por um portão de um estacionamento.
14	P14	34	2	2 adultos	Um homem entrando por um portão de um estacionamento, seguido de uma mulher.
15	P15	89	6	4 adultos	Um homem entrando por um portão, seguido de dois outros homens e uma mulher.
16	P16	56	4	2 adultos	Um homem entrando e uma mulher saindo pelo portão de um estacionamento.

Tabela 15 – Continuação

17	P17	100	6	2 adultos 1 criança	Uma criança e uma mulher saindo pelo portão, seguidos de outra mulher. A última retorna.
18	P18	139	9	7 adultos 1 criança	Um casal com uma criança saindo pelo portão. Um homem entra, seguido de duas mulheres. Outro homem passa na calçada.
19	P19	65	4	2 adultos 1 criança	Duas mulheres e uma criança saindo pelo portão

Tabela 16 – Descrição breve dos eventos usados para testes finais

Evento	Arquivo	Quadros	Tempo (s)	Atores	Descrição
01	O07	75	5	1 carro	Um carro entrando pelo portão.
02	O07	50	3	1 adulto	Uma mulher saindo do carro e caminhando na garagem
03	O07	213	13	3 adultos	Dois ladrões entrando por um portão, roubando a vítima e saindo correndo.
04	O09	52	3	1 adulto	Homem entrando por um portão
05	O09	59	4	1 adulto	Um homem saindo de um carro e entrando por um portão.
06	O09	127	8	2 adultos	Um homem saindo por um portão, seguido de outro homem e entram num carro.
07	O09	119	7	1 adulto	Um homem entrando por um portão e andando pelo pátio
08	O09	50	3	1 adulto	Um homem saindo de um carro e entrando por um portão.
09	O09	30	2	1 adulto	Um homem saindo por um portão e entrando num carro.
10	O09	39	2	1 adulto	Um homem saindo de um carro e entrando por um portão.
11	O09	147	9	1 adulto	Um homem saindo por um portão, entrando num carro e partindo.
12	O09	25	2	1 adulto	Um homem entrando por um portão e andando pelo pátio
13	O09	25	2	1 adulto	Um homem saindo por um portão.
14	O09	33	2	1 adulto	Um homem saindo por um portão.
15	O11	53	3	1 carro	Um carro entrando pelo portão.
16	O11	135	8	1 carro 1 adulto	Um ladrão entrando pelo portão, tentando assaltar o motorista. O motorista dá marcha ré.
17	O11	45	3	1 carro 1 adulto	Um motorista dando marcha ré e um ladrão tentando quebrar o vidro.

Tabela 16 – Continuação

Evento	Arquivo	Quadros	Tempo (s)	Atores	Descrição
18	O15	41	3	1 adulto	Um homem caminhando perto de um carro.
19	O15	129	8	1 carro	Um carro saindo pelo portão.
20	O16	33	2	1 adulto	Um homem caminhando perto de um carro.
21	O16	42	3	1 moto 1 adulto	Um homem de moto saindo por um portão.
22	O17	305	19	2 adultos	Um homem perto de um portão. Outro homem se aproxima.
23	O18	183	11	1 carro 3 adultos	Um carro saindo por um portão. A seguir dois homens e uma mulher saem.
24	O19	435	27	1 carro 3 adultos	Um carro entrando por um portão e um homem entra atrás. A seguir um homem e uma mulher entram.
25	O20	150	9	1 carro 1 adulto	Um carro entrando por um portão e um homem entra atrás.
26	O21	584	36	1 carro 3 adultos	Um carro saindo por um portão. Em seguida dois homens e uma mulher saem.
27	O22	466	29	1 carro 3 adultos	Um carro entrando por um portão e um homem entra atrás. Um homem também entra e posteriormente uma mulher.
28	O23	128	8	2 adultos	Dois homens entrando por um portão.
29	C02	13	1	2 carros	Dois carros entrando por um portão de um estacionamento.
30	C03	13	1	1 carro	Um carro entrando por um portão de um estacionamento.
31	C04	43	3	1 carro	Um carro entrando por um portão de um estacionamento.
32	C04	62	4	1 carro	Um carro entrando por um portão de um estacionamento.
33	C05	11	1	1 carro	Um carro entrando por um portão de um estacionamento.
34	C07	22	1	1 carro	Um carro entrando por um portão de um estacionamento.

Tabela 16 – Continuação

Evento	Arquivo	Quadros	Tempo (s)	Atores	Descrição
35	C09	165	10	1 carro	Um carro entrando e manobrando no estacionamento
36	C10	12	1	1 moto	Uma moto entrando por um portão de um estacionamento.
37	C11	12	1	1 carro	Um carro entrando por um portão de um estacionamento.
38	C12	11	1	1 carro	Um carro entrando por um portão de um estacionamento.
39	C13	6	0,5	1 carro	Um carro entrando por um portão de um estacionamento.
40	C15	23	1	1 carro	Um carro saindo pelo portão de um estacionamento.
41	C15	19	1	1 carro	Um carro saindo pelo portão de um estacionamento.
42	C17	113	7	1 carro	Um carro saindo pelo portão de um estacionamento.
43	C18	28	2	1 carro	Um carro saindo pelo portão de um estacionamento.
44	C19	90	6	1 carro	Um carro entrando por um portão de um estacionamento.
45	C20	13	1	1 carro	Um carro saindo pelo portão de um estacionamento.
46	C30	41	3	1 carro	Um carro entrando por um portão de um estacionamento.
47	C31	35	2	1 carro	Um carro entrando por um portão de um estacionamento.
48	C34	51	3	1 carro	Um carro entrando por um portão de um estacionamento.
49	C38	48	3	1 carro	Um carro entrando pelo portão de um estacionamento.
50	C39	68	4	1 carro	Um carro saindo por um portão de um estacionamento.
51	C41	41	3	1 carro	Um carro entrando por um portão de um estacionamento.
52	C43	31	2	1 carro	Um carro saindo pelo portão de um estacionamento.
53	C44	107	7	1 carro	Um carro saindo pelo portão de um estacionamento.
54	C45	61	4	1 carro	Um carro entrando por um portão de um estacionamento.

A.2 Tabelas das matrizes de confusão dos classificadores

Tabela 17 – Matriz Confusão Classificação por FO

Classes \ OffsetFo		-1,5	-0,75	-0,5	-0,25	0,0	0,25	0,5	1,0	2,5
Pessoas	VP	1335	1259	1224	1167	1120	1056	973	797	318
	FP	665	381	294	227	190	143	103	71	17
Carros	VP	1071	1355	1442	1509	1546	1593	1633	1665	1719
	FP	615	691	726	783	830	894	977	1153	1632
tempo (s)		309								

Tabela 18 – Matriz Confusão Classificação por HOG

Classes \ OffsetHog		-1,5	-0,75	-0,5	-0,25	0,0	0,25	0,5	1,0	2,5
Pessoas	VP	2081	1836	1613	1329	1012	771	512	215	3
	FP	1237	618	432	262	137	78	32	2	0
Carros	VP	499	1118	1304	1474	1599	1658	1704	1734	1736
	FP	10	255	478	762	0	1320	1579	1876	2088
tempo (s)		948								

Tabela 19 – Matriz Confusão Classificação por HOG→FO (OffsetFo=0,25)

OffSetHog / Classes		-1,5	-0,75	-0,5	-0,25	0,0	0,25	0,5	1,0	2,5
Pessoas	VP	1207	1064	940	779	606	461	298	119	2
	FP	234	129	84	49	15	7	4	0	0
Carros	VP	1502	1607	1652	1687	1721	1729	1732	1736	1736
	FP	884	1027	1151	1312	1485	1630	1793	1972	2089
tempo (s)		989								

Tabela 20 – Matriz Confusão Classificação por FO→HOG (OffsetFo=0,25)

OffSetHof / Classes		-1,5	-0,75	-0,5	-0,25	0,0	0,25	0,5	1,0	2,5
Pessoas	VP	1207	1064	940	779	606	461	298	119	2
	FP	234	129	84	49	15	7	4	0	0
Carros	VP	1502	1607	1652	1687	1721	1729	1732	1736	1736
	FP	884	1027	1151	1312	1485	1630	1793	1972	2089
tempo (s)		1031								

A.3 Tabela com resultados das simulações de geração de alarmes

Tabela 21 – Detecção do Evento Intrusão

Classif.	HOG 0,00				HOG FO 0,00				HOG -0,15 FO +0,25				HOG +0,31				HOG +0,50				HOG +0,59			
Evento	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP
01		X				X				X				X				X				X		
02	X				X				X				X				X				X			
03	X				X				X				X				X				X			
04	X				X				X				X						X				X	
05	X				X				X						X				X				X	
06	X				X				X				X				X				X			
07	X				X				X				X				X				X			
08	X				X				X						X				X				X	
09	X				X				X						X				X				X	
10	X				X				X						X				X				X	
11	X				X				X				X						X				X	
12			X				X		X						X				X				X	
13	X				X				X						X				X				X	
14	X				X				X						X				X				X	
15		X				X				X				X				X				X		
16			X				X		X						X				X				X	
17			X				X			X					X				X				X	
18	X				X				X				X						X				X	
19				X		X				X				X				X				X		
20			X				X				X				X				X				X	
21		X				X				X				X				X				X		
22	X				X				X				X				X				X			
23	X				X				X				X				X				X			
24a	X				X				X				X				X				X			
24b	X				X				X				X						X				X	
25			X				X				X				X				X				X	
26	X			X	X				X				X				X				X			
27	X				X				X				X				X				X			
28	X				X				X				X				X				X			
29		X				X				X				X				X				X		
30		X				X				X				X				X				X		
31				X		X				X						X		X				X		
32				X				X				X				X		X				X		
33				X		X				X						X				X				X
34		X				X				X				X				X				X		
35				X		X				X						X			X			X		

Tabela 21 – Continuação

Classif.	HOG 0,00				HOG FO 0,00 0,00				HOG -0,15 FO +0,25				HOG +0,31				HOG +0,50				HOG +0,59			
Evento	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP
36		X				X				X				X				X				X		
37				X		X				X						X		X				X		
38		X				X				X				X				X				X		
39		X				X				X				X				X				X		
40				X		X				X				X				X				X		
41		X				X				X				X				X				X		
42		X				X				X				X				X				X		
43				X		X				X				X				X				X		
44				X		X				X						X		X				X		
45				X		X				X						X				X		X		
46				X		X				X				X				X				X		
47				X		X				X						X				X				X
48		X				X				X				X				X				X		
49				X		X				X						X				X				X
50				X		X				X						X				X		X		
51		X				X				X				X				X				X		
52		X				X				X				X				X				X		
53		X				X				X				X				X				X		
Totais	20	15	5	15	20	28	5	1	**22**	**29**	**2**	**1**	14	19	11	10	10	23	16	5	10	26	15	3
	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP	VP	VN	FN	FP
Classif.	HOG 0,00				HOG FO 0,00 0,00				**HOG -0,15 FO +0,25**				HOG +0,31				HOG +0,50				HOG +0,59			

A.4 Tabelas do desempenho da arquitetura FOGI ao se ativar ou desativar o classificador FO

Tabela 22 – Arquitetura FOGI com classificador FO ativado

Arquitetura FOGI (Classificadoresa FO ativado e HOG ativado)					
gatilhoP (Presença)	gatilhoA (Ausência)	VP	VN	FN	FP
1	0	22	14	2	20
2	0	22	22	2	12
3	5	22	26	2	5
3	3	**22**	**27**	**2**	**4**
3	1	22	27	2	4
4	3	21	27	3	2
3	0	21	24	3	4
5	3	17	28	7	2
5	1	17	28	7	2
5	0	16	30	8	1

Tabela 23 – Arquitetura FOGI com classificador FO desativado

Arquitetura FOGI (Classificadoresa FO desativado e HOG ativado)					
gatilhoP (Presença)	gatilhoA (Ausência)	VP	VN	FN	FP
2	0	21	15	3	17
3	5	21	16	3	16
3	3	21	17	3	15
3	1	21	18	3	14
4	3	20	20	4	12
3	0	20	17	4	15