Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

CONTAGEM DE PESSOAS POR VÍDEO USANDO CÂMERAS EM POSIÇÃO ZENITAL

Aluno: Victor Hugo Cunha de Melo
Matricula: 08.1.4047

Orientador: David Menotti

Ouro Preto
15 de setembro de 2011

Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

CONTAGEM DE PESSOAS POR VÍDEO USANDO CÂMERA EM POSIÇÃO ZENITAL

Proposta de monografia apresentada ao curso de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto, como requisito parcial para a conclusão da disciplina Monografia II (BCC391).

Aluno: Victor Hugo Cunha de Melo
Matricula: 08.1.4047

Orientador: David Menotti

Ouro Preto
15 de setembro de 2011

Resumo

Detecção, rastreamento e contagem de pessoas são de grande utilidade para diversas aplicações comerciais, como monitoramento de espaços públicos, estádios de futebol, ou estações de ônibus. Neste projeto, propõe-se estudar e implementar métodos de contagem de pessoas por vídeo usando câmeras em posição zenital (rotacionada em 180 graus), que preservam a privacidade das pessoas. Ainda, propõe-se avaliar densamente os métodos estudados com vários vídeos adquiridos em ambientes distintos.

Palavras-chave: Reconhecimento de padrões. Contagem de pessoas. Câmera zenital.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Detecção, rastreamento e contagem de pessoas são de grande utilidade para diversas aplicações comerciais, como monitoramento de espaços públicos, estádios de futebol, ou estações de ônibus. Possui grandes implicações em segurança, e permite coletar informações dos sistemas de forma que possam ser utilizados para identificar padrões em tráfego por horário, otimizar agendamento de trabalhos, monitorar a efetividade de eventos, *etc*.

Além de sensores de imagens, formas mecânicas e outras tecnologias de sensores são utilizadas para contagem de pessoas. Os sistemas que utilizam contadores mecânicos, como catracas, contam apenas uma pessoa por vez e podem obstruir a passagem, ocasionando congestionamentos se há muitos transeuntes. Devido ao seu projeto, está sujeita a subcontagens. Sistemas que utilizam raios infravermelhos ou sensores de calor não bloqueiam as portas, mas não apresentam precisão para identificar pessoas em um grupo. É notável a necessidade de sistemas mais precisos, por isso câmeras foram selecionadas como instrumento de detecção.

Existem vários métodos propostos na literatura para a contagem de pessoas por vídeo [Bescos et al., 2003, Chien et al., 2004, Huang and Chow, 2003, Snidaro et al., 2005, Velipasalar et al., 2006]. No entanto, é difícil encontrar uma validação efetiva e extensa dos métodos propostos. Ainda, estes trabalhos não levam em conta a privacidade das pessoas sendo filmadas.

Neste projeto, propõe-se estudar e implementar métodos de contagem de pessoas por vídeo usando câmeras em posição zenital[1] [Antic et al., 2009, Barandiaran et al., 2008, Chen et al., 2008]. Ainda, propõe-se avaliar densamente os métodos estudados com vários vídeos sob diversas condições.

[1] rotacionada azimutalmente em 180 graus

2 Justificativa

O projeto em questão é relevante tanto nas esferas social e ambiental, quanto na computacional.

O produto final que pode ser gerado com a conclusão deste projeto é um sistema de controle de acesso de pessoas as dependências do DECOM. Tal sistema, por meio da quantificação das pessoas nas dependências do DECOM, pode oferecer maior segurança aos alunos, professores e funcionários da universidade. Ainda, aumentará a visibilidade dentro da universidade tanto da pesquisa (PROPP e PPGCC) quanto do Departamento de Computação (DECOM).

Em caso de sucesso pleno, o projeto pode ainda ser encubado em uma empresa júnior para se tornar um produto construído dentro da própria UFOP a ser comercializado no Brasil.

Além disso, este projeto demonstra o potencial da área de processamento de imagem nas mais diversas aplicações, e seus resultados poderão servir de estímulo para que áreas, além da automação, passem a fazer uso de seus recursos.

Finalmente, conceitos da área de processamento de imagem, visão computacional e reconhecimento de padrões podem ser expandidos durante o estudo da literatura. O anseio maior deste projeto é obter um novo método para a solução desse problema particular: a contagem de pessoas por vídeo.

3 Objetivos

3.1 Objetivo geral

O objetivo geral deste projeto é pesquisar, caracterizar e implementar um método para a contagem de pessoas por vídeo usando câmeras em posição zenital.

3.2 Objetivos específicos

Os objetivos específicos a serem atingidos são:

1. Fazer uma revisão da literatura sobre métodos de contagem de pessoas por vídeo usando câmeras em posição zenital.
2. Fazer uma revisão da literatura sobre processamento de imagem, visão computacional e reconhecimento de padrões, visando:
   (a) Representação digital de imagens;
   (b) Métodos de filtragem de imagem;
   (c) Técnicas de identificação de objetos em imagens;
   (d) Técnicas de remoção de *background*;
   (e) Técnicas de rastreamento de objetos em vídeo;
3. Implementar os métodos de contagem de pessoas por vídeo;
4. Comparar e analisar os resultados obtidos pelos diferentes métodos implementados tendo como entrada vários vídeos obtidos nas dependências do DECOM em diversas épocas do dia e do ano;
5. Contribuir com a divulgação de técnicas de processamento gráfico / imagens e vídeo à solução de problemas de automação;
6. Produzir um artigo que possa ser publicado em um evento científico nacional e outro internacional e ainda outro que possa ser submetido a revista especializada;

4 Metodologia

As principais atividades previstas para esse projeto são:

1. Pesquisa de técnicas de processamento de imagem, reconhecimento de padrões e visão computacional e métodos para contagem de pessoas por vídeo;
2. Caracterização e classificação de cada método pesquisado;
3. Implementação dos métodos pesquisados;
4. Realização de testes em vídeos obtidos a partir das câmeras instaladas nas dependências do DECOM adquiridos em diversas épocas do dia e do ano;
5. Classificar os vídeos por quantidade de transeuntes e turno;
6. Desenvolver um método para estimar se a contagem de pessoas é precisa em vídeos longos, pois é inviável contar todas as pessoas do vídeo manualmente;
7. Substituir uma parte do pré-processamento dos vídeos para uma linguagem mais eficiente, como C/C++;
8. Organização dos resultados obtidos;
9. Análise dos resultados;
10. Possível proposição e implementação de um novo método;
11. Preparação de artigos e pôsteres.

Considerando as atividades descritas, de forma geral, a metodologia prevista, para cada atividade discutida é a seguinte:

1. Inicialmente será feito um estudo das técnicas de processamento de imagem, reconhecimento de padrões e visão computacional, seguido de estudo de métodos para a contagem de pessoas por vídeo. Durante este processo estar-se-á classificando os métodos estudados;
2. Em seguida, os métodos estudados serão avaliados, implementados e testados;
3. O sistema proposto para estimar a precisão de um método dado um vídeo de longa duração consiste, em suma: a partir de um ponto do vídeo, analisar manualmente a quantidade de pessoas que estão entrado/saindo e quantas estão se deslocando para cima/baixo de um *frame* para outro. Este procedimento será efetuado durante 10 minutos de vídeo;
4. Por fim, validar-se-á os métodos submetendo a eles vídeos contendo pessoas, visando avaliar a precisão dos métodos estudados. Espera-se também propor uma metodologia para avaliação automática dos métods. Dessa forma, espera-se poder testar e validar a eficácia dos métodos estudados.

Observação: Para a realização dessas atividades, o projeto conta com o Laboratório de Processamento Digital de Imagem (LaPDI), localizado à sala COM20 do ICEB, com toda a infra-estrutura necessária ao desenvolvimento do projeto, no caso, microcomputadores e *softwares* "livres", *scanner*, câmeras digitais e servidores com alta capacidade de armazenamento. Uma câmera necessária para aquisição dos vídeos já está instalada no corredor da secretaria do DECOM. Ainda, o laboratório conta com bibliografia especializada (livros internacionais) sobre processamento de imagem, reconhecimento de padrões e visão computacional.

5 Atividades Desenvolvidas

Dentre os objetivos listados, o processo de revisão da literatura nos tópicos de contagem de pessoas por vídeo usando câmeras em posição zenital e de processamento digital de imagens, visão computacional e reconhecimento de padrões é realizado constantemente.

Em atividades anteriores já havíamos implementado o método proposto por [Antic et al., 2009]. Sua descrição é detalhada na Seção 5.2.

O segundo método que optamos por implementar é a solução proposta por [Barandiaran et al., 2008]. Detalhes de sua implementação constam na Seção 5.3. Como havia uma dificuldade em concluí-lo, entramos em contato com os autores do artigo que nos auxiliaram em uma parte do procedimento final da contagem. Não obstante, o método permanece incompleto.

5.1 Base de Dados

A base de vídeos atualmente é composta por cinco vídeos. A câmera, situada no corredor do Departamento de Computação, capturou vídeos de duração de uma hora, durante o dia e a noite. Os vídeos tem resolução de 640×480, e utilizam o padrão JPEG. Pretendemos incorporar à base outros cinco vídeos.

Para a execução dos testes apresentados na Seção 5.4 foram utilizados dois vídeos curtos fornecidos pelos autores.

5.2 O Primeiro Método

O método para contagem de pessoas é dividido em: captura do vídeo, subtração do fundo, segmentação, rastreamento e contagem de pessoas (Figura 1). As operações nos *frames* do vídeo são feitas em blocos de pixels, o que reduz a quantidade de computações e o efeito obtido é o mesmo se essas operações fossem feitas pixel a pixel. O tamanho padrão para os blocos é 8x8.

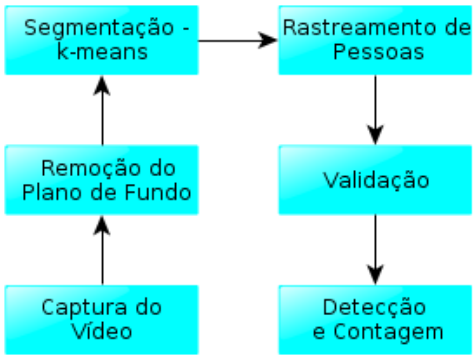


Figura 1: Fluxograma para representação do sistema

A primeira parte do método é a subtração do fundo. Essa operação é essencial para a detecção das pessoas que será realizada posteriormente, através da comparação dos blocos do *frame* atual com os blocos do *frame* pertencente ao fundo. As imagens (*frames*) que pertencem ao fundo do vídeo são obtidas através do seguinte filtro

$$F^{t+1} = (1 - \alpha) \cdot F^t + \alpha \cdot I^t \tag{1}$$

onde F e I representam, respectivamente, os *frames* de fundo e os *frames* do vídeo original; t é o número do *frame*; e α é uma taxa de aprendizado que pode variar entre 0.01 e 0.1. Essa taxa deve ser ajustada de acordo com a situação, porém foi escolhido 0.01 como padrão. O filtro é aplicado sobre todos os *frames* e todos os seus canais de cada *frame*.

O algoritmo utiliza fatores multiplicativos $\beta^t_{m,n,p}$, determinados através estimativa máxima de verossimilhança (MLE). MLE é um método estatístico utilizado para ajustar os dados a um modelo e fornecer estimativas para os parâmetros do modelo. Os índices (m, n) referem-se às coordenadas dos blocos e p aos canais da imagem (RGB - vermelho, verde e azul).

$$\beta^t_{m,n,p} = \frac{\sum I^t_{m,n,p} \cdot F^t_{m,n,p}}{\sum (F^t_{m,n,p})^2} \tag{2}$$

A detecção de pessoas nos *frames* é realizada através da diferença entre os fatores multiplicativos máximo e mínimo. São calculados o maior e o menor β entre os canais da imagem e a diferença entre eles é guardada em $\delta\beta^t$, para cada *frame*.

$$\delta\beta^t = \max_p \beta^t_{m,n,p} - \min_p \beta^t_{m,n,p} \tag{3}$$

Os fatores multiplicativos dos blocos do fundo tem valor aproximado de 1. Se $\delta\beta^t$ não é pequeno ou se algum fator multiplicativo é muito diferente de 1, o bloco pertence ao primeiro plano.

$$P^t = \begin{cases} 1, \text{se } \delta\beta^t > T_1 \vee |\beta^t_{m,n,p}| > T_2 \\ 0, \text{caso contrário} \end{cases} \tag{4}$$

P é a imagem com pessoas e T_1, T_2 são limites entre $[0.1, 0.2]$ e $[0.3, 0.6]$, respectivamente. Esses parâmetros também devem ser ajustados através de experimentos para cada situação específica.

Nesse momento, há uma imagem P para cada *frame* e são essas imagens que as pessoas aparecem. O próximo passo do algoritmo é a segmentação dessas pessoas. A segmentação é um problema difícil em Análise de Imagem, devido às várias características que representam uma pessoa. Como os vídeos em questão as pessoas aparecem na forma zenital (por cima), esse problema é reduzido. Assim as pessoas passam a ser vistas como formas geométricas (Figura 2), o que pode ser extraído através de técnicas tradicionais de *clustering* como o *k-means*.

No *k-means* Duda et al. [2000] existem k centróides, um para cada grupo - ou *cluster*. Cada indivíduo é associado ao centróide mais próximo e os centróides são recalculados com base nos indivíduos classificados. No entanto, o valor de k não é conhecido *a priori*. O valor de k é exatamente o número de pessoas na cena.

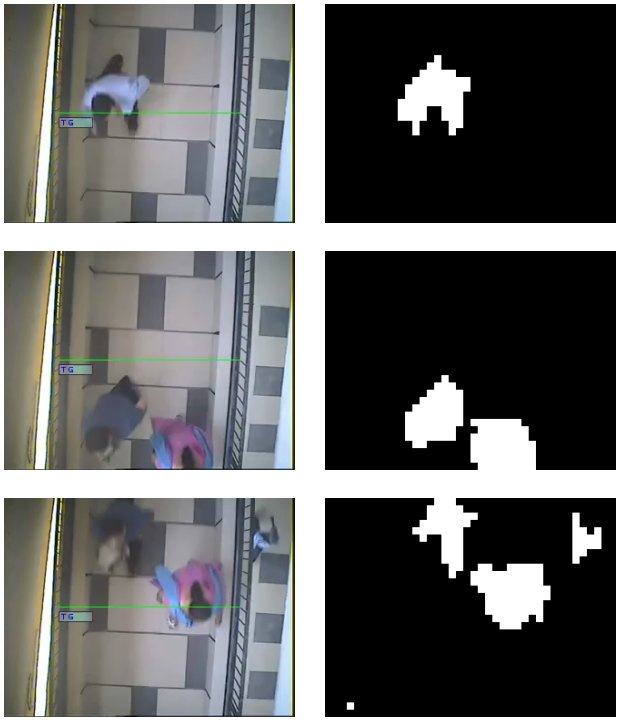
Então, o valor de k é estimado como o número máximo de *clusters* em que a distância dentro dos *clusters* é maior do que uma distância mínima D_{min}. Essa constante corresponde ao tamanho médio de uma pessoa na cena, e deve ser estabelecida através de experimentos. Em uma imagem com k *clusters*, cujos centróides são $C_i, i = 1, 2, ..., k$, a distância mínima dentro do *cluster* é definida como

$$d^k_{min} = \min_{1 \leq i < j \leq k} ||C_i - C_j|| \tag{5}$$

No caso de apenas um *cluster*, definimos formalmente $d^1_{min} = \infty$. O número atual de *clusters* $k*$ é então estimado como o máximo número de *clusters* que possuem a distância mínima dentro do *cluster* d^k_{min} maior que D_{min} .

$$k* = \max\{k | d^k_{min} \geq D_{min} \wedge d^{k+1}_{min} < D_{min}\} \tag{6}$$

No *k-means*, a inicialização dos centróides é muito importante pois pode-se melhorar a convergência do algoritmo. Sempre que possível, os centróides são inicializados na posição dos centróides encontrados na iteração anterior. Dessa forma os centróides são inicializados com uma posição muito provável de ser a melhor para os *clusters*, pois o deslocamento de uma pessoa no vídeo é pequeno.



(a) (b)

Figura 2: Resultados demonstrativos das etapas de subtração do fundo e segmentação de pessoas. (a) *Frame* original. (b) Subtração do fundo e segmentação de pessoas.

Nesse ponto do algoritmo são conhecidas as pessoas em cada *frame* do vídeo. A próxima parte é fazer o rastreamento dessas pessoas, ou seja, descobrir se a mesma pessoa está em vários frames consecutivos para então contá-las. Esse passo foi implementado de forma *gulosa*, analisando dois *frames* consecutivos por vez.

O algoritmo encontra os *clusters* correspondentes em dois *frames* consecutivos que possuem a menor distância. O objetivo é obter a menor distância Euclidiana quadrada entre os *clusters*. Então esses *clusters* com a distância mínima são marcados como a mesma pessoa em uma matriz binária, onde as linhas representam os *clusters* e as colunas representam os *frames*.

Dessa forma, se o *cluster* i do *frame* t corresponde ao *cluster* i do *frame* $t+1$, a matriz na posição (i,t) possui valor 1. Ao final de todas as iterações, essa matriz possui o valor 1 nos intervalos em que a mesma pessoa está em vários frames.

$$M_{i,t} = \begin{cases} 1, \text{se } c_i^t = c_i^{t+1} \\ 0, \text{caso contrário} \end{cases} \tag{7}$$

onde $M_{i,t}$ representa a matriz binária do *cluster* i, no *frame* t. c_i^t equivale ao *cluster* i do *frame* t.

O último passo é contar as pessoas. Essa parte é feita através da análise da matriz binária construída no passo anterior. Como cada linha dessa matriz representa um *cluster*, é necessário analisar cada linha separadamente. Uma pessoa é detectada a cada variação de 0 para 1 detectada na matriz binária.

A Figura 2 demonstra os principais passos do algoritmo. As imagens da primeira coluna mostram os *frames* originais. As imagens da segunda coluna ilustram a subtração do fundo através de blocos proposta (blocos de tamanho 8 x 8 pixels) seguida da segmentação de pessoas.

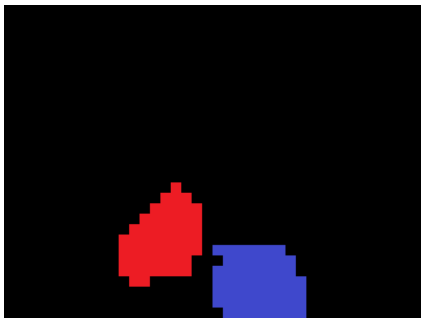


Figura 3: Após segmentação pelo *k-means*

A Figura 3 apresenta o resultado da segmentação de pessoas através do *k-means*, onde o número de *clusters* é automaticamente determinado usando a distância mínima inter-cluster. Nesse caso, o método segmentou corretamente encontrando o valor de $k = 2$, *i.e.*, duas pessoas.

5.3 O Segundo Método

A idéia principal por trás desta solução consiste em definir uma área de interesse (*ROI, region of interest*) nas imagens onde analisamos os movimentos das pessoas. Inserimos linhas virtuais ortogonais a direção de movimento.

O algoritmo é dividido em três passos diferentes. Primeiro detectamos o movimento e extraímos a região onde as pessoas passaram. Após este passo, uma contagem cumulativa é efetuada pelas linhas virtuais. Finalmente, cada linha é examinada para a contagem de pessoas.

5.3.1 Extração do Movimento

Para extrair as informações de movimentação do plano de fundo, o método utiliza a diferença entre dois *frames* consecutivos. Utiliza-se um limiar arbitrário para filtrar os ruídos da imagem final.

5.3.2 Contagem de Pessoas

A segunda parte do algoritmo consiste na contagem realizada independentemente para cada linha pertencente a área de interesse das imagens. Cada linha é representada por uma função l, onde os eixos x e y correspondem a posição dentro da linha com o número acumulado de *pixels* da imagem.

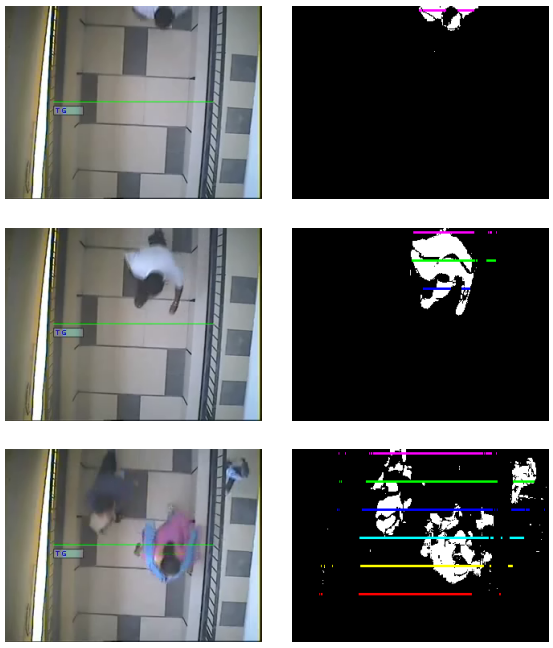
$$l_x^t = l_x^{t-1} + D^t \;\; \forall x \wedge 0 \leq x < czw \tag{8}$$

$$l_x^0 = 0 \;\; \forall x \wedge 0 \leq x < czw \tag{9}$$

onde x é o ponto na linha; czw é a largura na área de interesse, que é igual a largura das linhas.

Os *pixels* são acumulados a cada vez que uma pessoa passa através de uma linha (Equação 8).

A Figura 4 mostra estes passos do algoritmo. As imagens na coluna da Figura 4(a), exibe as imagens originais do vídeo. As imagens na coluna da Figura 4(b) ilustram a detecção de movimento de algumas pessoas por meio deste método. Inicialmente, não há nenhuma linha virtual pois ninguém cruzou a região de interesse. As linhas surgem quando uma pessoa cruza a região.



(a) (b)

Figura 4: Resultados do passo-a-passo do método de *multiple lines*.

Não apresentamos resultados dos testes porque faltam detalhes sobre a parte final do algoritmo para concluirmos sua implementação. Os autores esclareceram que esta etapa é como um esquema de votação. Cada linha vota na quantidade de pessoas que entraram e saíram separadamente. A contagem final é aquela que obteve o maior consenso.

Tabela 1: Resultados do primeiro método

	real (a)	método (a)	real (b)	método (b)
pessoas	6	7	6	5
TP	6	7	6	5
FP + FN	0+0	1+0	0+0	0+1
precisão	1.00	0.87	1.00	1.00
recall	1.00	1.00	1.00	0.83
F-score	1.00	0.93	1.00	0.90

5.4 Resultados Experimentais e Análise

Nesta seção apresentamos os resultados da implementação do primeiro método (Seção 5.2). O segundo método ainda está em fase de desenvolvimento, e por isso não apresenta nenhum resultado.

Para mensurar a corretude do método, foram utilizados as métricas de *recall* e precisão [Baeza-Yates and Ribeiro-Neto, 1999].

Analisando a *F-score*, que pode ser interpretada como uma média das métricas de precisão e *recall*, para ambos os vídeos o algoritmo obteve uma média de 0.91 de precisão. Esse resultado pode ser considerado satisfatório, porém a contagem de pessoas deve ser realizada com exatidão. Por isso o resultado do *F-score* precisa ser melhorado.

Analisando detalhadamente o algoritmo, esses erros ocorreram por dois motivos. O primeiro é o ajuste de parâmetros, como D_{min} que é essencial ao algoritmo, porém não é trivial determinar o valor apropriado. O outro motivo são os ruídos nas imagens. É necessário filtrá-las para aprimorar os resultados.

6 Cronograma de atividades

Na Tabela 2 é apresentada uma proposta de cronograma com as seguintes tarefas:

1. Coleta de vídeos;
2. Classificação e sistema para avaliação de precisão dos métodos;
3. Implementação;
4. Testes e análises;
5. Relatório de atividades;
6. Redação da monografia;
7. Apresentação do trabalho.

Atividades	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**
1		V	V	V			X	X		
2	V	V	V	V	V	V	X	X		
3							X	X		
4				V	V	V	X	X		
5				V						
6						V	X	X	X	
7										X

Tabela 2: Cronograma de Atividades.

Referências

B. Antic, D. Letic, D. Culibrk, and V. Crnojevic. K-means based segmentation for real-time zenithal people counting. In *IEEE International Conference on Image Processing (ICIP)*, pages 2565 –2568, 2009.

Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

J. Barandiaran, B. Murguia, and F. Boto. Real-time people counting using multiple lines. In *International Workshop on Image Analysis for Multimedia Interactive Services*, pages 159–162, 2008.

J. Bescos, J. M. Menendez, and N. Garcia. DCT based segmentation applied to a scalable zenithal people counter. In *IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 1005–1008, 2003.

C.H. Chen, Y.C. Chang, T.Y. Chen, and D.J. Wang. People counting system for getting in/out of a bus based on video processing. In *International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 565–569, 2008.

S.Y. Chien, Y.W. Huang, B.Y. Hsieh, S.Y. Ma, and L.G. Chen. Fast video segmentation algorithm with shadow cancellation, global motion compensation, and adaptive threshold techniques. *IEEE Transactions on Multimedia*, 6(5):732–748, 2004.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2000.

D. Huang and T. W. S. Chow. A people-counting system using a hybrid RBF neural network. *Neural Processing Letters*, 18:97–113, 2003.

L. Snidaro, C. Micheloni, and C. Chiavedale. Video security for ambient intelligence. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 35(1):133–144, 2005.

S. Velipasalar, Y.l. Tian, and A. Hampapur. Automatic counting of interacting people by using a single uncalibrated camera. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1265–1268, 2006.