一种适用于 2D/3D 转换的分段化结构重建技术

刘伟，吴毅红，郭复胜，胡占义

（中国科学院自动化研究所模式识别国家重点实验室　北京　100190）

wliu @nlpr.ia.ac.cn

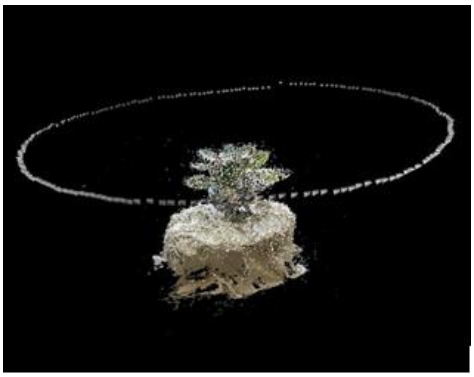
近些年来，3D 电影逐渐普及，成为人们日常娱乐生活的重要部分。相比传统的 2D 电影，3D 技术可以提供更加身临其境的逼真效果，并逐渐成为了当前发展的一种趋势。然而 3D 电影从题材的选择、拍摄、剪辑、洗印到发行放映，都有一些特殊的技术要求，制作成本较高、周期也相对较长。虽然现在已经有近百部的 3D 电影问世，但是在 2D 和 3D 技术共存的局面下，面对内容丰富的传统电影的挑战，3D 影片片源依然是杯水车薪。在这种情况下，将 2D 电影转为 3D 电影是解决此问题的有效途径，也是近年来产业界和学术界的关注热点。

在之前的综述中[1]，我们把 2D/3D 转换技术根据摄像机和场景的相对运动关系分为了四大类，每一类都对应着不同的研究方法，它们针对不同场景、利用不同的深度线索、采用了不同的转换方法。

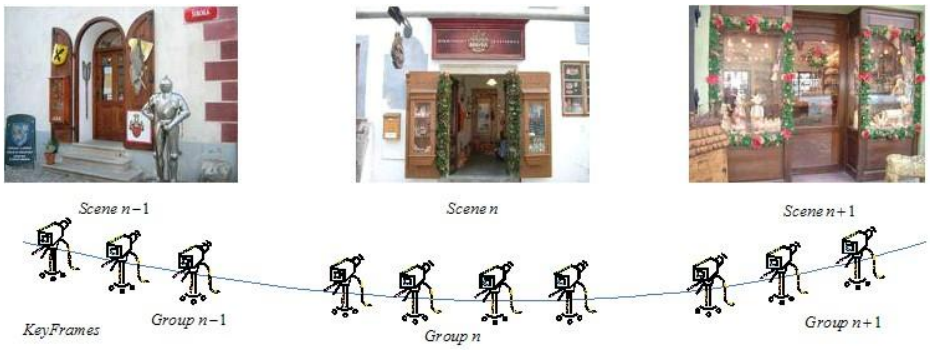
结构重建技术(Structure from Motion, SFM)是计算机视觉中一个重要的研究方向，广泛应用于古迹重建，电影制作，城市建模等领域。由于这种技术可以从静止的场景、运动的摄像机拍摄的图片集中获取场景的深度线索，正好符合 2D/3D 转换方法中的一类情况，所以我们对这种技术在 2D/3D 转换中的应用做了深入的研究，提出了一种分段化结构重建框架，力求解决一部分场景视频的转换问题。

1 2D/3D 转换中的结构重建技术特点

SFM 技术通过图片集中的匹配点来估计三维静止场景中运动摄像机的内外参数和该场景相对于一个参考坐标系的结构关系。我们利用这种技术来获取场景离散的深度信息。然而，基于 SFM 的 2D/3D 视频转换方法与传统的基于 SFM 的视频重建方法[2]相比具有两个明显的特点。首先，传统的基于 SFM 的视频重建方法所用的视频源专注于一个场景，如图 1a 所示；而在 2D/3D 转换技术中所处理的视频源往往包含多个连续的场景，如图 1b 所示。其次，视频重建的整个过程以一个统一的参考坐标系为基准，追求全局结构的优化；而对于 2D/3D 转换方法，在多场景的视频片段中深度图的生成仅仅依赖于对应场景的三维结构信息，而不需要获得整个视频所描绘的三维场景，更加强调局部结构的优化。基于上述分析，我们提出了分段化结构重建框架。



(a) 基于 SFM 的视频重建[2]



(b) 2D/3D 转换中视频流的连续场景和关键帧序列

图 1 基于 SFM 的视频重建和 2D/3D 转换常用视频拍摄方式对比

2 分段化结构重建

分段化结构重建指的是在视频流中，对每一个子序列分别执行 SFM 来恢复局部场景的结构和运动信息。本节我们将讨论分段化结构重建框架，整个框架流程如图 2 所示。

在本文中，视频流中所有属于同一个场景的视频帧被称为一个子序列，第 n 个子序列用 *Sequence n* 表示，子序列中的所有关键帧用 *Group n* 表示。三维空间中的点用齐次坐标 X 表示，它在图片中的投影用齐次坐标 x 表示。两者之间的关系为：$\lambda x=PX$，λ 为非零尺度因子，$P=K[R\ t]$。这里 R, t 是摄像机的外参，其中 R 是 3×3 旋转矩阵，t 是平移向量，K 是摄像机的内参，并且：

$$K=\begin{bmatrix} f & 0 & c_x \\ & rf & c_y \\ & & 1 \end{bmatrix} \tag{1}$$

其中 f 表示摄像机的焦距，r 表示像素长宽比，(c_x, c_y)是图像主点。在实验中，我们将 r 取为 1，(c_x, c_y)取为图像的中心坐标。

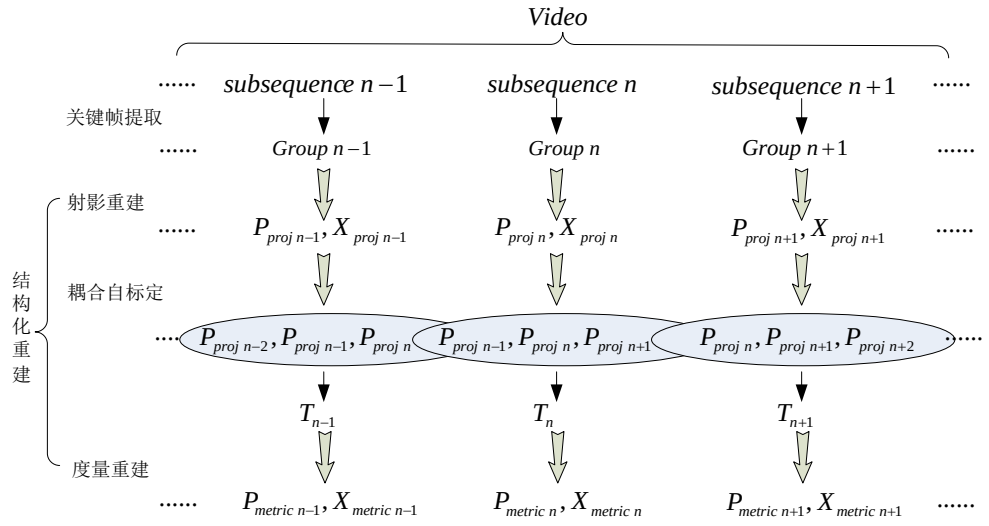


图 2 分段化结构重建框架

2.1 关键帧的提取

第一步是关键帧的提取，如图 3 所示，在这里我们采用了分层提取的方式。对于输入的视频，首先通过一种视频摘要算法[3]把视频流根据不同的场景划分为连续的子序列；然后基于 Gric(Geometric Robust Information Criterion)准则[4][5]在每个子序列中分别对关键帧进行提

取。GRIC 是一种通用的鲁棒模型选择准则,它对每一个运动模型(F,H)都分别构造了一个分值函数，通过比较一对图片不同模型的分值函数大小就可以判断它们是否处于退化状态。它的形式如下所示：

$$GRIC = \sum_i \rho(e_i^2)_i + \lambda_1 dn + \lambda_2 k$$
$$\rho(e_i^2) = \min(\frac{e_i^2}{\sigma^2}, \lambda_3(r-d)) \quad (1)$$

其中，e_i 表示留数，d 为选择模型的尺度（模型为 F 时 d=3，模型为 H 时 d=2),n 为两幅图片匹配点的个数，k 为选择模型的自由度（模型为 F 时 k=7，模型为 H 时 k=8),对于二维匹配点 r=4,σ^2 为误差的方差, λ_1=log(r), λ_2=log(rn), λ_3 限制留数的值。

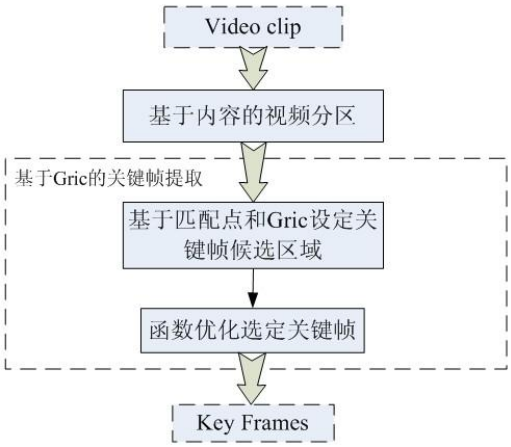


图 3 关键帧的提取

SFM 过程中有两类退化现象，一类是运动退化，例如只做旋转运动的摄像机，一类是结构退化，即拍摄的场景在三维空间中处于同一平面中。选择关键帧时应尽量避免这两种情况的出现。在应用 GRIC 准则时，如果 $GRIC_F > GRIC_H$,那么就认为这对图片处于一种退化的关系。除此之外，还考虑到视频流中连续帧之间的基线较短，拉长基线可以增强 SFM 的准确性，但同时会造成匹配点的减少，所以关键帧的选择应该考虑在保持一定匹配点数目的基础上尽可能的拉长两帧之间的基线。综合这两个方面，首先我们设计了下面的算法流程来选定关键帧候选区域：

Algorithm1. To find the *jth* Key Frame Candidate zone $\Phi(k_j)$ around Frame *m*

1: Input: Video stream in a scene zone
2: Output: the *jth* Key Frame Candidate zone $\Phi(k_j)$
3: **for** $i=m-N_{seachzone}$; $i<m+N_{seachzone}$; $i=i+1$
4: Match keypoints between frames k_i and k_{j-1}
5: Compute *H* and *F* using *RANSAC*
6: Discard outlier matches
7: Calculate correspondence radio R_c
8: **if** $R_c<T_{min}$ or $R_c>T_{max}$ then
9: continue
10: **end if**
11: **if** $GRIC_H(k_i, k_{j-1}) \leqq GRIC_F(k_i, k_{j-1})$ then
12: continue

13: **end if**

14: $\Phi(k_j) \leftarrow i$

15: **end for**

然后，我们又采用了函数优化的方式从关键帧候选区域中进一步选定合适的关键帧。假定一个视频序列为 *f(1), f(2),……, f(n),*$\Phi(k_j)$表示第 *j* 个关键帧候选区域，则关键帧优化函数为：

$$k_i = \arg\min_{j\in\varphi(k_i)}(f(j))$$
$$f(j) = \max(GRIC_F(k_{i-1}, j) - GRIC_H(k_{i-1}, j), GRIC_F(j,n) - GRIC_H(j,n)) \quad (2)$$

其中 k_i 表示第 *i* 个关键帧，$GRIC_F(i,j)$表示视频帧 *i* 和视频帧 *j* 当模型为 F 时的 *GRIC* 分值大小，$GRIC_H(i,j)$表示视频帧 *i* 和视频帧 *j* 当模型为 H 时的 *GRIC* 分值大小。这样在满足角点匹配率并排除了退化情况的前提下就能保证提取出的关键帧之间的基线尽可能的长。

在如图 4 所示的实验中，摄像机沿着红色轨迹在平移运动中拍摄了工作台的一角，整段视频共有 189 帧，首先基于场景内容的变化把视频流分为了多个子序列，而后在每个区域中提取出相应的关键帧，图 5 展示了其中一部分关键帧。

图 4 实验场景及摄像机运动轨迹示意图

(a) 子序列 1 部分关键帧

(b) 子序列 2 部分关键帧

(c) 子序列 3 部分关键帧

图 5 子序列中的部分关键帧

2.2 结构化重建

将视频分为子序列并提取出每个子序列的关键帧后，就可以在每个子序列中进行结构化重建过程。整个过程又分为射影重建、自标定和度量重建三个部分，下面分别予以介绍。

首先，对于每一组子序列利用上一步提取出的关键帧分别进行传统的射影重建[6]。

然后是摄像机的自标定。普通视频并不存储拍摄时摄像机的参数，况且在视频拍摄中摄像机参数会随着时间变化，所以摄像机自标定是分段化结构重建中的重要一步。我们改进了一种鲁棒自标定方法。因为一个子序列中的关键帧的数量是有限的，而参与自标定的视频序列越长标定的结果越可靠。所以在标定一组子序列的参数时，我们同时利用了相邻子序列的信息来增强标定的平滑性和稳定性。

在计算机视觉中，三维空间中绝对二次曲面在视平面中的投影对应着绝对二次曲线。它们的关系如下式所示：

$$\lambda KK^T = P\Omega^* P^T \tag{3}$$

其中Ω^*表示射影空间下的绝对二次曲面，它是一个 4×4 秩为 3 的对称矩阵，P 是一个投影矩阵。根据文献[2]提出的线性自标定方法，通过加入权重因子考虑了不确定因素后子序列中的每个关键帧对应着下面的方程组：

$$\begin{aligned}
\tfrac{1}{9v}\left(P_1\Omega^* P_1^T - P_3\Omega^* P_3^T\right) &= 0\\
\tfrac{1}{9v}\left(P_2\Omega^* P_2^T - P_3\Omega^* P_3^T\right) &= 0\\
\tfrac{1}{0.2v}\left(P_1\Omega^* P_1^T - P_2\Omega^* P_2^T\right) &= 0\\
\tfrac{1}{0.1v}\left(P_1\Omega^* P_2^T\right) &= 0\\
\tfrac{1}{0.1v}\left(P_1\Omega^* P_3^T\right) &= 0\\
\tfrac{1}{0.01v}\left(P_2\Omega^* P_3^T\right) &= 0
\end{aligned} \tag{4}$$

其中 P_i 表示投影矩阵 P 的第 i 个行向量，v 是尺度因子，初始化时设为 1，在随后的方程组每次迭代求解过程中设为 $P_3\Omega^* P_3^T$。对于每个子序列，如果把其中一个关键帧的投影矩阵设为 P= [I | 0]，那么(4)中的Ω^*就可以变换为下面的形式：

$$\Omega^* = \begin{bmatrix} KK^T & a \\ a^T & b \end{bmatrix} \tag{5}$$

把一个子序列中所有关键帧对应的方程组(公式 4)联立成一个更大的方程组，通过线性最小二乘法就能够求解Ω^*。这样对于子序列 n，可以推导出以下方程组：

$$[C_n \; D_n]\begin{bmatrix} k_n \\ \begin{bmatrix} a_n \\ b_n \end{bmatrix} \end{bmatrix} = 0 \tag{6}$$

其中 n 是子序列的序号，k_n 是矩阵 $K_n K_n^T$ 的向量化形式，a_n 是一个 3 维向量，b_n 是一个标量，C_n, D_n 包含了子序列 n 中所有关键帧对应的方程组所包含的系数。

在通常包含多场景的镜头中，摄像机的焦距是连续变化的，摄像机的参数也不会发生突然的变化。所以我们可以假设子序列 *Group n* 和它相邻的两个子序列(*Group n-1* 和 *Group n+1*)具有相同的摄像机内参。这样 *Group n* 的内参 k_n 就能够从下面的耦合方程组中解出：

$$\begin{bmatrix} C_n & D_n & 0 & 0 \\ C_{n-1} & 0 & D_{n-1} & 0 \\ C_{n+1} & 0 & 0 & D_{n+1} \end{bmatrix} \begin{bmatrix} k_n \\ \begin{bmatrix} a_n \\ b_n \end{bmatrix} \\ \begin{bmatrix} a_{n-1} \\ b_{n-1} \end{bmatrix} \\ \begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} \end{bmatrix} = 0 \tag{7}$$

通过(k_n, a_n, b_n)我们可以得到射影空间中的Ω_n^*。因为在计算机视觉中，存在变换矩阵 T_n 可以使射影空间下的绝对二次曲面上升到度量空间下的标准形式：$\Omega_n^* \to diag(1,1,1,0)$，它同样也可以将子序列 *Group n* 从射影空间上升到度量空间。这样通过约束$T_n\Omega_n^*T_n^T = diag(1,1,1,0)$求解出 T_n，就可以得到度量重建结果，完成最后一步：

$$P_{metric\,n} = P_{proj\,n}T_n^{-1} \quad and \quad X_{metric\,n} = T_n X_{proj\,n} \tag{8}$$

我们的结构重建方法具有以下两个特点：1) 通过耦合，自标定结果比从单个子序列中得到的结果更加鲁棒。同时由于射影重建是在每一个子序列中独立进行的，这样有效的避免了射影漂移现象[2]的产生。2) 我们的方法可以处理变焦的长镜头。在估计子序列 *Group n*-1 和 *Group n*+1 的数据信息，同时假设(*Group n*-1, *Group n*, *Group n*+1)具有相同的摄像机内参。这仅仅是为了增强自标定结果的鲁棒性，并不意味着在这三个连续的子序列中摄像机内参不变。事实上，当估计子序列 *Group n*+1 的摄像机内参时，我们又利用了 *Group n* 和 *Group n*+2 的数据信息，同时假设(*Group n*, *Group n*+1, *Group n*+2)具有相同的摄像机内参。

在另一组实验中，通过平移摄像机沿着工作台桌面从左到右进行了拍摄。定焦摄像机用标定板测得的焦距值为 715 像素，并把它作为焦距真值。在分段化结构重建过程中，自标定算法一共执行了 6 次。图 6 是每个子序列自标定得到的焦距估值，可以看到分段化结构重建框架保持了焦距估值的平稳性，能够提供可靠的深度线索。

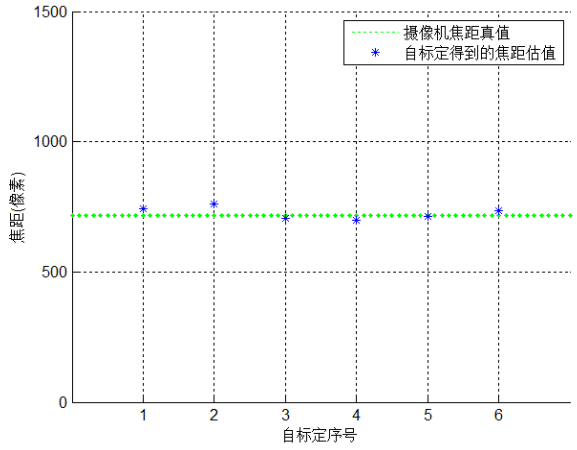


图 6 耦合自标定焦距估计

3 结论

我们结合了 2D/3D 转换技术的特点提出了一种分段化结构重建框架，该框架能够有效的从视频中获得连续场景的结构信息，为 2D/3D 转换提供必要的深度线索。在下一步的工

作中，我们将在此基础上利用分段化结构重建框架提供的深度线索进一步研究高效的转换方法，使 SFM 技术能够真正解决这一类视频的转换问题。

参考文献：

[1] Liu Wei, Wu Yihong, Hu Zhanyi. A Survey of 2D to 3D Conversion Technology for Film. Journal of Computer-Aided Design & Computer Graphics,2012,24(1):1-15 (in Chinese)
(刘伟，吴毅红，胡占义．电影 2D/ 3D 转换技术概述[J]．计算机辅助设计与图形学学报，2012，24(1):1-15)

[2] J. Repko, M. Pollefeys. “3D Models from Extended Uncalibrated Video Sequences: Addressing Key-frame Selection and Projective Drift,” in *Proc. 3DIM*, pp.150-157, 2005

[3] T. Liu and J.R. Kender, “Computational approaches to temporal sampling of video sequences,” *ACM Trans. Multi. Comput. Commun. Appl.*, vol. 2,no. 2, pp. 7-29, 2007

[4] M.T. Ahmed and M.N. Dailey, “Robust key frame extraction for 3D reconstruction from video streams,” in *Proc. VISAPP*, vol. 1, pp. 231-236, 2010.

[5] P.H.S. Torr. “An assessment of information criteria for motion model selection,” in *CVPR97*, pp.47-53, 1997.

[6] R. Hartley and A. Zisserman, *Multiple view geometry*. Cambridge: Cambridge University Press, UK, 2003, pp. 262-276