

Uma Implementação de um Sistema de Contagem de Pessoas Baseado em Vídeo Processamento

Suellen Silva de Almeida

susilvaalmeida@gmail.com

Victor Hugo Cunha de Melo

victorhcmelo@gmail.com

Jean Carlos Mendes

carlomendes@yahoo.com.br

David Menotti

menottid@gmail.com

Universidade Federal de Ouro Preto
Campus Universitário, Morro do Cruzeiro
Ouro Preto - Minas Gerais, Brasil

Abstract

This paper presents a evaluation of an approach to people segmentation, tracking and counting, developed for a system using a zenithal camera. The initial step consists of block-wise background subtraction, followed by k-means clustering to allow segmentation of single persons in the scene. The number of people in the scene is estimated as the maximal number of clusters with acceptable inter-cluster separation. Tracking of segmented people is addressed as a problem of dynamic cluster assignment between two consecutive frames and it is solved in a greedy fashion. Further details of this greedy solution can be found in this paper.

Resumo— O artigo apresenta a avaliação de uma abordagem para segmentação de pessoas, rastreamento e contagem, desenvolvido para um sistema utilizando câmera com posicionamento zenital. A partir do passo inicial de remoção do segundo plano por blocos, a clusterização por k-means é utilizada para permitir a segmentação de pessoas individuais em uma cena. O número de pessoas de uma cena é estimado como o número máximo de clusters com uma separação inter-cluster aceitável. O rastreamento de pessoas segmentadas é visto como um problema de associação de clusters dinâmicos entre dois frames consecutivos e é resolvido por uma estratégia gulosa. Maiores detalhes sobre esta estratégia podem ser encontradas neste artigo.

Palavras-chave— People counting, People segmentation, Background subtraction, People tracking.

1. Introdução

Deteção, rastreamento e contagem de pessoas [1] é de grande utilidade para diversas aplicações comerciais, como monitoramento de espaços públicos, estádios de futebol, ou estações de ônibus. Tem 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 [14]. 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 informativos, por isso câmeras foram selecionadas como instrumento de detecção.

A segmentação do plano de fundo é o primeiro passo em diversas aplicações de visão computacional. Usualmente é obtido em sistemas de detecção humana calculando a diferença pixel-a-pixel entre o *frame* atual e a imagem do plano de fundo, seguido por uma limiarização automática [14] [7] [13]. Se a acurácia que essa abordagem oferece não é garantida [10] [4], uma estratégia por blocos é preferível pois produz uma segmentação mais estável na presença de mudanças de luz e sombra.

Rossi e Bozzoli [12] utilizam características em tons de cinza sensíveis a mudanças de alta frequência na cena para detectar objetos em movimento. Eles usam *template matching* para rastrear as características extraídas. Huang

e Chow [10] utilizam características mais elaboradas para descrever os borrões do primeiro plano. Ao invés de rastrear os indivíduos, eles simplesmente contam o número de pessoas na região de interesse. Velipasalar *et al.* [14] propõe o uso do tamanho dos borrões detectados para segmentar pessoas individuais e do procedimento de *mean-shift* como forma de lidar com borrões mesclados. Em [3], Beleznaï *et al.* também empregam o procedimento *mean-shift* para desenvolver um sistema de rastreamento de pessoas geral.

Alguns autores [11] [9] utilizam o posicionamento de câmera oblíquo. Este permite detectar mais características, porém apresenta problemas com relação a oclusões e quanto a privacidade das pessoas. O posicionamento da câmera zenital, por sua vez, consiste em uma câmera posicionada sobre a cabeça das pessoas que remove eficientemente o problema de oclusões entre os objetos. Além disso, o posicionamento zenital oferece outras vantagens [14] [5], como:

- objetos aparecem em tamanhos relativamente constantes;
- proporciona melhor visão de pessoas no cenário;
- privacidade, pois não reconhece o rosto da pessoa;
- elimina a necessidade de calibração;
- simples e fácil de manter.

Entretanto, a segmentação resultante geralmente contém borrões mesclados, pertencentes a várias pessoas muito próximas.

Em [6] é proposto um método para contagem de pessoas que entram ou saem de um ônibus baseado no processamento de um vídeo, cujo posicionamento da câmera também é zenital. Cada *frame* capturado é dividido em vários blocos e cada bloco é classificado de acordo com seu vetor de movimentos. Se a quantidade de blocos com vetores similares de movimento maior que um limiar, esses blocos pertencem ao mesmo objeto em movimento. Como resultado, o número de objetos em movimento é o número de passageiros entrando ou saindo do ônibus. O problema de agitação na câmera e variação da iluminação no ônibus é superado nesse método, que apresentou 92% de precisão nos testes realizados.

Neste artigo é apresentada uma avaliação de um método para segmentação de pessoas, rastreamento e contagem [1]. Descrevemos com mais detalhes o método e, principalmente, o algoritmo guloso para rastreamento de pessoas.

O artigo é organizado da seguinte forma. A Seção 2 apresenta uma explicação detalhada sobre o sistema. Na Seção 3 encontra-se a análise dos resultados experimentais. Conclusões são apresentadas na Seção 4 e trabalhos futuros na Seção 5.

2. Arquitetura do Sistema

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.

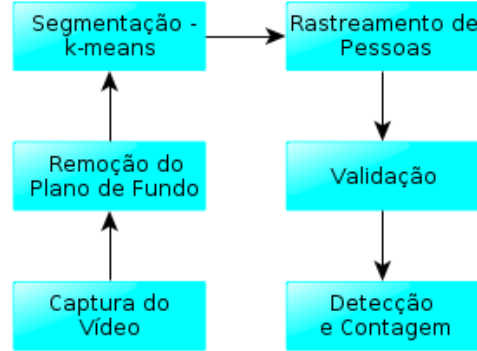


Figura 1. Fluxograma para representação do sistema

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

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

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 $\beta_{m,n,p}^t$, 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).

$$\beta_{m,n,p}^t = \frac{\sum I_{m,n,p}^t \cdot F_{m,n,p}^t}{\sum (F_{m,n,p}^t)^2} \quad (2)$$

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*.

$$\delta\beta^t = \max_p \beta_{m,n,p}^t - \min_p \beta_{m,n,p}^t \quad (3)$$

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.

$$P^t = \begin{cases} 1, & \text{se } \delta\beta^t > T_1 \vee |\beta_{m,n,p}^t| > T_2 \\ 0, & \text{caso contrario} \end{cases} \quad (4)$$

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* [8] existem k centróides, um para cada grupo (*cluster*). Cada indivíduo é associado ao centróide mais próximo e os centróides são recalculados com base nos indivíduos classificados. No entanto, não o valor de k não é conhecido a priori, e é esse valor que queremos descobrir, 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 $C_i, i = 1, 2, \dots, k$, a distância mínima dentro do *cluster* é definida como

$$d_{min}^k = \min_{1 \leq i < j \leq k} \|C_i - C_j\| \quad (5)$$

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

$$k^* = \max\{k | d_{min}^k \geq D_{min} \wedge d_{min}^{k+1} < D_{min}\} \quad (6)$$

No *k-means*, a inicialização dos centróides é muito importante pois pode-se melhorar a convergência do algoritmo. Então sempre que possível, os centróides do *k-means* foi inicializado como os centróides encontrados na iteração anterior. Dessa forma os centróides são sempre inicializados com uma posição muito provável de ser a melhor para os *clusters*.

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 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

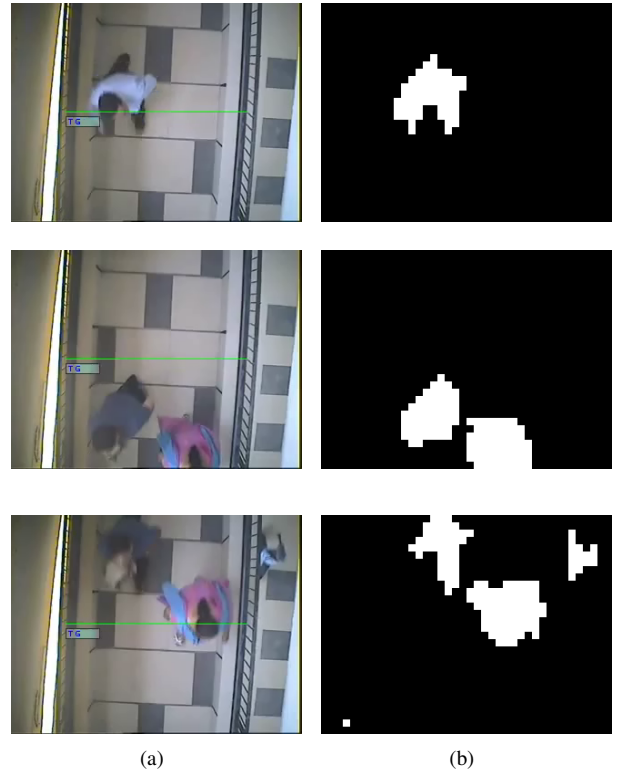


Figura 2. Resultados demonstrativos das etapas de subtração do fundo e segmentação de pessoas. (a) *Frame* original. (b) Subtração do fundo e segmentação de pessoas.

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.

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

$M_{i,t}$ 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. Ao percorrer essas linhas, de ocorre uma mudança de 0 para 1, é porque uma pessoa foi detectada e o contador é incrementando.

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.

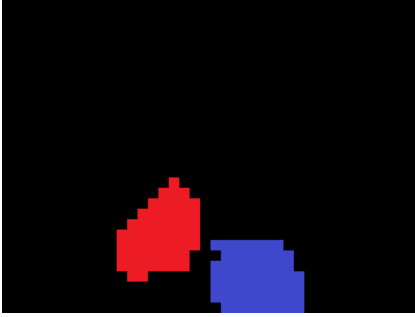


Figura 3. Após segmentação pelo *k-means*

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.

3. Resultados Experimentais e Análise

Para avaliar o desempenho do método proposto, foram utilizados dois vídeos gerados por câmeras colocadas na posição zenital. O primeiro vídeo foi filmado em um terminal rodoviário (b) (320 x 240 pixels, 145 frames) e o segundo, fornecido pelos autores do artigo [1], foi filmado em um escritório (a) (640 x 480 pixels, 524 frames). Ambas as sequências possuem no máximo três pessoas na cena ao mesmo tempo. O valor de D_{min} variou bastante entre os dois vídeos. Isso ocorreu devido à distância da câmera do chão que é diferente entre os dois, além das dimensões dos vídeos que também são diferentes.

Tabela 1. Comparação dos resultados do método apresentado com relação ao número real de pessoas

	verdade (a)	método (a)	verdade (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

Para avaliar a corretude do método, foram calculadas a precisão e o *recall* [2], métricas frequentemente utilizadas em algoritmos de Reconhecimento de Padrões. Ao usar precisão e recall, o conjunto de rótulos possíveis para uma determinada instância é dividido em dois subgrupos, um dos quais é considerado relevante para os objetivos da métrica. *Recall* é então calculado como a fração de instâncias corretas entre todas as instâncias que realmente pertencem ao subconjunto relevante. A precisão é a fração de instâncias corretas entre aqueles que o algoritmo considera pertencer ao subconjunto relevante. A precisão pode ser vista como uma medida de exatidão ou de fidelidade, enquanto o *recall* é uma medida de completude.

Os termos verdadeiro positivo (TP), verdadeiro negativo (TN), falso positivo (FP) e falso negativo (FN) são utilizados para comparar a classificação de um item (de acordo com um algoritmo) com a real classificação desse item.

Dessa forma, precisão e *recall* foram definidas como:

$$precisao = \frac{TP}{TP + FP} \quad (8)$$

$$recall = \frac{TP}{TP + FN} \quad (9)$$

Outra medida utilizada foi o *F-score*, que combina a precisão com o recall. Essa métrica pode ser interpretada como uma média ponderada da precisão e *recall*, onde uma pontuação do F-score atinge o seu melhor valor em 1 e o pior resultado em 0.

$$F = \frac{2 \times precisao \times recall}{precisao + recall} \quad (10)$$

A avaliação estatística para a segmentação proposta está na tabela acima. Os valores apresentados nessa tabela mostram que o método é eficiente por apresentar uma taxa de

Tabela 2. Resultados do artigo original [1]

	Método Original
pessoas	20
TP	20
FP+FN	0+1
<i>precisao</i>	1.00
<i>recall</i>	0.95
<i>F-score</i>	0.97

segmentação correta de aproximadamente 90% no caso do escritório. Já no terminal, essa taxa foi menor, de 83%. É possível perceber que para o vídeo do terminal (b), o resultado da precisão foi melhor, o que significa que para esse caso, o algoritmo obteve sucesso em relação à exatidão. No entanto, o outro vídeo (a), obteve melhor resultado no *recall*, mostrando que o algoritmo perdeu em relação à exatidão mas ganhou em relação à completude. Analisando então o *F-score*, que pode ser interpretada como uma média das duas outras métricas, para ambos os vídeos o algoritmo obteve uma média de 0.91 de precisão. Esse resultado pode ser considerado satisfatório pois o melhor resultado é 1, porém a contagem de pessoas deve ser realizada com exatidão. Com 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 determiná-lo não é trivial. O outro motivo são os ruídos nas imagens. É necessário filtrar mais as imagens para aprimorar os resultados.

Comparando os resultados apresentados com os resultados do artigo (Tabela 2), podemos observar que o algoritmo possui um desempenho melhor do que nossos experimentos mostraram. Os principais motivos já foram citados e mais uma característica pode ser analisada. No artigo [1], as pessoas só entram na contagem quando ultrapassam uma linha, isso eliminaria os ruídos do nosso método e possivelmente melhora o desempenho.

4. Conclusões

Neste artigo é apresentada uma avaliação de um método para segmentação de pessoas, rastreamento e contagem utilizando um sistema de câmera zenital. O algoritmo realiza a remoção do fundo seguida da segmentação de pessoas através do *k-means*. A solução gulosa para o problema de associação de *clusters* entre dois *frames* consecutivos é explorada para o rastreamento de pessoas. Sequências de vídeo foram utilizadas para avaliar os resultados. O resultado obtido não é tão preciso quanto ao artigo original devido ao problema para se determinar os parâmetros da aplicação e os ruídos não removidos dos *frames*.

5. Trabalhos Futuros

Como trabalho futuro, pretende-se implementar outros artigos para comparação de eficácia entre métodos. Além disso, desejamos substituir a segmentação via *k-means* por um algoritmo de *labeling* que poderá melhorar o desempenho (no MATLAB, implementado na função `bwlabel`). O ajuste dos parâmetros precisa ser aperfeiçoado e o filtro de remoção de ruídos deveria ser aplicado para melhorar a precisão do sistema. Nesse filtro, todos os três canais da imagem são multiplicados por uma constante dentro de um bloco (Fórmula 11).

$$I_{m,n,p}^t = \beta_{m,n,p}^t \cdot F_{m,n,p}^t + W_{m,n,p}^t \quad (11)$$

onde W denota o ruído aditivo Gaussiano branco (AWGN) e $\beta_{m,n,1} \approx \beta_{m,n,2} \approx \beta_{m,n,3}$.

Enfim, pretendemos estudar outros métodos de rastreamento, como Filtro de Partículas e Colônia de Formigas, para que avaliar sua precisão.

Referências

- [1] B. Antic, D. Letic, D. Culibrk, and V. Crnojevic. K-means based segmentation for real-time zenithal people counting. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 2565–2568, nov. 2009.
- [2] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
- [3] C. Beleznaï, B. Frühstück, and H. Bischof. Human tracking by fast mean shift mode seeking [10 font size blank 1]. 2006.
- [4] J. Bescos, J. Menendez, and N. Garcia. Dct based segmentation applied to a scalable zenithal people counter. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 3, pages III – 1005–8 vol.2, 2003.
- [5] M. Bozzoli and L. Cinque. A statistical method for people counting in crowded environments. *Image Analysis and Processing, International Conference on*, 0:506–511, 2007.
- [6] T.-Y. C. Chao-Ho (Thou-Ho) Chen, Yin-Chan Chang and D.-J. Wang. People counting system for getting in/out of a bus based on video processing. *ISDA*, pages 565–569, 2008.
- [7] 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. *Multimedia, IEEE Transactions on*, 6(5):732 – 748, 2004.
- [8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.
- [9] H. Elik, A. Hanjalic, and E. Hendriks. Towards a robust solution to people counting. In *Iterational Conference on Image Processing (ICIP)*. IEEE Computer Society, 2006.
- [10] D. Huang and T. W. S. Chow. A people-counting system using a hybrid rbf neural network. *Neural Process. Lett.*, 18:97–113, October 2003.

- [11] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos. Estimating pedestrian counts in groups. *Comput. Vis. Image Underst.*, 110(1):43–59, 2008.
- [12] M. Rossi and A. Bozzoli. Tracking and counting moving people. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 3, pages 212 –216 vol.3, Nov. 1994.
- [13] L. Snidaro, C. Micheloni, and C. Chiavedale. Video security for ambient intelligence. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 35(1):133 – 144, 2005.
- [14] S. Velipasalar, Y.-L. Tian, and A. Hampapur. Automatic counting of interacting people by using a single uncalibrated camera. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1265 –1268, 2006.