Segmentação de Imagens Aéreas em Ambientes Internos Utilizando Visão Estéreo

Vinicius Graciano Santos
DCC - UFMG
Belo Horizonte - MG
vgs@dcc.ufmg.br

***Resumo*—Este trabalho trata o problema da localização de objetos em imagens aéreas através do uso de visão estéreo e segmentação. O algoritmo desenvolvido calcula o mapa de disparidade após a aplicação de um filtro que remove regiões com baixo nível de textura, que podem gerar erros no processo de correlação. O mapa de disparidade é tratado para que apenas os objetos com as maiores alturas sejam identificados.**

Foram feitos experimentos em imagens retiradas de um ambiente controlado e os resultados mostraram que a tanto a localização quanto a forma dos objetos são encontradas de forma aproximada, assim um pós-processamento seria necessário para ele ser utilizado em aplicações reais.

I. Introdução

O problema da localização de objetos em imagens aéreas é um tema muito pesquisado na área de visão computacional. Suas aplicações estendem-se pelas mais diversas áreas, como cartografia, planejamento militar e urbano, computação gráfica e realidade virtual [1]. Neste trabalho, esse problema é tratado restrito a ambientes internos, desejando-se localizar apenas os objetos mais proeminentes do solo, ou seja, objetos que não se destacam na cena com relação a sua altura devem ser filtrados.

A localização de objetos nesses ambientes pode ser útil para aplicações comerciais e experimentais. Por exemplo, na contagem de veículos de grande porte que passam por uma ponte, na determinação do fluxo de pessoas que acessam uma determinada área ou então para a geração automática de um mapa de navegação para um robô móvel.

A abordagem escolhida para o problema utiliza dois conceitos muito trabalhados nas áreas de processamento digital de imagens e visão computacional: segmentação e visão estéreo.

Por definição, a segmentação de imagens é o problema que lida com a composição da imagem em regiões que são semânticamente uniformes. Entretanto, como as próprias imagens conseguem prover pouca informação semântica, normalmente os algoritmos desenvolvidos para esse propósito necessitam de intervenção humana ou são específicos para algum tipo de aplicação [2].

Visão estéreo refere-se a possibilidade de inferir informações sobre a estrutura 3-D e a profundidade da cena a partir de duas ou mais imagens retiradas de diferentes pontos de vista [3]. Considerando-se um par de câmeras, podemos dizer que os dois problemas principais da visão estéreo são a correspondência, que consiste em determinar qual é o pixel correspondente de uma imagem na outra; e a reconstrução, que precisa inferir a estrutura 3-D a partir das correspondências entre pixels. A diferença da posição de pontos correspondentes, cada qual em relação a sua respectiva imagem, é denominada disparide e a partir dessa informação é possível inferir certas informações sobre a profundidade da cena.

II. Trabalhos Relacionados

Grande parte da literatura trata o problema considerando-se imagens externas e tais algoritmos podem ser alterados ou utilizados diretamente em ambientes internos.

Um algoritmo de detecção e reconstrução de prédios utilizando-se visão estéreo é proposto em [1]. O método utiliza um mapa de disparidade para segmentar os objetos de interesse e uma seleção baseada em hipóteses através da localização de ciclos muito próximos espacialmente em um grafo não-direcionado formado pelas quinas e junções dos objetos para inferir quais deles são prédios ou estruturas mais elevadas de tamanho significante. Os experimentos utilizados mostraram que o algoritmo é eficiente e robusto em cenas urbanas, onde as faces e bordas das estruturas possuem um caráter retilíneo.

Em [2] é apresentada uma estensão do algoritmo de *watershed* para efetuar a segmentação de imagens aéreas em baixas altitudes. O método introduz o conceito de similaridades de cores das regiões como mais um critério no processo de junção (*merging*) do algoritmo. Os experimentos realizados comparam o método com os seus antecessores e mostram que boa parte dos problemas de *over-segmentation* foram eliminados. Apesar de não abordar diretamente o mesmo problema deste trabalho, a segmentação pode ser utilizada como uma base de entrada para um algoritmo que tenha por finalidade identificar os que estão mais proeminentes do solo.

O trabalho apresentado em [4] trata o problema de gerar um mapa de navegação para um robô móvel através de uma única imagem aérea em um ambiente interno e controlado. É mostrado que detectores de bordas não são robustos para a localização, já que a maioria dos ambientes internos possuem padrões repetidos de textura no chão, que não são eliminados por esse processo. Como o ambiente é controlado, os tons de cinza dos obstáculos foram definidos

de tal forma que fosse possível uma segmentação com base nos pontos de máximo do histograma da imagem. A proposta é muito limitada vendo-se que restringe o tipo de objeto que pode ser encontrado, mas o trabalho mostrou que um sistema específico para obter automaticamente um mapa de navegação nesse tipo de ambiente pode ser muito útil para experimentações na área de robótica móvel.

III. Metodologia

Como já exposto, o objetivo do trabalho é identificar a posição de objetos proeminentes do solo em imagens aéreas retiradas em ambientes internos. Para isso, será utilizado um sistema estéreo calibrado e será assumido o modelo de reflectância Lambertiana para as superfícies desses objetos. Também, serão levantadas as seguintes restrições:

- Todos os objetos a serem mapeados devem estar totalmente no campo de visão de ambas as câmeras.
- A porcentagem da área do solo ocupada pelos objetos não pode ser muito significativa.
- A altura média dos objetos da cena não pode ser muito maior do que a distância entre as duas câmeras.

A primeira restrição implica que o problema de oclusão não será tratado pelo algoritmo, já a segunda é necessária; pois se uma grande quantidade de objetos estiverem na cena, a estimação do solo pode ser dificultada. Essa questão será explicada com maior detalhes adiante. Quanto a terceira, ela é naturalmente atendida por imagens capturadas em ambientes internos. Um par de entrada que não atende essa restrição seriam fotos retiradas de grandes altitudes a partir de um avião.

A. Calibração do Sistema Estéreo e Retificação das Imagens

A calibração do sistema é feita anteriormente a execução do algoritmo, sendo feita uma única vez. O ato de calibrar uma câmera consiste em encontrar os parâmetros necessários da transformação que mapeia pontos em um referencial da cena para pixels na imagem. Os parâmetros intrínsecos regem a transformação de um ponto no referencial da câmera para os pixels de uma imagem enquanto que os extrínsecos mapeiam pontos da cena para o referencial da câmera. No caso de um sistema estéreo, os parâmetros extrínsecos são a matriz de rotação $\mathbf{R}$ e o vetor de translação $\mathbf{T}$ que mapeiam pontos no referencial da câmera da direita para o da esquerda.

A geometria epipolar rege o sistema estéreo e define restrições quanto a localização da correspondência de um ponto em uma imagem na outra. Sem entrar em maiores detalhes, a restrição epipolar define que o correspondente de um pixel $\mathbf{p}$ pertence a linha epipolar da outra imagem. A retificação consiste em determinar a transformação de cada imagem tal que pares de linhas epipolares conjugadas tornem-se colineares e paralelas a um dos eixos da imagem. A importância desse processo é evidente no problema da correspondência entre pontos, já que será apenas necessário executar uma busca linear em vez de uma bidimensional nas imagens. Maiores informações sobre a geometria epipolar, retificação e calibração de câmeras podem ser encontradas em [3]. Um par de imagens estéreo é mostrado na Figura 1 e as correspondentes imagens retificadas e transformadas são mostradas na Figura 2. Note que essa transformação faz com que pixels correspondentes pertençam a uma mesma reta horizontal, como indicado na figura.

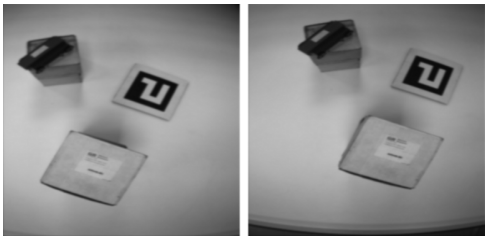


Figura 1. Par estéreo capturado

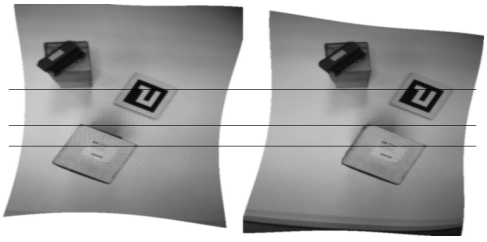


Figura 2. Retificação e warping do par estéreo

B. Segmentação de Regiões com Pouca Textura

Um dos grandes problemas dos algoritmos que tentam solucionar o problema de correspondência são regiões da imagem que possuam um nível de textura baixo, já que esses métodos normalmente se baseiam na comparação de janelas de pixels utilizando alguma fórmula para medir a similaridade entre elas, o que em regiões homogêneas pode acarretar em erros no cálculo do mapa de disparidade. Para evitar esse problema, é desejável que essas áreas sejam segmentadas para não serem consideradas durante essa etapa.

Regiões de baixa textura são caracterizadas por áreas onde o gradiente da imagem seja relativamente pequeno, ou seja, dada uma imagem $\mathbf{I}$ e uma região $\mathbf{R} \subseteq \mathbf{I}$, no caso ideal de uma região homogênea temos:

$$w(\mathbf{p}) = \sum_{p \in R} ||\nabla \mathbf{I}(\mathbf{p})|| = 0.$$

Os componentes de $\nabla \mathbf{I}$ são encontrados convolvendo a imagem com os operadores horizontais e verticais de Sobel [5]. Então, é definido um limiar τ e uma máscara $\mathbf{M}$, com as mesmas dimensões de $\mathbf{I}$, é gerada tal que

$$\mathbf{M}(\mathbf{p}) = \begin{cases} 0, & w(\mathbf{p}) < \tau \\ 1, & w(\mathbf{p}) \geq \tau. \end{cases}$$

O filtro da mediana é aplicado sobre $\mathbf{M}$ para suavizar a máscara e remover pequenas regiões formadas por ruído ou de tamanho insignificante perante a janela de correlação explicada no próximo passo.

O procedimento é executado para ambas as imagens do par estéreo, obtendo-se como saída duas máscaras $\mathbf{M_E}$ e $\mathbf{M_D}$. A Figura 4a mostra $\mathbf{M_e}$ aplicada em sua respectiva imagem fonte.

C. Cálculo do Mapa de Disparidade

A partir da geometria de um sistema estéreo onde os eixos ópticos das câmeras são paralelos é possível mostrar facilmente que existe uma relação entre a profundidade Z de um ponto $\mathbf{P}$ na cena com a diferença entre a as suas respectivas projeções no referencial de cada imagem [3]. Essa diferença é denominada disparidade e dada a distância T e uma distância focal f, com relação as duas cãmeras, a equação que relaciona essas variáveis é dada por:

$$Z = f\frac{T}{d}.$$

O mapa de disparidade consiste em uma imagem que codifica as disparidades de cada par correspondente de pixels. Logo, dado um pixel $\mathbf{p_e}$ na imagem da esquerda, queremos encontrar o seu correspondente $\mathbf{p_d}$ na outra imagem, para obtermos a disparidade. O algoritmo de correlação é muito utilizado para gerar mapas de disparidade densos e sua idéia básica é comparar uma janela de tamanho $2W+1$ centrada em $\mathbf{p_e}$ com várias janelas dentro de uma região $\mathbf{R}(\mathbf{p_e})$ da outra imagem. Assim, devemos computar, para cada pixel $\mathbf{p_e} = [i,j]^T$, o vetor de deslocamento $\mathbf{d} = [d_1, d_2]^T \in \mathbf{R}(\mathbf{p_e})$ que minimize

$$\sum_{k=-W}^{W}\sum_{l=-W}^{W}\Psi(I_E(i+k,j+l), I_D(i+k-d_1, j+l-d_2)).$$

A função $\Psi(u,v)$ define uma medida de comparação entre as janelas e foi definida como $\Psi(u,v) = (u-v)^2$, que é conhecida como SSD (*sum of squared differences*). A escolha dessa fórmula foi devido a grande recomendação da literatura devido a sua maior robustez em regiões com valores de intensidade muito discrepantes.

Como as imagens estão retificadas, não é necessário uma busca bidimensional com relação a $\mathbf{d}$; devido à restrição epipolar. Logo, a ordem de complexidade do algoritmo nesse tipo de imagem é menor do que em sua forma original apresentada.

Como já citado, o algoritmo de correlação pode não retornar casamentos aceitáveis em regiões mais homogêneas da imagem. Dessa forma, a correlação é feita apenas considerando-se os pixels $\mathbf{p}$ tais que $M(p) \neq 0$, para cada imagem e a sua respectiva máscara. A Figura 4b mostra o mapa de disparidade para o par da Figura 2.

D. Filtrando o Mapa de Disparidade

O mapa de disparidade nos indica um valor que relaciona-se com altura dos objetos relativa a câmera. Como desejamos localizar apenas aqueles que estão mais proeminentes do solo, devemos eliminar as disparidades abaixo de um certo limiar. A forma mais direta de encontrar um valor aceitável para o *threshold* é verificando o histograma $H(x)$ do mapa de disparidade. O limiar μ é o valor tal que

$$\mu = \arg\max H(x).$$

Como $H(x)$ não define uma curva suave devido ao ruído na imagem e aos erros do processo de correlação, o filtro da mediana é usado para suavizar o mapa de disparidade, eliminando alguns possíveis *outliers* que podem gerar variações em certos pontos de $H(x)$.

Assim, dado o mapa de disparidade original $\mathbf{D}$, é gerado um novo mapa $\mathbf{D}'$ de acordo com a relação abaixo.

$$\mathbf{D}'(\mathbf{p}) = \begin{cases} 0, & \mathbf{D}(\mathbf{p}) \leq \mu \\ \mathbf{D}(\mathbf{p}), & \mathbf{D}(\mathbf{p}) > \mu \end{cases}$$

Mesmo após esse *threshold*, $\mathbf{D}'$ ainda possui pequenas regiões com alta disparidade que não pertencem aos objetos a serem localizados, novamente, devido aos erros gerados pelos passos anteriores. Por esses motivos, tais regiões devem ser filtradas.

$\mathbf{D}'$ é convertido em uma imagem binária $\mathbf{D'_b}$ de forma que

$$\mathbf{D'_b}(\mathbf{p}) = \begin{cases} 0, & \mathbf{D}'(\mathbf{p}) = 0 \\ 1, & \mathbf{D}'(\mathbf{p}) > 0. \end{cases}$$

O problema de identificar as regiões conexas R_i tais que $\mathbf{D'_b}(\mathbf{p}) = 1, \forall\ \mathbf{p} \in R_i$ e $R_i \cap R_j = \emptyset, \forall\ i,j \in [1,n]$, com $i \neq j$ é conhecido na literatura como *region labelling*. Algoritmos eficientes para resolver esse problema, não apenas limitados a imagens binárias, são descritos em [6] e [7].

Como, a princípio, estamos considerando um único par de imagens em um processamento *offline*, o tempo de execução do algoritmo não é uma das prioridades principais. Dessa maneira, um algoritmo ingênuo foi implementado apenas para a prova de conceito do método desenvolvido. Logo, em uma aplicação com restrições quanto a complexidade de tempo, este passo seria um gargalo no sistema e deve ser otimizado.

O algoritmo é como se segue:

- Encontre o conjunto Φ de contornos contidos em $\mathbf{D'_b}$. Por exemplo, utilizando-se o detector de bordas de Canny [3], [8].
- Defina uma imagem de saída (I_s), com as mesmas dimensões de $\mathbf{D'_b}$. Para cada pixel $\mathbf{p} \in \mathbf{I_s}$, considerando-se um contono $C_i \subseteq \Phi$ e $R(C_i)$, a região dentro de C_i, faça:

$$\mathbf{I_s}(\mathbf{p}) = \begin{cases} i, & p \in R(C_i) \\ 0, & p \notin \bigcup_{i=1}^{n} R(C_i) \end{cases}$$

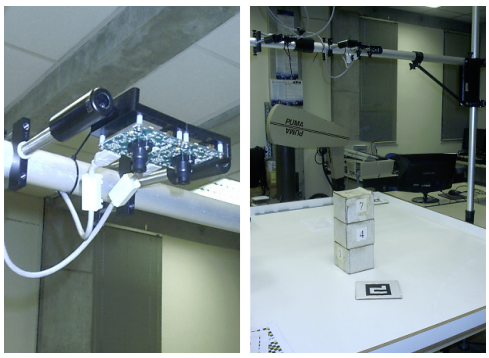


Figura 3. Sistema estéreo utilizado nos experimentos

Tendo-se a classificação das regiões, calcula-se a área de cada uma e aquelas que forem muito menores do que a média são eliminadas. A saída é a imagem $\mathbf{S}$ tal que $\mathbf{S}(\mathbf{p}) = 0$ nas regiões filtradas e $\mathbf{S}(\mathbf{p}) = \mathbf{D}'_{\mathbf{b}}(\mathbf{p})$ no caso contrário. $\mathbf{S}$ irá conter as regiões que aproximam a localização e o formato dos objetos na imagem da esquerda. O processo pode ser repetido para o outro membro do par se for necessário. As figuras 4c e 4d mostram as saídas desta última etapa.

IV. Experimentos

Os experimentos foram feitos utilizando-se um sistema estéreo composto de duas câmeras *Point Grey Research Dragonfly*© com resolução de 640×480, instaladas no *VerLab* do *DCC-UFMG*. As cenas montadas para os testes são composições de alguns objetos simples com alturas distintas localizados sobre uma mesa que representa o solo. O sistema e o exemplo de uma cena podem ser vistos na Figura 3.

Os passos de calibração, retificação e *warping* das imagens foram feitos utilizando-se o *Camera Calibration Toolbox for Matlab*, desenvolvido por Jean-Yves Boughet do *California Institute of Technology*.

O conjunto de testes considerou dois grupos de imagens. O primeiro contém cenas onde existem objetos com uma grande diferença relativa de altura e uma padrão homogêneo de textura no solo, enquanto as cenas do segundo grupo contêm padrões repetidos de textura. Os passos e resultados para dois pares de imagens podem ser vistos na Figura 4 e 5.

Os resultados mostraram que o algoritmo é robusto quanto a eliminação dos objetos menos proeminentes, assim como de padrões de textura contidas no chão. A localização é dada de forma aproximada, ou seja, o resultado são regiões que aproximam a verdadeira forma e posição dos objetos. Também, é importante ressaltar que a segmentação é efetuada comparando-se as alturas relativas desses objetos. Logo, o algoritmo poderá ter problemas com imagens onde a quantidade de objetos seja suficiente para ocupar boa parte da área da imagem, pois devido às comparações relativas, se houver uma grande plano definido pelos objetos, apenas os mais altos a partir desse plano serão identificados. Por esses motivos, o algoritmo fica restrito a cenas que não possuem essa característica e a segunda restrição ao problema foi imposta, como descrito nas seções anteriores.

V. Conclusão

O objetivo deste trabalho era apresentar uma solução ao problema da localização de objetos a partir de imagens aéreas retiradas em ambientes internos. Um algoritmo utilizando-se visão estéreo com um par de câmeras calibradas foi desenvolvido e a segmentação resultou em uma aproximação da posição e da forma dos objetos mais proeminentes do solo. Dessa maneira, um pós-processamento sobre as imagens de saída deve ser feito para a sua utilização em aplicações reais. Por exemplo, utilizando-se o operador morfológico de dilatação [5] podemos expandir as regiões da imagem binária resultante, garantindo maior confiabilidade em uma mapa de navegação. Entretanto, regiões muito próximas pertencentes a diferentes objetos poderiam ser unidas, mas, dependendo da escala do robô com relação a cena, isso poderia não ser um problema.

No geral, a solução proposta foi condizente com as restrições impostas inicialmente e o algoritmo mostrou-se robusto a padrões de texturas contidas no chão. Trabalhos futuros incluem uma implementação mais eficiente do algoritmo, para que ele possa ser utilizado em aplicações que necessitam de processamento *online*, como em uma seqüência de vídeo. Além disso, formas de pós-processamento da saída serão estudadas e aplicadas para aumentar a confiabilidade e robustez dos resultados com relação a posição e a forma dos objetos.

Referências

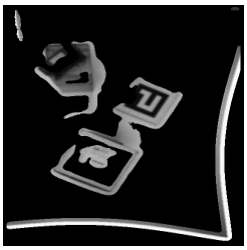
[1] D. min Woo A, Q. dat Nguyen A, Q. dung Nguyen Tran A, D. chul Park A, and Y. kee Jung B, "Building detection and reconstruction from aerial images," in *Proceedings of the Int. Soc. for Photogrammetry and Remote Sensing*, 2008.

[2] A. B. Huguet, M. C. d. Andrade, R. L. Carceroni, and A. d. A. Araujo, "Color-based watershed segmentation of low-altitude aerial images," in *SIBGRAPI '04: Proceedings of the Computer Graphics and Image Processing, XVII Brazilian Symposium*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 138–145.

[3] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

[4] A. Segovia and M. Rombaut, "Path finding from a spot image for a mobile robot," in *Intelligent Vehicles '93 Symposium*, 14-16 1993, pp. 277 –282.

[5] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006.



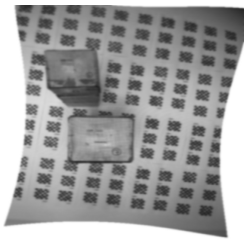
(a) Máscara de regiões texturadas

(b) Mapa de disparidade

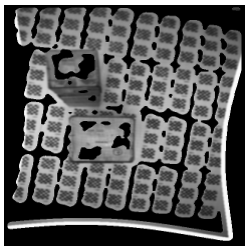
(c) Threshold a partir do histograma

(d) Resultado após eliminação de regiões pequenas

Figura 4. Passos do algoritmo para o par apresentado na Figura 1

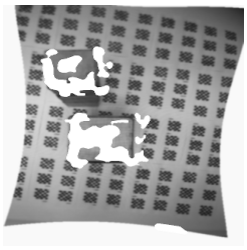


(a) Um dos pares retificados



(b) Máscara de regiões texturadas

(c) Mapa de disparidade



(d) Resultado

Figura 5. Passos do algoritmo sobre uma cena com um padrão repetido no solo

[6] M. B. Dillencourt and H. Samet, "A general approach to connected-component labeling for arbitrary image representations," *Journal of the ACM*, vol. 39, pp. 253–280, 1992.

[7] Z. Lin, J. Jin, and H. Talbot, "Unseeded region growing for 3d image segmentation," in *VIP '00: Selected papers from the Pan-Sydney workshop on Visualisation*. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2001, pp. 31–37.

[8] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.