VICTOR HUGO CUNHA DE MELO

Orientador: David Menotti

UMA METODOLOGIA PARA AVALIAÇÃO DE MÉTODOS DE CONTAGEM DE PESSOAS BASEADA EM ANÁLISE DE VÍDEO

Ouro Preto

Dezembro de 2011

Universidade Federal de Ouro Preto
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

UMA METODOLOGIA PARA AVALIAÇÃO DE MÉTODOS DE CONTAGEM DE PESSOAS BASEADA EM ANÁLISE DE VÍDEO

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

VICTOR HUGO CUNHA DE MELO

Ouro Preto
Dezembro de 2011

UNIVERSIDADE FEDERAL DE OURO PRETO

FOLHA DE APROVAÇÃO

Uma Metodologia para Avaliação de Métodos de Contagem de Pessoas baseada em Análise de Vídeo

VICTOR HUGO CUNHA DE MELO

Monografia defendida e aprovada pela banca examinadora constituída por:

Dr. David Menotti – Orientador
Universidade Federal de Ouro Preto

Dr. Guillermo Cámara Chávez
Universidade Federal de Ouro Preto

Dr. José Maria Ribeiro Neves
Universidade Federal de Ouro Preto

Ouro Preto, Dezembro de 2011

Resumo

Contagem de pessoas baseada em análise de vídeo é muito útil para diversas aplicações comerciais, como o monitoramento de espaços públicos ou eventos desportivos. No entanto, os métodos presentes na literatura geralmente apenas verificam se a contagem total é correta, independente do momento em que cada contagem acontece. Neste trabalho, propomos uma metodologia para avaliação de métodos de contagem de pessoas baseados em câmeras de vídeo na posição zenital. *A priori*, é necessário indicar manualmente em um dado vídeo quando cada pessoa entra ou sai da zona de contagem, gerando os dados *ground-truth*. A partir destes dados de referência e a saída de um método de contagem por vídeo, propomos um algoritmo guloso para resolver o problema de indicar as melhores pessoas rastreadas a partir da referência para a contagem de saída. O problema é modelado como um grafo bipartido, e assim métricas como *recall*, *precision* e *F-score* são introduzidas. Usando esta metodologia, é possível quantificar automaticamente a contagem de falsos positivos e negativos dos métodos e também identificar quando esses erros acontecem.

Abstract

People counting based on video analysis is very useful for many commercial applications, such as monitoring of public spaces or sporting events. However, the methods in the literature usually only check if the total counting is correct, regardless when each counting happens. In this paper, we propose a methodology for assessment of people counting methods based on video from cameras in zenith position. Initially, it is required to manually indicate in a given video when each people get in to and out from the counting zone, generating the ground-truth data. From this reference data and the output of a people counting method for a given video, we propose a greedy algorithm to solve the problem of indicating the best tracked people from the reference to the output counting, which is modelled as a bipartite graph, and so measures such as precision, recall and F-score are introduced. By using this methodology, it is possible to automatically quantify the false positive and negative counts of the methods and also identify when these errors happen.

Aos meus pais, Porfírio e Ana Vera, ao meu irmão Caio Hess e à minha família.

Agradecimentos

Primeiramente agradeço a Deus pela força e perseverança que recebi para a realização deste trabalho.

À todos os professores do DECOM por toda a contribuição para meu conhecimento e por sua amizade, especialmente ao meu orientador David Menotti por ter acreditado em mim e pela oportunidade de ser seu orientando.

Aos meus pais, pelo amor incondicional e por sempre me incentivarem.

Ao meu irmão Caio Hess, grande amigo e companheiro para todos os momentos.

À Mari, por todo o carinho e atenção, principalmente nesta etapa final.

Aos meus amigos que participaram desta jornada. Obrigado pelas madrugadas viradas em conjunto, por este tempo que passamos juntos e pelo que vocês me ensinaram.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Detecção, rastreamento e contagem de pessoas é de grande utilidade para diversas aplicações comerciais, como o monitoramento de espaços públicos ou eventos desportivos. Estações de metrô em megalópoles tem um tráfego intenso de pessoas por dia, podendo utilizar esses sistemas para medir o fluxo de pessoas. As informações coletadas a partir do processo de contagem também auxiliam a identificar padrões de tráfego de veículos e a monitorar o público em eventos. Além disso, sistemas de vigilância podem recorrer a estes métodos para atribuir o número exato de pessoas em lugares chave de segurança, e planejar modos de evacuação eficiente.

Os métodos presentes na literatura podem ser divididos em três grandes categorias (Velipasalar et al., 2006). Na primeira categoria encontram-se os sistemas baseados em contadores mecânicos, tais como as catracas. Estas possibilitam a contagem de apenas uma pessoa por vez e podem obstruir o caminho, causando congestionamento se houver um tráfego intenso de pessoas. Devido ao seu formato, são suscetíveis a efetuar uma subcontagem – quando as pessoas não respeitam o obstáculo saltando ou passando por baixo.

A segunda categoria é composta por sistemas baseados em sensores, tais como raio infravermelho e sensores de calor. Estes não obstruem o caminho, mas também estão sujeitos ao problema de subcontagem devido a sobreposição de pessoas.

A terceira categoria é composta por métodos baseados em visão utilizando câmeras. Diversas tentativas presentes na literatura obtiveram sucesso em superar os inconvenientes das outras categorias (Velipasalar et al., 2006; Chen et al., 2008; Huang e Chow, 2003).

Um problema comum destes métodos de contagem é o posicionamento da câmera. Alguns trabalhos, como visto em Kilambi et al. (2008); Elik et al. (2006), utilizam o posicionamento oblíquo da câmera. Embora permita a detecção de mais características, apresenta problemas com relação a oclusões e a privacidade dos indivíduos.

O posicionamento zenital, por sua vez, consiste em uma câmera sobre as pessoas, rotacionada azimutalmente em 180 graus (Figura 1.1). Ela remove efetivamente o problema de oclusão entre objetos, além de oferecer vantagens adicionais (Velipasalar et al., 2006; Bozzoli e

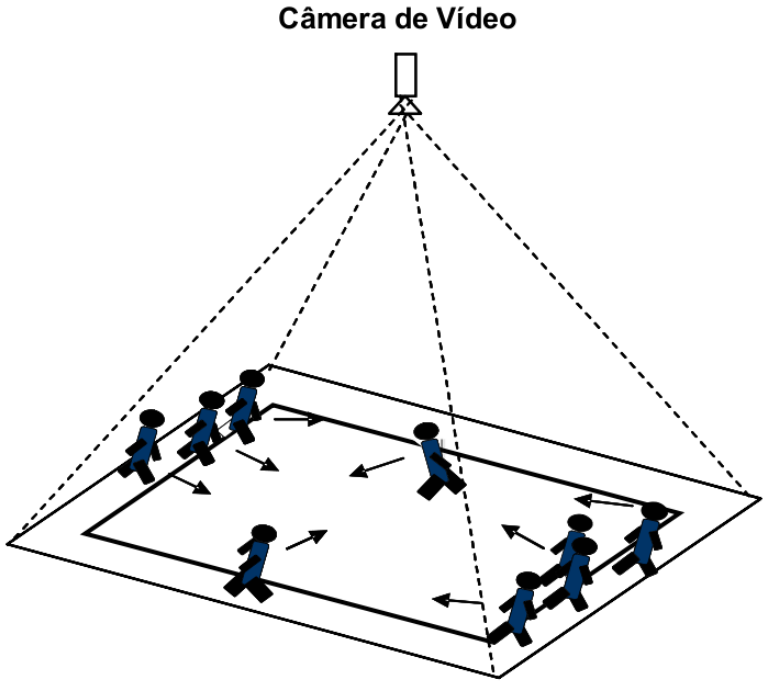


Figura 1.1: Representação de uma câmera zenital. Fonte: da Silva (2008)

Cinque, 2007), tais como não invadir a privacidade dos indivíduos e o tamanho relativamente constante dos objetos em cena.

O estudo de métodos baseados em visão com uso de câmeras apresentam dificuldades com relação à reprodução de experiências e a verificação dos resultados. Outra dificuldade é a ausência de um banco de dados de vídeo para avaliar os diferentes métodos. Como visto em Committee (2009), apenas se os dados do trabalho, bem como seus resultados, são compartilhados, é possível que novos pesquisadores verifiquem a exatidão dos dados e possam efetuar a análise dos resultados, a fim de utilizar o trabalho como base para futuras pesquisas.

A fim de superar tais dificuldades, neste trabalho, apresentamos uma metodologia para avaliação automática de métodos de contagem de pessoas. Nossa proposta de metodologia pode ser descrita resumidamente em três etapas:

- para um dado vídeo, os dados *ground-truth* são gerados manualmente, apenas uma vez, produzidos indicando quando cada pessoa entrou e saiu da zona de contagem;
- a partir destes dados de referência e a saída indicada por um método típico de contagem de pessoas, tendo como entrada um dado vídeo, é estabelecida a indicação das melhores pessoas rastreadas a partir da referência para a contagem de saída. Este problema de indicação é instanciado como um grafo bipartido, e uma estratégia gulosa simples é empregada para resolvê-la;

- finalmente, métricas como *recall*, *precision* e *F-score* (Antic et al., 2009) são introduzidas. A partir dos dados de correspondência podemos analisar várias situações específicas de contagem.

As principais vantagens desta metodologia são: quantificar automaticamente a contagem de falsos positivos e negativos dos métodos e também identificar quando estes erros acontecem. Ademais, uma vez que os erros foram identificados, é possível retornar ao método para corrigir suas causas e aprimorar sua precisão.

A fim de validar a nossa metodologia de avaliação, avaliamos o método proposto por Antic et al. (2009) em três vídeos de 10 minutos, gravados nos corredores do Departamento de Computação da Universidade Federal de Ouro Preto. Analisamos também os seus resultados com base nas métricas propostas.

1.1 Organização do Texto

O restante deste trabalho está organizado da seguinte forma. Trabalhos relacionados à métodos de contagem de pessoas baseados em câmera zenital são brevemente discutidos no Capítulo 2. O Capítulo 3 descreve a metodologia de avaliação proposta e o método de contagem de pessoas implementado. A análise dos resultados experimentais é apresentada no Capítulo 4. Finalmente, no Capítulo 5 são apontadas as conclusões e os trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

2.1 Métodos de Segmentação

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 quadro atual e a imagem do plano de fundo, seguido por uma limiarização automática (Velipasalar et al., 2006; Snidaro et al., 2005). Se a precisão que essa abordagem oferece não é garantida (Li et al., 2008; Bescos et al., 2003), uma estratégia por blocos é preferível pois produz uma segmentação mais estável na presença de mudanças de luz e sombra.

Rossi e Bozzoli (1994) 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 (2003) 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. (2006) 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 C. Beleznai e Horst (2006), também empregam o procedimento *mean-shift* para desenvolver um sistema de rastreamento de pessoas geral.

2.2 Metodologias de Avaliação Existentes

Na literatura, a principal forma encontrada para a avaliação dos métodos de contagem por vídeo é a precisão, um percentual entre o número real de pessoas e o número de pessoas contadas (Huang e Chow, 2003; Velipasalar et al., 2006; Septian et al., 2006; Hsieh et al., 2007; Yu et al., 2007a). Embora seja a principal informação desejada sobre o método, não permite descobrir onde ocorrem os erros de contagem. Muito menos, não é possível concluir se há falsos positivos ou negativos na contagem. Isto permite que um método obtenha uma precisão elevada, porém sem detectar as pessoas que realmente passaram pela cena. Portanto,

o resultado do método poderá variar com relação a diferentes cenários.

Outros autores, como Bescos et al. (2003); Xu et al. (2007), incluíram aos resultados de contagem números de falsos positivos, falsos negativos e verdadeiros positivos. Barandiaran et al. (2008); Antic et al. (2009) aplicaram o uso de métricas como *precision*, *recall*, e *F-Score*, que são definidos com base nestes dados anteriores. Tais valores permitem relacionar a precisão do método ao número real de pessoas que atravessou a zona de contagem. Entretanto, não permite detectar em qual momento ocorreu.

Já Chen et al. (2009) apresenta seus resultados analisados com relação a precisão, com detalhes sobre velocidade e direção de movimento das pessoas, separados pelo número de pessoas na cena. Entretanto, tais valores não são gerados de forma automática pelo método. Há comparação de resultados com quatro outros métodos de contagem por vídeo.

2.3 Caracterização dos Métodos

Nesta seção caracterizamos os métodos segundo o *hardware*, configurações, entre outros, realizada por Mendes e Menotti (2011).

Tabela 2.1: Caracterização dos métodos da literatura com relação aos parâmetros utilizados

Método	**Parâmetros de configuração**
Velipasalar et al. (2006)	Processamento a 33fps.
Septian et al. (2006)	Webcam montada a 2.3 m acima do solo. Processamento a 3 fps. Cita, sem detalhar, uso de filtro de passa-baixa para a remoção de ruídos.
Xu et al. (2007)	Processamento a 8fps. Imagens em tons de cinza. Câmera montada a 4m acima do solo. Imagens com 640x480 pixels.
Hsieh et al. (2007)	Sem informações sobre a montagem.
Yu et al. (2007b)	Imagem com resolução de 160x120. Processamento a 10 fps.
Bozzoli e Cinque (2007)	Câmera montada no teto da estação (não cita a distância relativa ao solo).
Yu et al. (2008)	Windows, Pentium IV 2.0 Ghz.
Ying Xin (2008)	Sem informações sobre a montagem e sobre a câmera
Chao-Ho (Thou-Ho)	Câmera de vídeo colorida, montada a 245 cm acima do solo. Tamanho de 320x240 pixels e taxa de 30fps. Processamento em um Intel Core2Duo T5500.
Yahiaoui et al. (2008)	Câmera stereo.
Jaijing et al. (2009)	Imagens de 192x144 pixels. Câmera montada a 2.5 m acima do solo. Área capturada de 2.1 x 3.8 m, em ambiente interno iluminado com lâmpadas fluorescentes. Processamento feito a 7.5 fps.
Antic et al. (2009)	Camera montada a 3m acima do solo em ambiente fechado com iluminação por lâmpadas fluorescentes e a luz do dia. Codigo em Matlab e posteriormente portado para C++. Performance de 12 fps em um PC com 3 GHz.
Huang e Chow (2003)	Intel Pentium IV 1.3 GhZ com 256Mb de RAM. Matlab e Visual C++ 6.0. Separada uma região de interesse com 300x90 pixels. Remocao de ruídos (durante o processamento)
Bescos et al. (2003)	Câmera a 3.8m de altura.
Barandiaran et al. (2008)	Usado em um PC comum, câmera de circuito interno de TV. Distancia focal de 3.5mm e imagens com 352x288 pixels
Chen et al. (2009)	Câmera montada a 4.2m acima do solo, videos com 320x240 pixels capturados a 30fps.

Capítulo 3

Desenvolvimento

Neste capítulo, descrevemos na Seção 3.1 a nossa metodologia de avaliação. Na Seção 3.2, abordamos o método de contagem de pessoas por câmera em posição zenital, proposto por Antic et al. (2009), que implementamos para validar nossa metodologia.

3.1 A Metodologia de Avaliação

Como referido anteriormente, a metodologia proposta pode ser dividida em três etapas principais:

1. a geração da *ground-truth*;
2. a indicação das melhores pessoas rastreadas a partir de referência com relação a saída do método de contagem;
3. calcular as métricas segundo estas indicações.

Cada uma destas etapas serão particularizadas nas subseções seguintes.

3.1.1 Geração da Referência

A referência é gerada através da análise de todo o vídeo. As anotações são feitas apenas para os quadros onde são detectados eventos. Um evento é considerado como uma pessoa entrando ou saindo da zona de contagem ou região de interesse (*ROI*). Uma vez que um evento acontece, ele é inserido em uma tabela de referência da seguinte forma. Se o evento consiste de uma pessoa entrando na *ROI*, nós adicionamos o seu identificador numérico único com um sinal positivo na coluna de direção correspondente, *i.e.*, **Up** ou **Down**. Caso contrário, se consiste em alguém saindo da *ROI*, a mesma identificação que foi criada para esta pessoa ao entrar na ROI é utilizada, porém com um sinal negativo e na coluna de direção correspondente. Note-se que esta convenção *Up* e *Down* pode ser adaptada para vídeos onde as pessoas vêm e

Tabela 3.1: Exemplo de geração da referência

Frame	**Up**	**Down**
1004	0	1
1019	2	0
1083	-1	0
1113	0	-2
2058	3	0
2067	4	0
2114	-3	0
2150	0	-4

vão para esquerda e direita. Também observamos que duas ou mais pessoas podem entrar ou sair da ROI no mesmo quadro, sem perda de generalidade da nossa convenção. Um exemplo de geração de referência para 4 pessoas é mostrado na Tabela 3.1.

Na Tabela 3.1, a pessoa cujo identificador corresponde ao "1", entrou pela parte inferior da zona de contagem no quadro 1004, e saiu pela parte superior da zona de contagem no quadro 1083. Um caso particular do sistema é representada pela pessoa "3", que adentra a ROI por cima e retorna por cima. Este caso representa um movimento *bidirecional*, pois primeiro ela seguia para baixo, depois mudou a rota voltando para cima.

3.1.2 Problema de Indicação

A indicação das melhores pessoas rastreadas a partir da referência com relação à contagem de saída funciona do seguinte modo. Da referência e saída contagem de dados, extrai-se uma simples representação. cada pessoa rastreada é representado como um composto triplo de sua ID e o número do quadro onde ele entra e sai do ROI. Juntamente a estes quadros, colocamos as informações de direção. A i-ésima pessoa rastreada dos dados de referência pode ser instanciada como $(RID^i, RF^i_{in}, RF^i_{out})$, enquanto a j-ésimo pessoa rastreada pelo método, como $(MID^j, MF^j_{in}, MF^k_{out})$. Observe que a saída do método de contagem das pessoas avalidado pode respeitar as mesmas regras e convenções das impostos para a geração da referência, mas a cardinalidade destes dois conjuntos pode ser diferente. Um exemplo que ilustra de tal transformação na representação de os dados de referência na Tabela 3.1 é mostrado na Tabela 3.2.

Cada conjunto de pessoas rastreadas para a referência R e o método M podem ser vistos como conjuntos disjuntos, onde o peso de conexão entre o elementos desses conjuntos são proporcionais às suas sobreposição na domínio do tempo, e o problema de indicação de pessoas rastreadas a partir da referência para o método se torna um problema de grafos. Para calcular as bordas de peso, W_{ij}, entre RID^i e MID^i propomos a calcular a sua intersecção e união de

Tabela 3.2: Exemplo de vértices da tabela de referênciaExample of reference vertices da Tabela 3.1 usada para a indicação

ID	$Frame_{in}$	$Frame_{out}$
1	-1004	+1083
2	+1019	-1113
3	+2058	+2114
4	+2067	-2150

intervalos de tempo, emph i.e.,

$$W^{ij} = \frac{|RInt^i \cap MInt^j|}{|RInt^i \cup MInt^j|} \tag{3.1}$$

onde $RInt^i$ e $MInt^j$ correspondem ao intervalo de tempo de RID^i e MID^j das pessoas rastreadas, respectivamente, e $0 \leq W^{ij} \leq 1$. Quando a direção do movimento das pessoas rastreadas para RID^i e MID^j são diferentes de zero, é atribuído a suas arestas seu respectivo peso. A partir desta definição, temos que quanto maior o intervalo de tempo entre a intersecção RID^i e MID^j, maior é o peso da aresta. Em contraste, quanto maior é a união dos intervalos de tempo entre RID^i e MID^j, menor é o peso da aresta. A Figura 3.1 ilustra um gráfico resultante desta procedimento, onde as arestas mais escuras representam um peso maior, enquanto as arestas mais claras representam um peso menor.

À primeira vista, podemos pensar que o nosso problema é a indicação direta relacionada com a correspondência máxima em grafos bipartidos ou a máxima correspondência em um fluxo líquido Bondy e Murty (1976). além de ter complexidade de tempo exponencial e pertencente a classe NP-completo Garey e Johnson (1979), esses problemas lidam com uma solução global. Além disso, os "bons" algoritmos disponíveis na literatura para resolver estes problemas podem obter soluções que não saturam todos os vértices da referência. Em nosso caso, preferimos uma solução que obtivesse o maior número de indicações dos vértices de referência para os de saída método independente do valor obtido para todo o grafo, *i.e.*, a solução global.

Assim, propomos um algoritmo guloso para selecionar a indicação das melhores pessoas rastreadas da referência à contagem de saída. Depois calcular o peso das arestas para todos os pares possíveis de RID^i e MID^j para $0 < i < |RID|$ and $0 < j < |MID|$, onde $|RID|$ e $|MID|$ representa a cardinalidade de RID e MID, respectivamente, podemos classificar as arestas em ordem decrescente de peso. Uma vez classificados, remove-se a partir deste conjunto as arestas W^{ij} com o maior peso e se ambos os seus vértices não correspondem a uma indicação, a indicação entre eles é estabelecida. Este processo é repetido até que não haja mais vértices livres de indicações em R ou em M ou até que não haja mais arestas para analisar.

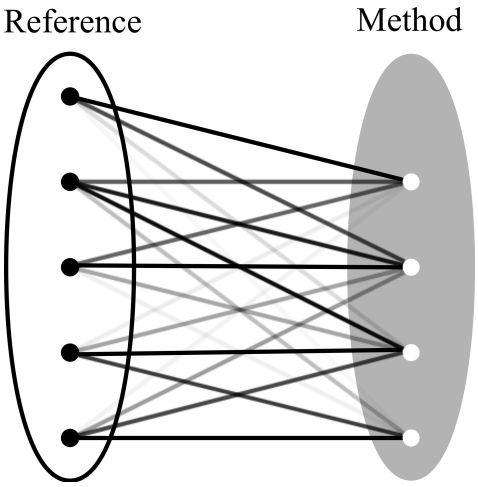


Figura 3.1: Grafo bipartido representando o problema de indicação. Cada aresta possui um peso baseado na Equação 3.1. Arestas mais escuras possuem peso maior; arestas mais claras possuem menor peso.

3.1.3 Métricas

Ao final do processo descrito na subseção anterior, todos as possíveis indicações a partir de R e M foram realizadas. A não indicação dos vértices dos conjuntos R e M são diretamente considerados como falsos negativos (FN) e positivos (FP), respectivamente, na contagem de avaliação.

A partir do número de vértices indicados a partir de R para M (verdadeiros positivos), o número de vértices não indicado em R (FN) e em M (FP), podemos definir três métricas, *i.e.*, *precision*, *recall*, e *F-score*.

$$precision = \frac{TP}{TP + FP}, \tag{3.2}$$

$$recall = \frac{TP}{TP + FN}, \tag{3.3}$$

e

$$F - score = \frac{2 \times precision \times recall}{precision + recall}. \tag{3.4}$$

Como já afirmado, esta não é a primeira vez na literatura de contagem de pessoas com base em métodos de análise de vídeo que as métricas como as definidas são utilizados para avaliar os resultados de contagem. No entanto, devido a nossa metodologia de avaliação, é possível determinar automaticamente em quais situações, segundo o número de pessoas presentes na zona de contagem, estes erros acontecem.

A fim de calcular automaticamente estas situações a partir dos dados referenciais, em primeiro lugar temos que calcular o número esperado de pessoas na zona de contagem em cada quadro. Esta informação pode ser facilmente estimada acumulando o tempo de intervalo de cada pessoa controladas em um total vetor de tempo. Uma vez que este vetor de tempo

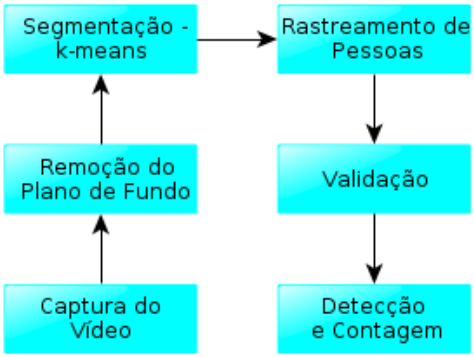


Figura 3.2: Fluxograma para representação do sistema

total é calculado, podemos obter o número esperado de pessoas para cada pessoa rastreada na referência de dados e na saída do método, tendo a frequência máxima no vetor de tempo total para o seu intervalo de tempo próprio. Como os conjuntos FN e FP são conhecidos, para a construção do histograma de situações é suficiente acumular cada erro na sua posição correspondente do histograma. Semelhante processo pode ser executado para obter a situações onde as pessoas TPs rastreadas acontecem.

3.2 O Método de Contagem de Pessoas

Nesta seção descrevemos em maiores detalhes o método para contagem de pessoas que utilizamos para validar nossa metodologia de avaliação. Ele é dividido em: captura do vídeo, subtração do fundo, segmentação, rastreamento e contagem de pessoas (Figura 3.2). As operações nos quadros 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.

A primeira etapa 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 quadro atual com os blocos do quadro pertencente ao fundo. As imagens ou quadros que pertencem ao fundo do vídeo são obtidas através do seguinte filtro

$$F^{t+1} = (1 - \alpha) \cdot F^t + \alpha \cdot I^t \tag{3.5}$$

onde F e I representam, respectivamente, os quadros de fundo e os quadros do vídeo original; t é o número do quadro; 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 do vídeo. Optamos por 0.01 pois os ruídos são reduzidos em nossos vídeos. O filtro é aplicado sobre todos os quadros e todos os seus canais RGB.

O algoritmo utiliza fatores multiplicativos $\beta^t_{m,n,p}$, determinados através da 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).

$$\beta^t_{m,n,p} = \frac{\sum I^t_{m,n,p} \cdot F^t_{m,n,p}}{\sum (F^t_{m,n,p})^2} \tag{3.6}$$

A detecção de pessoas nos quadros 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 armazenada em $\delta\beta^t$, para cada quadro.

$$\delta\beta^t = \max_p \beta^t_{m,n,p} - \min_p \beta^t_{m,n,p} \tag{3.7}$$

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.

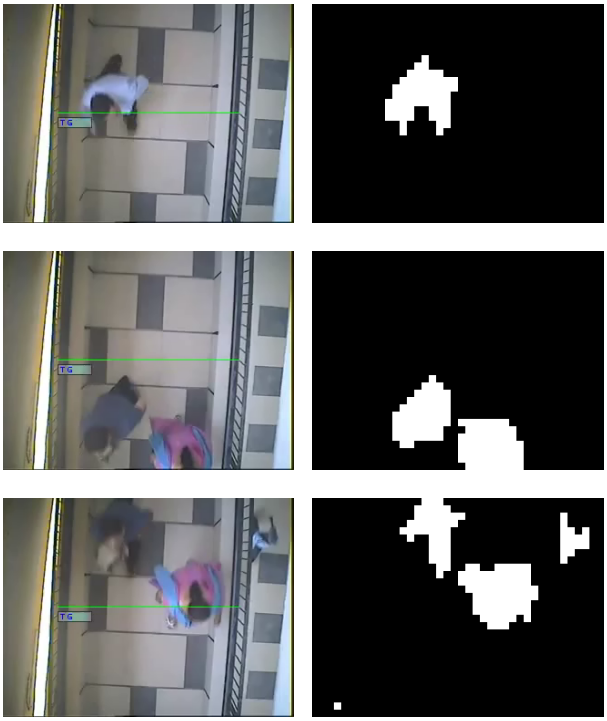
$$P^t = \begin{cases} 1, \text{se } \delta\beta^t > T_1 \vee |\beta^t_{m,n,p}| > T_2 \\ 0, \text{caso contrário} \end{cases} \tag{3.8}$$

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. Observamos que, para nossos vídeos, o melhor ajuste destes parâmetros consistiu em $T_1 = 0.2$ e $T_2 = 0.6$, por reduzir o ruído.

Nesse momento, há uma imagem P para cada quadro, as quais contém apenas as pessoas. O próximo passo do algoritmo é a segmentação destas pessoas. A segmentação é um problema difícil em Análise de Imagem, devido às várias características que representam uma pessoa. Como nos vídeos em questão aparece apenas a parte superior das pessoas, esse problema é reduzido. Assim as pessoas passam a ser vistas como formas geométricas (Figura 3.3), 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.

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



(a) (b)

Figura 3.3: Resultados demonstrativos das etapas de subtração do fundo e segmentação de pessoas. (a) quadro original. (b) Subtração do fundo e segmentação de pessoas.

$$d^k_{min} = \min_{1 \leq i < j \leq k} ||C_i - C_j|| \tag{3.9}$$

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} .

$$k* = \max\{k | d^k_{min} \geq D_{min} \wedge d^{k+1}_{min} < D_{min}\} \tag{3.10}$$

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 com a posição dos centróides encontrados na iteração anterior. Dessa forma os centróides são inicializados com uma posição mais próxima de ser a melhor para os *clusters*, pois o deslocamento de uma pessoa no vídeo é pequeno.

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

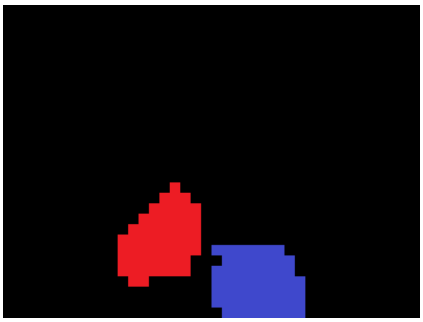


Figura 3.4: Após segmentação pelo *k-means*

O algoritmo encontra os *clusters* correspondentes em dois quadros consecutivos que possuem a menor distância. O objetivo é obter a menor distância Euclidiana quadrada entre os *clusters*. Os *clusters* cuja distância é a menor são considerados correspondentes a uma mesma pessoa. Deste modo, armazenamos a rota descrita por uma pessoa, *i.e*, o centróide de cada *cluster* correspondente a uma mesma pessoa é armazenado.

Após obter a rota de cada pessoa pela zona de contagem, que consiste de n pares (x, y), efetuamos a contagem sobre aqueles com um tamanho mínimo percorrido. Todas as rotas que possuírem tamanho suficiente, serão contabilizadas pelo algoritmo. Este procedimento evita que ruído ou má segmentação interfira na contagem, pois geralmente são menores que o tamanho mínimo necessário. Este tamanho consiste na metade da largura da zona de contagem, segundo Antic et al. (2009).

A Figura 3.3 demonstra os principais passos do algoritmo. As imagens da primeira coluna mostram os quadros 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.

A Figura 3.4 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.

3.2.1 Ajuste de Parâmetros

A Tabela 3.3 apresenta os parâmetros finais utilizados para este método em todos os três vídeos. O parâmetro que requer maior atenção é o tamanho mínimo de uma pessoa (D_{min}), sendo o principal responsável por segmentações incorretas. As linhas superior e inferior representam a delimitação da zona de contagem no vídeo, utilizada no processo de contagem pelo método.

Tabela 3.3: Valor dos parâmetros utilizados

Parâmetro	Valor
α	0.01
Bloco	8x8
T_1	0.2
T_2	0.6
D_{min}	50px
Linha superior	30
Linha inferior	450

Capítulo 4

Experimentos

Os vídeos utilizados nos experimentos têm resolução de 640x480, 30*fps* e 10 minutos. Todos os valores apresentados nesta seção foram gerados automaticamente por nossa metodologia.

A Tabela 4.1 apresenta, para os três vídeos, um resumo com a contagem esperada, a contagem realizada pelo método e o número de indicações feitas pelo algoritmo da metodologia. Para o primeiro vídeo, o método contou duas pessoas a mais do que o esperado. Além disso, o número de indicações pela nossa metodologia de avaliação mostra que 4 pessoas rastreadas pelo método não correspondem a qualquer rota traçada por uma pessoa, descrita na referência. Assim, o método contou quatro objetos que não correspondem a qualquer pessoa - *i.e.*, falsos positivos.

Para o segundo e o terceiro vídeo, a contagem realizada pelo método foi além do valor real. No entanto, o número de correspondências entre a metodologia de referência e a saída do método foram muito próximos do valor esperado para ambos os vídeos. Apenas *stm3* possui uma rota sem correspondência, enquanto em *stm2* houve a correspondência entre todos. De acordo com a nosso metodologia de avaliação, apenas o número de rotas indicadas entre referência e saída do método eram pessoas reais andando pela zona de contagem. Portanto, havia 25 falsos positivos nos vídeos *stm2* e *stm3*.

A Tabela 4.2 apresenta em maiores detalhes o número de falsos negativos (FN) e falsos positivos (FP) para diferentes números de pessoas em todo o vídeo. A coluna com o número de pessoas 1 mostra a soma cumulativa das pessoas erroneamente contadas quando havia

Tabela 4.1: Tabela com o número real de pessoas em todo o vídeo, a contagem realizada pelo método e o número de indicações

Vídeo	Esperado	Realizado	Indicações
stm1	33	35	29
stm2	16	41	16
stm3	22	46	21

Tabela 4.2: Número de falsos negativos (FN) e falsos positivos (FP) em quadros com diferentes números de pessoas

Vídeo		Número de Pessoas			
		0	1	2	3
stm1	**FN**	x	1	2	1
	FP	0	4	2	0
stm2	**FN**	x	0	0	0
	FP	5	20	0	0
stm3	**FN**	x	0	1	0
	FP	0	21	4	0

apenas uma pessoa andando pela zona de contagem. O valor total da FP em *stm2* e *stm3* é exatamente 25, como dito antes. Em quadros com 0 pessoas, não é possível ter qualquer FN.

Para o vídeo de entrada *stm1*, o método não encontrou as pessoas quando os quadros eram vazios, indicando que o método não identificou ruídos como uma pessoa. Quando 1 ou 2 pessoas estavam pela zona de contagem, o método contou equivocadamente algumas pessoas a mais. Portanto, o método segmentou de forma incorreta uma pessoa como duas ou mais, devido à parâmetros da aplicação.

Para a entrada *stm2*, o método considerou o ruído como uma pessoa, como mostrado na coluna com 0 pessoas. Além disso, ele segmentou incorretamente uma pessoa como mais de uma, erro repetido para a entrada de *stm3*.

Quando havia pessoas reais nos quadros, apenas *stm1* e *stm3* não contaram algumas das pessoas passando pelo zona de contagem (FN).

Finalmente, a Tabela 4.3 apresenta as métricas de *precision*, *recall* e *F-score*. Apenas em *stm1* o método obteve mais que 50% em *precision*, pois foi o único vídeo que chegou mais próximo da contagem correta, registrando poucos falsos positivos. Os vídeos *stm2* e *stm3* contaram muitos falsos positivos e por este motivo obtiveram um valor inferior ao vídeo *stm1*.

Por outro lado, em *recall* o método apresentou mais de 85% para todos vídeos. Isso mostra que, para todas as entradas, o método contou quase todas as pessoas que cruzaram a zona de contagem, com poucos ou nenhum falso negativo. O vídeo *stm2* obteve 100% nesta métrica, indicando que não deixou de contar nenhuma das pessoas que cruzaram a zona de contagem.

Tabela 4.3: *Recall*, *precision* e *F-score* calculado automaticamente

Video	stm1	stm2	stm3
precision	82.9	39.0	45.7
recall	87.9	100.0	95.5
$F - score$	85.3	56.1	61.8

Analisando a métrica *F-score*, que pode ser interpretada como uma média das duas outras

métricas, apenas *stm1* atingiu mais de 80%. Para os outros dois vídeos, o método precisa de uma correção nos parâmetros para que as pessoas sejam segmentadas corretamente e, assim, contar apenas o número de pessoas no quadro.

Capítulo 5

Conclusões

Neste trabalho, propomos uma metodologia para avaliar automaticamente métodos de contagem de pessoas. Por nossa metodologia de avaliação é possível quantificar automaticamente a contagem de falsos positivos e negativos e também identificar quando estes erros ocorrem.

Os resultados experimentais mostram que apenas para o vídeo *stm1* o resultado da métrica *F-score* foi superior a 85%, indicando que o método não deixou de contar muitas pessoas que cruzaram a zona de contagem, muito menos contabilizou falsos positivos em excesso. Portanto, a contagem se aproximou número real de pessoas que cruzou a zona de contagem.

Para os outros dois vídeos, os resultados para a métrica *F-score* foram 56.1% e 61.8%, para os vídeos *stm2* e *stm3* respectivamente, valores baixos com relação ao resultado do primeiro vídeo. Pela nossa metodologia de avaliação é possível determinar que o motivo destes valores é a segmentação errônea das pessoas, pois o método contabilizou falsos positivos em excesso.

A partir desta informação, é possível determinar que o parâmetro que deve ser reajustado é o D_{min}, o tamanho mínimo de uma pessoa, de modo a corrigir o processo de segmentação pelo *k-means* e aprimorar os resultados.

No entanto, usando nossa metodologia não é possível avaliar situações onde as pessoas transitam separadamente na zona de contagem ou se ocorre algum caso de *merge-splitting*, uma vez que não leva em conta as posições reais das pessoas durante sua passagem.

5.1 Trabalhos Futuros

Propomos como futuras direções de trabalho tentar superar a desvantagem de não poder detectar pessoas que transitam separadamente ou casos de *merge-splitting* da nossa metodologia. Uma solução seria efetuar a segmentação perfeita de cada pessoa durante a sua passagem através da zona de contagem ou indicar o centro de massa durante a sua passagem.

Pretendemos concluir a implementação de outros dois métodos de contagem de pessoas (Barandiaran et al., 2008; Yu et al., 2008) e também temos o intuito de implementar outros três (Chen et al., 2009), a fim de realizar uma avaliação extensa dos métodos usando a

metodologia aqui proposta.

Além dos métodos, pretendemos deixar disponíveis uma maior quantidade de vídeos, com maior tempo de duração e coletados em diferentes locais, usando câmeras e condições diferentes de iluminação. Além disso, é necessário coletar vídeos variando o número de pessoas na zona de contagem. Isto é, desde os momentos com alto tráfego, onde podemos ter facilmente cinco ou mais pessoas na zona de contagem, até os momentos de baixa atividade. De tais vídeos, acreditamos que uma avaliação mais realista dos métodos propostos na literatura pode ser realizada.

Referências Bibliográficas

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

Barandiaran, J.; Murguia, B. e Boto, F. (2008). Real-time people counting using multiple lines. In *International Workshop on Image Analysis for Multimedia Interactive Services (IAMIS)*, pp. 159–162.

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

Bondy, J. A. e Murty, U. S. R. (1976). *Graph Theory With Applications*. Elsevier Science Ltd. ISBN-13: 978-0444194510.

Bozzoli, M. e Cinque, L. (2007). A statistical method for people counting in crowded environments. In *IEEE International Conference on Image Analysis and Processing (ICIAP)*, pp. 506–511.

C. Beleznai, B. F. e Horst, B. (2006). Human tracking by fast mean shift mode seeking. *Journal of Multimedia*, 1(1):1–8.

Chao-Ho (Thou-Ho) Chen, Yin-Chan Chang, T.-Y. C. e Wang, D.-J. (2008). People counting system for getting in/out of a bus based on video processing. *ISDA*, pp. 565–569.

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

Chen, T.-Y.; Chen, T.-H. e Wang, D.-J. (2009). A cost-effective people-counter for passing through a gate based on image processing. *International Journal of Innovative Computing, Information and Control (ICIC International)*, 5(3):785–800.

Committee, D. A. (2009). *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age*. National Academy Press.

da Silva, L. S. (2008). Sistema computacional para contagem automática de pessoas baseado em análise de sequências de imagens. Mestrado, Centro Federal de Educação Tecnológica de Minas Gerais.

Duda, R. O.; Hart, P. E. e Stork, D. G. (2000). *Pattern Classification.* Wiley-Interscience, 2 edição.

Elik, H.; Hanjalic, A. e Hendriks, E. (2006). Towards a robust solution to people counting. In *IEEE International Conference on Image Processing (ICIP)*, pp. 2401–2404.

Garey, M. R. e Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman and Co. ISBN-13: 978-0716710455.

Hsieh, J.-W.; Peng, C.-S. e Fan, K.-C. (2007). Grid-based template matching for people counting. In *IEEE Workshop on Multimedia Signal Processing (MMSP)*, pp. 316–319.

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

Jaijing, K.; Kaewtrakulpong, P. e Siddhichai, S. (2009). Object detection and modeling algorithm for automatic visual people counting system. In *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, volume 2, pp. 1062–1065.

Kilambi, P.; Ribnick, E.; Joshi, A. J.; Masoud, O. e Papanikolopoulos, N. (2008). Estimating pedestrian counts in groups. *Computer Vision and Image Understading*, 110(1):43–59.

Li, M.; Zhang, Z.; Huang, K. e Tan, T. (2008). Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *IEEE International Conference on Pattern Recognition (ICPR)*, pp. 1–4.

Mendes, J. C. e Menotti, D. (2011). Uma análise de precisão e tempo de processamento em métodos de contagem de pessoas por vídeo. In *Seminário de Pós-Graduação em Ciência da Computação da Universidade Federal de Ouro Preto (PPGCC-UFOP).*

Rossi, M. e Bozzoli, A. (1994). Tracking and counting moving people. In *IEEE International Conference on Image Processing (ICIP)*, pp. 212–216.

Septian, H.; Tao, J. e Tan, Y.-P. (2006). People counting by video segmentation and tracking. In *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1–4.

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

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

Xu, X.-W.; Wang, Z.-Y.; Liang, Y.-H. e Zhang, Y.-Q. (2007). A rapid method for passing people counting in monocular video sequences. In *IEEE International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 3, pp. 1657–1662.

Yahiaoui, T.; Meurie, C.; Khoudour, L. e Cabestaing, F. (2008). A people counting system based on dense and close stereovision. In *International Conference on Image and Signal Processing (ICISP)*, volume 5099 (LNCS), pp. 59–66.

Ying Xin, Guangmin Sun, Q. W. (2008). A preprocessing method for tracking and counting pedestrians in bus video monitor. *INDIN*, pp. 1689–1693.

Yu, H.; He, Z. e Liu, J. (2007a). A vision-based method to estimate passenger flow in bus. In *IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 654–657.

Yu, H.; He, Z. e Liu, J. (2007b). A vision-based method to estimate passenger flow in bus. In *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*.

Yu, S.; Chen, X.; Sun, W. e Xie, D. (2008). A robust method for detecting and counting people. In *IEEE International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 1545–1549.