Metodologia Híbrida para Detecção de Pessoas em Vídeos

Hugo Alberto Perlin
Instituto Federal do Paraná - Campus Paranaguá
hugo.perlin@ifpr.edu.br

Heitor Silvério Lopes
CPGEI - Universidade Tecnológica Federal do Paraná (UTFPR)
hslopes@pesquisador.cnpq.br

Resumo

The human detection is a key step in the design of an automatic video analysis, which is necessary both accuracy and processing speed. This paper aims to propose a hybrid method of particle swarm optimization (PSO) and support vector machines (SVM), using as features the histograms of oriented gradients (HOG) reduced by principal component analysis (PCA). Tests show a detection rate about 80% at 35 frames per second (FPS), showing that the method is efficient both in terms of computational and quality, could be applied to real-time problems.

1. Introdução

A análise de vídeos realizada por humanos, para segurança por exemplo, é uma tarefa ineficiente, visto que a quantidade de dados a serem analisados geralmente é bem maior do que o número de pessoas disponíveis. Além disso, na maioria das vezes os vídeos são utilizados como um tipo de evidência em algum caso, pois não se consegue fazer a análise em tempo real dos vídeos para que se possa tomar uma decisão no momento em que os fatos ocorrem.

O desenvolvimento de sistemas computacionais que auxiliem na análise dos vídeos, bem como na tomada de decisão, possibilita um uso mais eficiente dos dados, pois é possível analisar uma quantidade muito maior de dados com um número reduzido de pessoas .

Um passo fundamental para o emprego destes sistemas é a detecção de pessoas dentro de um quadro do vídeo. A detecção consiste em determinar a localização de uma ou mais pessoas, informando sua posição relativa no quadro. Como os seres humanos possuem características que são comum a todos, é possível estabelecer algumas características e utilizá-las como forma de diferenciar uma pessoa de qualquer outro objeto dentro de um quadro do vídeo [4] [9] .

A detecção de pessoas é um assunto bastante discutido na área de visão computacional, sendo que diversas proposições de solução podem ser encontradas. Em [5], são utilizados dois conjuntos de características, os Histogramas de Gradiente Orientados (HOG) e também matriz de covariância juntamente com redes neurais. Em [8] e [11], o HOG é combinado com Padrões Binários Locais (LBP) para a representação de pessoas, sendo classificados através de uma Máquina de Vetores de Suporte (SVM). Em [10], o HOG é combinado com uma versão incremental da Análise de Componente Principal (PCA) os quais são utilizados em um filtro de partículas para criar um *framework* de rastreamento. Em [1], o HOG e o SVM são utilizados em conjunto com a otimização por Enxame de Partículas (PSO), para realizar a detecção de pessoas em imagens estáticas.

Inspirado por estes trabalhos, o objetivo aqui é propor um sistema de reconhecimento de padrões (SRP) que seja capaz de realizar a detecção de uma pessoa em um vídeo.

Este artigo está dividido em 5 seções que se seguem. Na seção 2 é feita uma revisão sobre trabalhos correlatos. Na seção 3, são explicadas as técnicas que foram utilizadas para a construção do sistema. O processo para a realização dos testes e validação dos resultados é mostrado na seção 4. Por fim, as conclusões são discutidas na seção 5.

2. Trabalhos Correlatos

Um modelo básico de SRP pode consistir nos módulos de aquisição de dados; pré-processamento, redução de dimensionalidade e predição [3]. Diversas técnicas estão presentes na literatura e podem ser empregadas para cada um destes módulos quando se trata de detecção de pessoas em vídeos.

2.1. Histogramas de Gradientes Orientados

A fase de pré-processamento consiste na preparação dos dados para a posterior predição. É nesta fase que são extraídos um conjunto de características que permita definir o padrão desejado.

Proposto por [2], a idéia do HOG está na possibilidade de representar um objeto através dos gradientes locais de intensidade. Isso é feito dividindo-se a imagem em subconjuntos denominados células, para as quais são computados histogramas da direção do gradiente. De posse dos histogramas é feita uma normalização com base nos valores dos histogramas das células adjacentes, chamados blocos, buscando invariância a luminosidade.

O HOG é computado em janelas de 64x128 pixels, as quais são divididas em células de 8x8 pixels agregadas em blocos de 2x2 células. Para melhorar os resultados, os blocos são sobrepostos, gerando desta forma um conjunto de 7x15 blocos para cada janela. O histograma de uma célula é composto por 9 partições, variando a direção do gradiente de 0^o a 180^o. Assim, o vetor HOG é composto por 3780 componentes (4 células x 9 posições do histograma x 105 blocos) [2] [1].

2.2. Otimização por Enxame de Partículas

Na fase de predição, uma heurística de otimização pode ser empregada para encontrar uma solução de forma eficiente, buscando manter a qualidade da solução. Dentre os métodos de otimização, destacam-se os bioinspirados, que utilizam conceitos observados na natureza para compor um algoritmo de otimização. Nesta classe de métodos se encontra o PSO que é inspirado no comportamento de cardumes de peixes e bandos de pássaros.

O PSO é um algoritmo populacional cooperativo, onde um conjunto de agentes, chamados de partículas, trocam informação entre si buscando encontrar a melhor solução dentro de um espaço de busca. Cada uma das partículas do PSO representa uma possível solução para o problema. Como a idéia é de otimização, deve ser determinada uma forma de avaliar a qualidade de uma solução, isto é feito pela definição de uma função de *fitness*.

Em um processo iterativo, as soluções do PSO são avaliadas e atualizadas baseadas na qualidade da solução da partícula e do melhor do grupo. A atualização de uma partícula é feita de acordo com as equações 1 e 2. O processo se repete até que um determinado limiar de qualidade de solução seja encontrado ou um número máximo de iterações atingido, como mostrado na figura 1 [6].

$$\begin{aligned} Vel_i^{t+1} = w * Vel_i^t + \\ c_1 * r_1 * (melhorP_i^t - x_i^t) + \\ c_2 * r_2 * (melhorG_i^t - x_i^t) \end{aligned} \quad (1)$$

$$x_i^{t+1} = x_i^t + Vel_i^{t+1} \quad (2)$$

onde: w é o momento de inércia, c_1 e c_2 são constantes de aceleração definidas pelo usuário, r_1 e r_2 são números aleatórios distribuídos aleatóriamente, $melhorP$ é a melhor solução encontrada até o momento pela partícula, $melhorG$ é a melhor solução da população até o momento, e x_i representa a solução corrente.

3. Metodologia

Este trabalho busca hibridizar algumas técnicas de forma a desenvolver um método que permita alcançar melhores resultados tanto na qualidade de detecção de pessoas bem como a redução do tempo de processamento.

Como conjunto de características para representar uma pessoa foi empregado o HOG, visto seu uso difundido na literatura dada a sua robutez quando aplicado ao problema.

Analisando o tamanho do vetor de características gerado por meio do HOG, nota-se que o mesmo quando comparado com a dimensão de uma imagem normal, é pequeno. Entretanto, seria interessante uma maior redução eliminando possíveis componentes desnecessárias, visto que o tamanho do vetores de característica tem influência no esforço computacional para a classificação. Vale lembrar, que tal redução idealmente não deve afetar a qualidade da classificação. Para isto foi empregada a PCA, a qual é vastamente utilizada na literatura em problemas que necessitam redução de dimensionalidade, mostrando uma boa qualidade no mapeamento.

Dentro de uma imagem, uma pessoa pode aparecer em qualquer posição bem como em qualquer nível de escala. Assim, uma pessoa em uma imagem pode ser representada por uma tripla (X, Y, S), onde X e Y são as componentes do ponto central de uma janela de detecção e S é a componente de escala. Ou seja, esta tripla representa uma solução para o problema de detecção de pessoas dentro de uma imagem.

A combinação de todos os valores desta tripla, dependendo do tamanho da imagem bem como do intervalo de escala, pode gerar um espaço de busca bastante grande. Desta forma, é necessário realizar algum tipo de busca na imagem para determinar os valores das componentes que indiquem a presença de uma pessoa.

Esta busca pode ser vista como um processo de otimização, onde para cada solução associa-se um valor de qualidade de solução. Maximizando a qualidade da solução será encontrada uma pessoa dentro da imagem. Neste trabalho a função de qualidade é determinada pela resposta contínua de um SVM treinado para reconhecer pessoas. Nos termos dos algoritmos bioinspirados esta é a função de *fitness*.

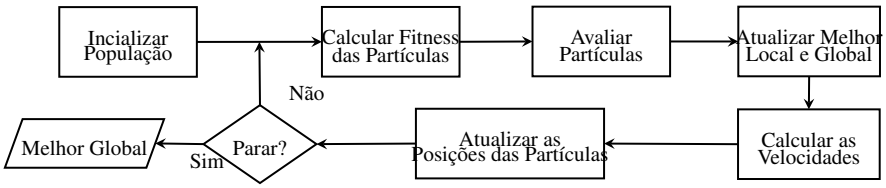

Figura 1. Fluxograma mostrando os passos de um algoritmo básico para o PSO.

Grande parte dos trabalhos utiliza o HOG juntamente com o SVM como detector de pessoas. Porém, o algoritmo utilizado é a força bruta, onde uma janela de detecção é deslizada sobre uma imagem alvo. O espaço de busca é o tamanho da imagem alvo multiplicado por um intervalo de escala. A força bruta garante o resultado ótimo, porém com um grande esforço computacional. Uma forma mais eficiente de realizar a busca é utilizando uma heurística, que aqui é o PSO.

O fluxograma do método proposto é apresentado na figura 2. Nota-se que a sequência básica das etapas é a mesma do PSO convencional. A diferença está na forma de cálculo do valor de *fitness* para cada uma das partículas. É neste momento que a hibridização entre HOG, SVM, PCA e PSO ocorre. A região dentro da linha tracejada, indica as etapas necessárias para o cálculo do *fitness*.

Para melhorar a qualidade de reconhecimento, uma forma de inicialização não aleatória das partículas foi empregada. Este tipo de inicialização é possível visto que a posição de uma pessoa durante uma sequência de quadros de um vídeo muda poucos pixels. Desta forma, dado um limiar definido de forma empírica, as partículas do PSO são inicializadas para a busca em um novo quadro em uma região muito próxima ao melhor resultado do quadro anterior, exceto no primeiro quadro, onde a inicialização é totalmente aleatória.

Como a função de *fitness* é a saída de um SVM, é necessário realizar o treinamento do mesmo. O treino segue o fluxograma mostrado na figura 3. Após esta fase, são gerados os vetores de suporte pelo SVM, os quais servirão como entrada para uma futura classificação e também os vetores de mapeamento do PCA, que permitirão a redução de dimensionalidade.

4. Testes e Resultados

Na implementação foi empregado o pacote de visão computacional OpenCV, em sua versão 2.3.1. A escolha deste pacote se deve ao seu amplo uso na comunidade científica, bem como na qualidade dos algoritmos implementados, que incluem o HOG, PCA e SVM.

A implementação da heurística PSO foi complementar ao pacote e implementada com o uso da linguagem de programação C++. Para aumentar a eficiência computacional, foi feito o uso de processamento paralelo, buscando usar a vantagem de ambientes computacionais com processadores de vários núcleos.

O ambiente computacional para a realização dos testes foi uma estação de trabalho HP Z210, processador Intel Xeon QuadCore 3,3 Ghz, 8 GB de RAM, GPU Nvidia Quadro 2000 e sistema operacional Ubuntu/Linux 2.6.35-32.

O conjunto de treinamento utilizado é a base de imagens INRIA [1], é um conjunto de imagens de pessoas, mostrados na figura 4, dividido em um conjunto de treinamento e de testes, juntamente com um conjunto de imagens que não possuem pessoas. Esta base é vastamente utilizado na literatura como benchmark de sistemas de detecção de pessoas. O conjunto de treinamento é composto por 2416 imagens positivas, ou seja, que contém pessoas. O conjunto negativo pode ser gerado aleatoriamente por meio das imagens negativas, ou seja, que não contém pessoas. Foram utilizadas 7392 imagens negativas durante o treinamento.

Figura 4. Exemplos do conjunto de treinamento utilizado.

O conjunto de parâmetros empregado foi o proposto pela literatura. No caso do PCA foram testadas diferentes possibilidades de mapeamentos, visto que isto tem grande influência tanto na qualidade da solução e principalmente no custo computacional. De forma empírica três mapeamentos

1 Disponível para download em http://pascal.inrialpes.fr/data/human/

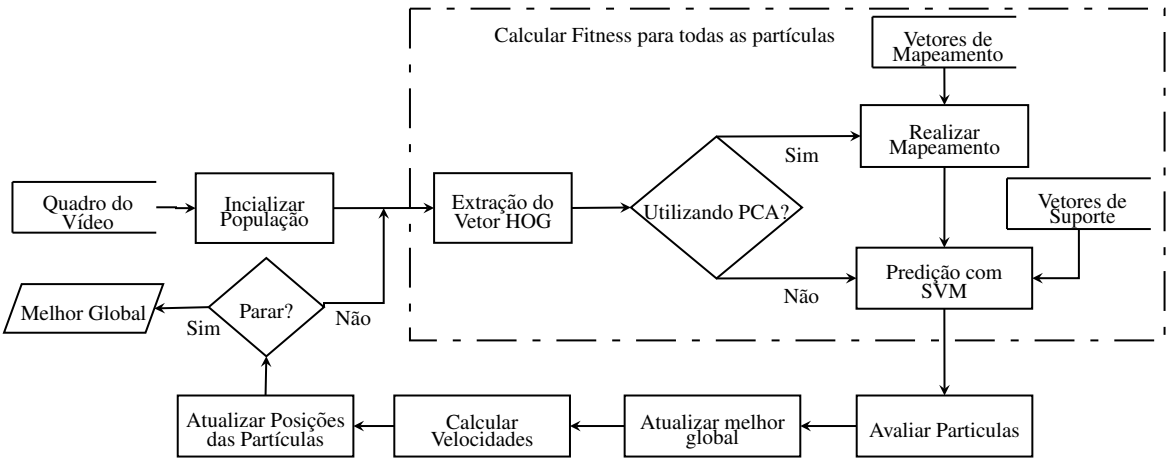

Figura 2. Fluxograma mostrando a sequência de passos do PSO para a busca de uma pessoa em um vídeo.

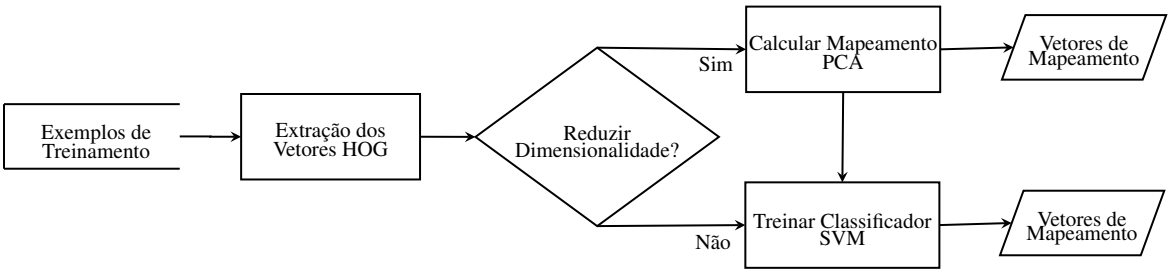

Figura 3. Fluxograma mostrando as etapas realizadas durante a etapa de treinamento do SRP proposto, onde as saídas geradas são os vetores de suporte e também os vetores de mapeamento.

foram escolhidos, de 3780 para 1890, 378 e 38 componentes, ou seja, 50% ,10% e 1% do tamanho do vetor original.

Para o PSO foi utilizada uma população de 10 partículas, cada uma com 3 dimensões, as quais representam X,Y e a escala. O critério de parada foi utilizado um máximo de 10 iterações. O intervalo de escala foi definido entre $0,5$ e $2,0$.

Para a realização dos testes foram empregados dois vídeos, disponibilizados em [7]. O primeiro vídeo [2], mostra uma pessoa em uma sala com uma câmera fixa. Foram utilizados apenas os 264 primeiros quadros do vídeo, os quais possuem 352x288 pixels. O segundo vídeo [3] é composto por 286 quadros de 320x240 pixels, onde na sua maioria aparece uma ou mais pessoas nos quadros em um ambiente externo sem nenhum tipo de controle. Uma característica adicional deste vídeo é a movimentação da câmera o que gera dificuldades adicionais para o sistema realizar o reconhecimento.

Como o treinamento e execução utilizam conjunto de imagens diferentes, pode-se entender que a validação do sistema se comporta como uma validação cruzada. A medida de qualidade para determinar o melhor modelo foi a quantidade de acertos. Por acerto é considerada a solução que esteja em um raio de +/- 10 pixels tanto para X quanto para Y, em relação a solução de referência. Como as informações de referência sobre a posição das pessoas em cada quadro do vídeo não satisfaziam as necessidades, a determinação foi feita manualmente.

Os resultados para o primeiro vídeo são mostrados na tabela 1 e para o segundo vídeo na tabela 2. Nas tabelas são mostrados, o tempo em segundos para a execução de uma rodada do método. a quantidade de quadros por segundo (FPS), o número de acertos e a porcetagem de acerto relativa a quantidade total de quadros do vídeo. Nas colunas estão os diferentes mapeamento do PCA que foram testa-

2 Disponível para download em (http://www.openvisor.org/ video_details.asp?idvideo=199)

3 Disponível para download em (http://www.openvisor.org/ video_details.asp?idvideo=322)

dos. Os valores apresentados são equivalentes à média de 30 rodadas independentes do método proposto.

	Componentes do PCA			
	38	378	1890	3780
Tempo (s)	7,60	20,27	82,07	17,17
FPS	34,74	13,03	3,22	15,38
Acertos	207,90	187,46	210,27	204,20
% Acerto	78,75%	71,01%	79,65%	77,35%

Tabela 1. Resultados médios de 30 rodadas independentes para o vídeo 1.

	Componentes do PCA			
	38	378	1890	3780
Tempo (s)	8,1	23,57	88,27	19,77
FPS	35,31	12,14	3,24	14,47
Acertos	230,53	217,23	220,37	215,53
% Acerto	80,61%	75,96%	77,05%	75,36%

Tabela 2. Resultados para de 30 rodadas independentes para o vídeo 2.

As figuras 5 e 6, mostram o resultado do método proposto, onde a solução encontrada é representada pelo retângulo azul. Pode-se notar que no vídeo 2 existe movimento tanto das pessoas quanto da câmera, o que não influênciou fortemente a detecção.

De maneira geral, o método apresenta um nível satisfatório de detecção, com base nos valores das taxas de acerto. Ao analisar os dados deve-se levar em consideração dois pesos para classificar a qualidade dos modelos que são a quantidade de quadros por segundo e a porcentagem de acerto. Sob este ponto de vista, pode-se notar que o modelo que utiliza PCA com redução de 3780 para 38 componentes forneceu o melhor resultado em ambos os casos de teste. É notável que a maior influência da aplicação do PCA foi na quantidade de quadros por segundo, visto que a redução realizada pelo PCA permite uma grande redução na quantidade dos vetores utilizados pelo SVM.

A referência [1] emprega PSO + HOG + SVM para realizar a detecção de pessoas em imagens estáticas. Os autores definiram a população do PSO com 20 partículas avaliadas por 10 gerações. Nestas condições, conseguiram atingir a quantidade de 12 quadros por segundo de processamento resultando em uma taxa de acerto de cerca de 70%. Comparando os resultados, nota-se um grande aumento na quantidade de quadros avaliados, quase 3 vezes mais, com um ganho considerável na taxa de detecção, cerca de 10%.

Por utilizar uma heurística para otimizar a posição da janela de detecção, não se tem a garantia que o método sempre irá encontrar uma pessoa no quadro. As execuções que não foi detectada a posição de uma pessoa na cena, mesmo existindo uma ou mais, deve-se pelo fato de o PSO ficar preso em um ótimo local e não conseguir. Além disso, uma classificação errada fornecida pelo SVM também compromete o processo de busca realizado pelo PSO.

A metodologia desenvolvida aqui foi pensada para realizar a detecção de apenas uma pessoa por execução. Nos quadros onde aparecem mais de uma pessoa, a tendência do método é continuar seguindo a pessoa previamente detectada. Pode ocorrer que durante a busca o método detecte uma ou outra pessoa, devido a característica não-determinística do PSO.

5. Conclusões

A detecção de pessoas em vídeos é um ponto fundamental para que um sistema de análise automática possa funcionar. O método proposto busca encontrar uma solução com baixo custo computacional e um bom nível de desempenho, o que pode ser comprovado pelos resultados obtidos. Os artigos que serviram de base para o desenvolvimento deste trabalho, empregam as técnicas apenas para imagens estáticas. O método proposto amplia o uso das técnicas para vídeos, sendo assim uma proposta de solução para a detecção de pessoas em vídeos.

Sabe-se que para o uso efetivo deste método necessita-se aumentar as taxas de detecção, bem como permitir que mais de uma pessoa seja localizada em um único quadro. Para o primeiro ponto, o emprego de outros classificadores, bem como a utilização de um conjunto de treinamento mais amplo podem fornecer resultados melhores. Para o segundo ponto é necessário a alteração na arquitetura do SDP, mais especificamente na implementação do PSO, tornando este multimodal. Estes pontos serão objeto de trabalhos futuros.

Referências

[1] S.-T. An, J.-J. Kim, J.-W. Lee, and J.-J. Lee. Fast human detection using gaussian particle swarm optimization. In *Digital Ecosystems and Technologies Conference (DEST), 2011 Proceedings of the 5th IEEE International Conference on*, pages 143–146, 31 2011-june 3 2011.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

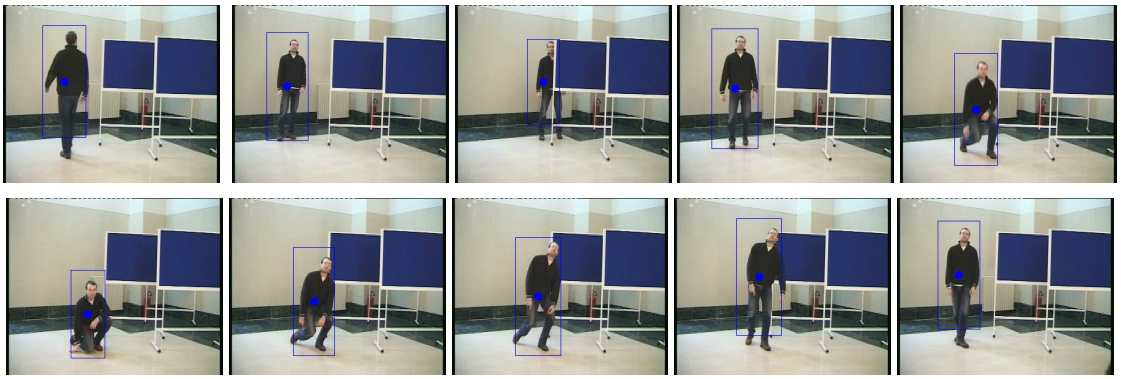

Figura 5. Alguns quadros do vídeo 1 mostrando os resultados obtidos. A solução encontrada pelo método é o retângulo azul.

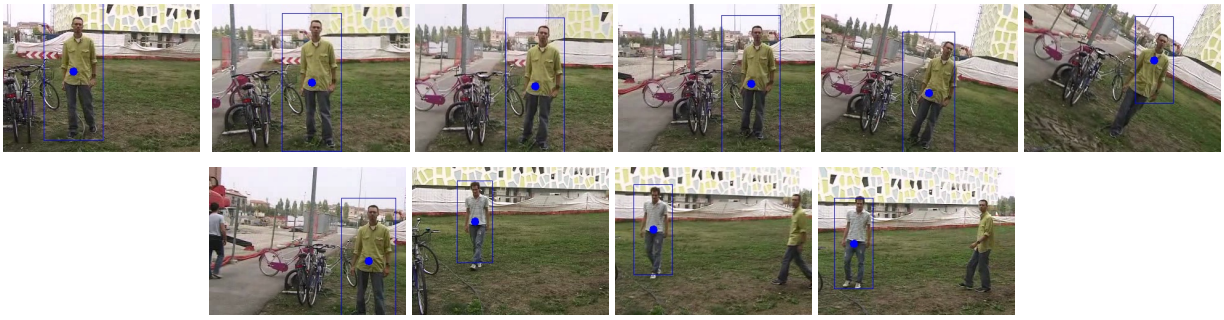

Figura 6. Resultados do método para o vídeo 2, onde além da movimentação das pessoas existe também a movimentação da câmera. Nota-se que mesmo com um certo nível de rotação o método foi capaz de detectar a pessoa com acurácia razoável. A solução encontrada pelo método é o retângulo azul.

[3] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2001.

[4] C. Fernández, P. Baiget, F. X. Roca, and J. Gonzílez. Augmenting video surveillance footage with virtual agents for incremental event evaluation. *Pattern Recogn. Lett.*, 32(6):878–889, Apr. 2011.

[5] O. L. Junior, D. Delgado, V. Goncalves, and U. Nunes. Trainable classifier-fusion schemes: An application to pedestrian detection. *2009 12th International IEEE Conference on Intelligent Transportation Systems*, (1):1–6, 2009.

[6] H. A. Perlin, H. S. Lopes, and T. M. Centeno. Particle swarm optimization for object recognition in computer vision. In *Proceedings of the 21st international conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems: New Frontiers in Applied Artificial Intelligence*, IEA/AIE '08, pages 11–21, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] R. Vezzani and R. Cucchiara. Video surveillance online repository (visor): an integrated framework. *Multimedia Tools Appl.*, 50:359–380, November 2010.

[8] X. Wang and T. X. Han. An hog-lbp human detector with partial occlusion handling. *Computer Engineering*, pages(Iccv):32–39, 2009.

[9] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.*, 115(2):224–241, Feb. 2011.

[10] H. Yang, Z. Song, and R. Chen. An incremental pca-hog descriptor for robust visual hand tracking. *Lecture Notes in Computer Science*, 6454:687–695, 2010.

[11] C. Zeng, H. Ma, and A. Ming. Fast human detection using mi-svm and a cascade of hog-lbp features. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3845 –3848, sept. 2010.