Sincronização de seqüências de vídeo baseada em rastreamento de pontos de interesse

André Lima Gaspar Ruas.

May 21, 2009

1 Introdução

Na reconstrução de cenas baseada em múltiplas visões é preciso que tenhamos um conjunto de imagens obtidas no mesmo instante de tempo, para reconstruir a cena por triangulação. Isso requer o conhecimento dos instantes de tempo de múltiplas seqüências de vídeo. Em alguns casos essa sincronização pode ser realizada diretamente através de hardware, em outros casos pode ser feita manualmente. Porem em diversos casos não está disponível uma sincronização feita por hardware e a medida que o número de seqüências aumenta se torna mais custoso sincroniza-las manualmente. Além disso em muitos casos, como aplicações de tempo real, é impossível sincronizar manualmente as seqüências. Nesses casos métodos de sincronização automáticos se tornam necessários.

2 Trabalhos Relacionados

2.1 Sincronização de vídeos:

Wolf e Zomet.[**?**] propuseram um algoritmos de sincronização que determina a diferença de tempo através de uma minimização do posto de uma matriz empilhada com dados de rastreamento de duas câmeras. Caspi et al.[**?**] propuseram um método de sincronização baseado em correspondência entre trajetória de objetos. Yan e Pollefeys[**?**] propuseram um método de sincronização automática baseado em correlação entre distribuições de pontos de interesse no tempo e espaço entre seqüências de vídeo e Tresadern e Reid[**?**] desenvolveram um método que estima o offset de tempo entre seqüências de vídeo baseado em rastreamento de objetos não rígidos com acurácia de subquadro.

2.2 Detectores e descritores de pontos de interesse:

Um dos primeiros detectores de pontos de interesse em imagens data de Moravec[**?**] seu método considerava um córner como sendo um ponto com baixa autosimilaridade, O trabalho de Moravec foi posteriormente melhorado por Harris and Stephens[**?**] que aumentaram a repetibilidade do método de Moravec sendo um detector amplamente utilizado. Lowe[**?**] em 1999 estendeu o trabalho de Harris tornando o detector independente da escala. Posteriormente Lowe[**?**] criou um detector e descritor de pontos de interesse em imagens, invariante a rotação e

escala apelidado de SIFT, que obteve grande sucesso em aplicações como rastreamento e reconhecimento. Mais recentemente Herbert et al.[**?**] propuseram o SURF (speeded up robust features) que foi mostrado ser mais eficiente [**?**, **?**, **?**] e obter uma taxa de acerto tão boa ou melhor que a do SIFT tanto em ambientes internos quanto externos.

Dado o sucesso obtido pelo SURF na representação e correspondência de pontos de interesse, matching de imagens e reconhecimento. Esse trabalho se propõe a estudar como o descritor do surf pode ser adaptado e utilizado para resolver o problema de sincronização de múltiplas seqüências de vídeo da mesma cena.

3 Formulação do Problema

Assumindo que apenas duas seqüências devem ser alinhadas no tempo podemos definir uma matriz de medição. $S_1 = \begin{pmatrix} I_1^1 & I_2^1 & ... & I_n^1 \end{pmatrix}$ e $S_2 = \begin{pmatrix} I_1^2 & I_2^2 & ... & I_m^2 \end{pmatrix}$ obtidas da mesma cena a uma taxa de quadros constante. A taxa de quadros não necessariamente precisa ser a mesma chamando de $\theta_1(n)$ o tempo em que a n-ésima imagem da seqüência 1 foi capturada e $\theta_2(n)$ o tempo de captura da n-ésima imagem da seqüência 2 podemos relacionar os tempos das duas seqüências através da equação

$$\theta_1(n) = c \cdot \theta_2(n) + \delta t$$

Onde c é uma constante definida pela razão da taxa de quadros e δt é a diferença de tempo entre o início das capturas.

Se tivermos n pontos característicos da cena rastreados nas duas seqüências podemos definir uma matriz de medição $\mathbf{W}$[**?**]:

$$\mathbf{W} = \begin{bmatrix} u_1^1 & u_2^1 & ... & u_n^1 \\ v_1^1 & v_2^1 & ... & v_n^1 \\ u_1^2 & u_2^2 & ... & u_n^2 \\ v_1^2 & v_2^2 & ... & v_n^2 \end{bmatrix}$$

para cada par de imagens nas seqüências onde $(u_n^i, v_n^i)^T$ corresponde a n-ésima característica observada na i-ésima visão.

A matriz de medição é então normalizada com respeito ao seu centróide de forma que cada linha possui média 0. Conforme foi mostrado por Tomasi e Kanade[**?**] assumindo uma correspondência exata e projeção afim podemos decompor a matriz de medição em:

$$\mathbf{W} = \begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & ... & \mathbf{X}_n \end{bmatrix}$$

onde P_i é a matriz de projeção afim 2×3 relacionada ao i-ésimo ponto de vista e X_n 3×1 corresponde ao ponto 3D correspondente a n-ésima característica.

Sob condições ideais [**?**] a matriz W possuí posto igual a 3 porém devido a ruídos e correspondências inexatas a matriz W quase sempre terá posto 4.

Quando a matriz W contem somente pontos obtidos no mesmo instante de tempo o quarto valor singular depende apenas dos ruídos nas medições portanto podemos esperar, se as correspondências forem boas e o ruído for pequeno, que o quarto valor singular de W seja baixo para imagens obtidas no mesmo instante de tempo.

4 Metodologia.

tomando como referencia uma das seqüências definimos a matriz W(F,f) como sendo a matriz $4 \times NM$:

$$W(F,f) = \begin{pmatrix} u^F_{ref,1} & \dots & u^F_{ref,n} & u^{F+1}_{ref,1} & \dots & u^{F+M-1}_{ref,n} \\ v^F_{ref,1} & \dots & v^F_{ref,n} & v^{F+1}_{ref,1} & \dots & v^{F+M-1}_{ref,n} \\ u^f_{tg,1} & \dots & u^f_{tg,n} & u^{f+1}_{tg,1} & \dots & u^{f+M-1}_{tg,n} \\ v^F_{tg,1} & \dots & v^f_{tg,n} & v^{f+1}_{tg,1} & \dots & v^{f+M-1}_{tg,n} \end{pmatrix}$$

1. para cada seqüência de frames S_1 , S_2 encontre e rastreie pontos de interesse $p = (u_n, v_n)$.
2. calcule as correspondências entre os pontos nas duas seqüências.
3. para cada frame F na primeira seqüência f na segunda seqüência calcule a matriz W(F,f).
4. para cada frame F ache o f que minimize o quarto valor singular de W.
5. faça uma regressão linear para os valores calculados de

$$\theta_1(F) = c \cdot \theta_2(f) + \delta t$$

Encontrar os pontos de interesse nas cenas pode ser feito utilizando um dos diversos métodos descritos na literatura como o de harris[**?**], o SURF[**?**] ou o SIFT[**?**]. esses pontos podem ser rastreados nas cenas usando um rastreador simples que utiliza filtragem de Kalman ou um filtro de Partículas.

O passo das correspondências pode ser feito utilizando os próprios métodos de comparação do SURF e do SIFT reforçando que se um ponto **p** em um frame possui um correspondente $\mathbf{p}'$ os correspondentes nos frames subseqüentes devem se manter ou seja mesmo ponto **p** em todos os frames da seqüência deve ter o mesmo correspondente $\mathbf{p}'$ pode-se assim evitar muitas falsas correspondências. já que essas podem ser verificadas com as correspondências nos outros quadro

References

[1] Herbert Bay, Beat Fasel, and Luc Van Gool. Interactive museum guide: Fast and robust recognition of museum objects. In *Int. Workshop on Mobile Vision*, 2006.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Ninth European Conference on Computer Vision*, 2006.

[3] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Fourth Alvey VisionConference*, pages 147–151, Manchester, 1988.

[4] David G. Lowe. Object recognition from local scale-invariant features. In *In International Conference on Computer Vision*, pages 1150–1157, Corfu Greece, 1999.

[5] David G. Lowe. Distinctive image features from scale-invariant key points. *International Journal of Computer Vision*, 2:91–110, 2004.

[6] Hans Moravec. Rover visual obstacle avoidance. In *International Joint Conference on Artificial Intelligence*, pages 785–790, Vancouver, 1981.

[7] Denis Simakov, Yaron Caspi, and Michal Irani. Feature-based sequence-to-sequence matching. In *Workshop on Vision and Modelling of Dynamic Scenes*, Copenhagen, May 2002.

[8] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography:a factorization approach. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[9] Phil Tresadern and Ian Reid. Synchronizing image sequences of non-rigid objects. In *British Machine Vision Conference*, Norwich, 2003.

[10] Christopher Valgren and Achim Lilienthal. Sift, surf and seasons: Long-term outdoor localization using local features. In *European conference in mobile robots*, 2007.

[11] Lior Wolf and Assaf Zomet. Correspondence-free synchronization and reconstruction in a non-rigid scene. In *Workshop on Vision and Modelling of Dynamic Scenes*, Copenhagen, May 2002.

[12] Jingyu Yan and Marc Pollefeys. Video syncronization via space-time interest point distribution. In *Advanced Concepts for Intelligent Vision Systems*, 2004.