新兴数字视频稳像中相机运动估计技术综述

魏闪闪[1,2]，谢 巍[2]，贺志强[3]

(1. 北京航空航天大学 计算机学院，北京 100191；2. 联想研究院 SoC 中心，北京 100085；3. 联想集团 生态系统与云服务业务群组，北京 100085)

摘 要：为了解当前数字视频稳像中相机运动估计技术的发展现状、存在问题和发展前景，对其进行了研究。通过对相关文献的归纳和分析，将新兴数字视频稳像中相机运动估计技术分为 3D 运动估计、传感器辅助运动估计以及新兴 2D 运动估计，并分别讨论了各种运动估计技术的研究现状、研究成果、不足和挑战；最后，结合技术和市场因素对数字视频稳像中相机运动估计技术的发展前景进行了展望。研究表明，当前数字视频稳像中相机运动估计技术在实时应用和估计精度上都有待改善。

关键词：新兴数字视频稳像；相机运动估计；综述

中图分类号：TP391.41 **文献标志码**：A **文章编号**：1001-3695(2017)02-0321-07

doi:10.3969/j.issn.1001-3695.2017.02.001

Camera motion estimation in emerging digital video stabilization: a survey

Wei Shanshan[1,2], Xie Wei[2], He Zhiqiang[3]

(1. *School of Computer Science & Engineering, Beihang University, Beijing* 100191, *China*; 2. *SoC Center, Lenovo Corporate Research & Development, Beijing* 100085, *China*; 3. *Business Groups of Ecological System & Cloud Services, Lenovo, Beijing* 100085, *China*)

Abstract: To find out the present situation, problems and future development of camera motion estimation (CME) in digital video stabilization (DVS), this paper made a careful study of it. This paper surveyed and analyzed current related articles, and divided CME in the emerging DVS into 3D method, sensor-based method and emerging 2D method. For each method, this paper discussed its research status, achievements, inefficiencies and challenges. Considering both the technology and market factors, this paper prospected its future development. The result of the research shows that the CME techniques in DVS are still needed to be improved both in real-time application and in estimation accuracy.

Key words: emerging digital video stabilization; camera motion estimation; survey

0 引言

视频稳像技术的目的是消除或减少视频抖动，生成人眼视觉上稳定的视频。它总体上可以分为三大类：机械稳像、光学稳像和数字稳像[1]。机械稳像是早期摄像机常用的稳像技术，采用稳定整个摄像机的方法；光学稳像系统对光路进行重定向或者移动成像板；除了机械稳像和光学稳像技术之外，还有两类视频稳像技术：电子稳像技术和纯数字稳像技术。这两类技术很相似，区别在于电子稳像技术使用硬件传感器(如陀螺仪等)来检测相机抖动，而纯数字稳像技术通过图像处理方法分析连续视频帧的运动来估计相机抖动。在得到相机运动向量之后，两者都进行运动补偿，最后根据补偿后的运动进行图像修补。在本文中，电子稳像和纯数字稳像技术被统一称为数字稳像技术，因为两者在图像补偿阶段都采用了数字方法。

相机在 3D 空间中运动，而图像的像素在 2D 平面上运动；用户运动引起相机运动，而相机运动又导致图像上像素的运动。可见 2D 图像上的像素运动与相机空间 3D 运动二者之间是存在运动关联的，因此常见的数字稳像技术采用三步法：运动估计、运动补偿和图像修补。其中运动估计是第一步也是最重要的步骤，它是确定运动向量的过程，这些运动向量是描述 2D 图像(通常是连续的视频帧)间运动转换的量。相机运动向量可能跟整张图像相关(全局运动估计)，也可能跟图像某一部分相关，如矩形块、任意形状块甚至是每个像素。

根据所选场景运动模型的复杂程度，可将当前数字视频稳像技术分为 2D 数字稳像技术(简称 2D 方法)和 3D 数字稳像技术(简称 3D 方法)两类。相机运动估计是指在相应场景运动模型下，从图像信息中恢复相机姿态的问题。早期数字视频稳像技术采用 2D 方法：使用 2D 变换模型来表示相机运动，不考虑恢复场景的 3D 几何信息。2D 方法计算简单、算法稳健，但是它忽略了相机本身的 3D 空间运动信息，有时可能会有平行视差问题的存在，因此稳像效果有限。为解决此问题，研究者提出了 3D 稳像技术，它通过恢复相机 3D 空间运动，提高运动估计的质量和视频稳像的效果。由于 3D 运动估计需要从 2D 图像信息恢复 3D 运动信息，所以 3D 运动估计是病态问题(ill-posed problem)。3D 方法稳像效果显著，但由于 3D 运动场景恢复计算复杂且受限于诸多因素，对于当前流行智能设备的计算能力而言并不适合实际应用。因此近几年研究者把目光又转向了 2D 方法。与早期 2D 方法不同，它们不再采用传统

收稿日期：2016-03-28；**修回日期**：2016-05-12

作者简介：魏闪闪(1986-)，男，河北石家庄人，博士研究生，主要研究方向为图像视频处理、机器视觉(wswss11986@qq.com)；谢巍(1974-)，男，江西南城人，教授级高工，博士，主要研究方向为集成电路设计、图像处理；贺志强(1963-)，男，山西太原人，研究员，博导，硕士，主要研究方向为计算机系统结构、计算机应用技术.

“三步法”,而是借助传感器或者特征轨迹来进行运动估计。本文对近年来新兴的数字稳像技术中的运动估计技术进行了综述。

1 特征点问题

与数字视频稳像紧密相关的一类问题是特征点匹配问题。首先绝大多数方案的运动估计与之相关,它们通常的几何假设前提是:匹配特征点问题已解决,因为特征点的匹配关系中包含了相机运动信息;其次大多数图像修补技术也需要特征点匹配来辅助完成。

目前存在多种特征点匹配算法,适合于视频稳像运动估计的也有很多。Amisha[2]对2D稳像应用中经常采用的特征点匹配方法进行了对比分析,列出了它们的优点和缺点。在新兴数字视频稳像应用中,对特征点匹配精度和稳健性的要求更高,最常用的特征点匹配方法相比2D应用而言要少许多。最常用的算法包括尺度不变特征转换方法(scale-invariant feature transform,SIFT)、加速稳健特征算法(speeded up robust features,SURF)以及用于特征点跟踪的KLT(Kanade-Lucas-Tomasi)算法。其他2D运动估计中常用的特征点方法由于在旋转或缩放不变性或者抗噪能力方面表现较差而未被新兴稳像技术所采用。

Lowe[3]在前人研究的基础上,将斑点检测、特征描述以及特征匹配搜索结合在一起优化,提出了里程碑式的SIFT算法。该算法首先发表于1999年计算机视觉国际会议(International Conference on Computer Vision,ICCV),2004年再次经作者整理完善后发表于计算机视觉国际期刊《International Journal of Computer Vision,IJCV》。对于斑点检测,Lowe提出了LoG的近似算法DoH,提高了检测效率。在Lowe的基础上,Bay等人[4]在2006年提出了加速稳健特征算法SURF,对DoH进行了简化和近似。KLT算法由Kanade和Lucas[5]在20世纪80年代提出,并由Shi等人[6]在1994年作了进一步的完善。

新兴稳像方案中常用的特征点方法总结如表1所示。除了主流应用的算法外,SIFT流方法[7]以及粒子视频法[8]也有应用。

表1 新兴稳像方案中常用的特征点方法

稳像方案	主流特征点方法	其他特征点方法
3D稳像	SIFT, KLT	SURF
基于特征轨迹稳像	KLT	SIFT流,粒子视频

2 3D稳像运动估计

Buehler等人[9]提出的基于图像渲染(rmage-based rendering)技术的视频稳像方案被视为最早的3D稳像方案。作者在文中指出基于图像的渲染技术能够用于视频稳像应用的三个必备技术条件:帧间匹配特征点;帧间插值;场景中虚拟相机轨迹操作。这为以后的3D稳像技术提供了框架:通过2D图像信息来重构相机的3D运动信息,然后对恢复的相机3D运动进行平滑处理去除抖动,生成平滑后新的相机运动信息,最后根据这些新的运动信息来完成图像的修补。即3D稳像方案采用三步法:相机3D运动估计、相机3D运动补偿和2D图像修补。进行3D运动估计首先需要确定相机3D运动模型。

2.1 运动模型

当前主流3D稳像方案通常采用针孔相机模型下的投影变换运动模型。针孔相机模型如图1所示。针孔相机模型适合于很多计算机视觉应用,它完成中心投影[10]。模型中C为光心(焦点),光心垂直于成像平面与其交于点P(光心投影点),CP长度为相机焦距f。

模型中四个坐标系的定义如下:

a)3D世界坐标系$W(X,Y,Z)$

点坐标用齐次坐标表示,形式为$\tilde{X}\sim(X,Y,Z,1)^{\mathrm{T}}$。

b)3D相机坐标系$C(X_c,Y_c,Z_c)$

点坐标用齐次坐标表示,形式为$\tilde{X}_c\sim(X_c,Y_c,Z_c,1)^{\mathrm{T}}$。

c)2D图像成像板坐标系$P(x,y)$

点坐标用齐次坐标表示,形式为$\tilde{x}\sim(x,y,1)^{\mathrm{T}}$。

d)2D图像像素坐标系$I(u,v)$

点坐标用齐次坐标表示,形式为$\tilde{u}\sim(u,v,1)^{\mathrm{T}}$。

相机坐标系到世界坐标系的转换可以用$[T|R^{\mathrm{T}}]$来表示,即三维旋转和平移操作。在此针孔模型下,世界坐标系$W(X,Y,Z)$中一个3D点映射到2D图像像素坐标要经过三次坐标间的转换:

a)3D世界坐标系到3D相机坐标系。

$$W(X,Y,Z)\rightarrow C(X_c,Y_c,Z_c)$$

这个转换就是相机的外参数。将世界坐标系中$\tilde{X}\sim(X,Y,Z,1)^{\mathrm{T}}$转换到相机坐标系中的点$\tilde{X}_c\sim(X_c,Y_c,Z_c,1)^{\mathrm{T}}$,用矩阵运算表示为

$$[X_c\quad Y_c\quad Z_c\quad 1]^{\mathrm{T}}\sim\begin{bmatrix}R & t\\ 0 & 1\end{bmatrix}[X\quad Y\quad Z\quad 1]^{\mathrm{T}}\tag{1}$$

b)3D相机坐标系到2D成像板坐标系。

$$C(X_c,Y_c,Z_c)\rightarrow P(x,y)$$

将相机坐标系中的点$\tilde{X}_c\sim(X_c,Y_c,Z_c,1)^{\mathrm{T}}$转换到图像成像板坐标系中的点$\tilde{x}\sim(x,y,1)^{\mathrm{T}}$,用矩阵运算表示为

$$[x\quad y\quad 1]^{\mathrm{T}}\sim\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\end{bmatrix}[X_c\quad Y_c\quad Z_c\quad 1]^{\mathrm{T}}\tag{2}$$

c)2D图像成像板坐标系到2D图像像素坐标系。

$$P(x,y)\rightarrow I(u,v)$$

将图像成像板坐标系中的点$\tilde{x}\sim(x,y,1)^{\mathrm{T}}$转换到图像像素坐标系中的点$\tilde{u}\sim(u,v,1)^{\mathrm{T}}$,用矩阵运算表示为

$$[u\quad v\quad 1]\sim K\tilde{x}=\begin{bmatrix}f&s&u_0\\0&f&v_0\\0&0&1\end{bmatrix}[x\quad y\quad 1]^{\mathrm{T}}\tag{3}$$

其中:K即为相机内参数,s为倾斜因子,光心投影坐标落在图像像素坐标的(u_0,v_0)处。

综合三个变换,即

$$\tilde{u}\sim P\tilde{X}\tag{4}$$

$P\sim K[R\ T]$称为投影变换矩阵。

基于针孔相机模型的相机投影变换运动如图2所示。

如图2所示,在针孔相机投影变换运动模型下,当相机位置从世界坐标系中一点C移动到C'时,成像物体从位置O移动到O',物体在相机上成像像素位置也相应随之变化,从P移动到P'。如果在两个相机位置,物体静止,则相应成像板上像素的移动是由相机运动导致的;如果期间物体运动,则相应图

像上像素的移动除了考虑相机运动,还应考虑物体本身运动导致的像素运动。

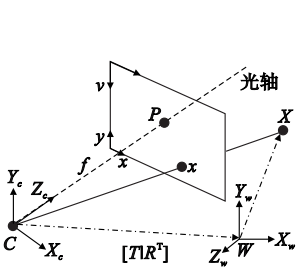

图1　针孔相机模型

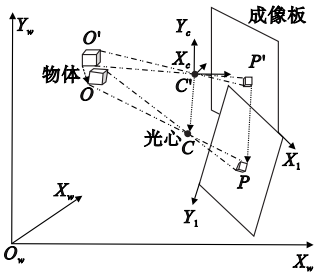

图2　针孔模型投影变换运动模型

以上是3D运动估计的模型,在视频稳像中进行相机3D运动估计即在已知若干连续视频帧条件下求解这些帧对应的相机3D运动的过程。

2.2　3D运动估计算法

因为投影模型具有尺度不变性,这里的相机运动估计是指相对姿态(related camera pose)估计,这对于视频稳像目的来说恰恰是足够的。正如最早3D稳像视频方案[9]中所采用的,求解3D运动估计的传统方法是运动到结构算法(structure from motion,SFM)。

2.2.1　运动到结构算法

SFM算法从二维视频序列中恢复出相应三维相机运动信息,其中包括成像相机的运动参数以及场景的结构信息。

SFM问题可以通过双视角、三视角或者多视角方案来解决,其中最常用也是3D稳像中使用的是双视角方法:给定两个视角相机的图像,恢复两个成像时刻相机的位置信息。其描述如图3所示。图3中,给定一个欧氏空间点 X_0,其在相机 C 坐标系中的坐标为 X,在相机 C' 中的坐标为 X',两者之间的转换关系可以表示为

$$X = RX' + T \tag{5}$$

其中:R 为3×3旋转矩阵,T 为三维向量,两者即为3D运动估计需要求解的运动信息。它的求解需要数学技巧,将式(5)两边都乘以 $X^T[T]_\times$ 可将等式简化为

$$X^T[T]_\times RX' = X^T EX' = 0 \tag{6}$$

这里 $[T]_\times$ 是向量 T 的叉积矩阵,即

$$T = [T_x \quad T_y \quad T_z], [T]_\times = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix}$$

其中:$E \sim [T]_\times R$ 为3×3的本征矩阵(essential matrix),它只决定于 R 和 T,且是确定在相差一个尺度的意义下的。式(6)对于图像成像板坐标同样成立,即

$$x^T[T]_\times Rx' = x^T Ex' = 0 \tag{7}$$

式(7)即为计算机视觉中的极限约束条件。进一步,根据图像成像板坐标与图像像素坐标的关系,如果相机内参 K 已知,则

$$x \sim K^{-1}u \tag{8}$$

将式(8)代入式(7)可以得到极限约束条件的像素坐标表达式:

$$(K^{-1}u)^T E(K'^{-1}u') = u^T(K^{-1}EK'^{-1})u' = u^T Fu' = 0 \tag{9}$$

$F \sim K^{-1}EK'^{-1}$ 称为基础矩阵[11](fundamental matrix),它是3×3矩阵,秩为2,且包含了相机姿态参数 R 和 T。于是相机相对姿态获取问题就转换为基础矩阵分解的问题。该问题在计算机视觉领域被称为最小问题(minimal problems),其解法分为相机标定和未标定两大类,如果相机参数未标定,则需要在算法中确定相机参数;反之,则相机参数作为已知参与运算。相对姿态的未知量包括3个旋转量、2个平移量(确定在相差一个尺度意义下)和若干相机内参数 K(参数个数取决于相机模型的选择),因此若相机已标定,则可以将求解未知参量数量降为5个。又由于基础矩阵的自由度为8[10],相机未标定的解法包括8点算法、7点算法和6点算法;相机已标定的解法为5点算法。Hartley[12]在1997年最早证明8点算法的可行性,算法的前提是特征点匹配的问题已解决。8点算法顾名思义是指完成一次相机姿态估计需要求解8对特征点坐标构成的8元一次方程组。之后,研究者逐渐在8点算法基础上增加约束条件从而减少算法依赖的最少特征点数,依次提出了7点算法、6点算法。2003年Hartley等人[10]提出7点算法,它与8点算法相似,不同的是增加了基础矩阵和本质矩阵是奇异矩阵的约束条件,因而将方程组元数由8降低为7。2005年Stewenius等人[13]提出6点算法,2012年Kukelova等人[14]提出基于多项式特征值的6点算法。以上都是解决未标定最小问题的算法,对于已标定相机的解决方案,即5点算法的经典算法包括2004年Nister[15]提出的5点算法、2006年Li等人[16]提出的更易于实现的Nister的5点算法的改进版,以及2012年Kukelova等人[14]提出的基于多项式特征值的5点算法。这些传统图像处理方法是应用最广泛的相机姿态估计方法,它们都依赖于图像特征点的获取,而图像特征点的获取过程中会引入误差和噪声,因此在实际运算过程中一般使用随机采样一致性算法RANSAC[17]来抑制这些误差和噪声。Fischler等人[18]对这几种算法进行了系统的实验比较。

基础矩阵 F 分解的结果即为相机姿态 R 和 T 的解,也就完成了两张图像所对应的相机相对运动的估计。整段视频的运动估计可通过取连续帧进行帧间相对运动估计,最终整合为相机运动链。

在Buehler等人[9]提出最早的3D稳像方案后,由于当时(2001年)设备计算能力的限制以及适合3D稳像技术的图像修补算法尚未出现,3D稳像方案并没有迅速发展,直到2009年Liu等人[19]提出保存内容的3D视频稳像方案。该方案被视为最经典的3D稳像方案,它首先使用SFM算法恢复相机3D原始运动以及3D场景点云,然后进行自动或交互的运动平滑来获取稳定的相机运动,最后使用3D场景点云和参考帧图像完成运动平滑后的图像形变修复。为防止修复的图像产生畸变,作者提出了保存内容的形变技术(content-preserving warps),并将稀疏场景点的位移视为软约束条件。另外,Zhang等人[20]基于相机3D运动模型提出一种稳像方案,首先提取SIFT特征点(经RANSAC算法去除偏值),然后使用SFM算法恢复3D运动信息,之后将视频稳像问题建模为平滑和相似约束条件下的二次成本函数,通过平衡平滑性和相似性来获取视频的稳定性。

2.2.2　近似算法

除了SFM算法,还有的3D方案使用近似估计的方法来减少计算量。Wang等人[21]提出一种适用于手持设备的3D稳像方案,方案基于连续帧间平移与旋转运动量相对较小的事实,假设旋转角度极小,通过对旋转矩阵进行简化精简了相机3D运动的估计过程。

2.3 问题与挑战

3D 视频稳像技术可以通过非常稳定的 3D 相机路径来获取高质量的相机运动,能够有效解决平行视差问题并获取效果显著的稳像结果,但是其实际应用受到 3D 重建过程的限制。3D 重建一般通过 SFM 算法来完成。表 2 对当前 3D 稳像运动估计代表方案进行了总结。

表 2 3D 稳像运动估计总结

提出时间	方案	特征点方法	3D 重建方法	效率
2001	文献[9]	KLT	SFM	线下处理
2009	文献[19]	KLT	SFM	线下处理
2009	文献[20]	SIFT	SFM	线下处理
2009	文献[21]	SIFT	近似算法	线下处理

SFM 算法虽然发展很快,但目前能提供稳健、高效并且通用的方案仍然面临着极大挑战。

1)SFM 算法失败问题

视频拍摄时经常存在一些问题导致 SFM 算法无法获取足够的信息进行 3D 重建,尤其对于业余拍摄者拍摄的视频来说,这些问题包括以下几点:

a)缺少视差。很多情况下视频缺少视差或视差不足以进行 SFM 运算,或者拍摄场景较平坦如背景为颜色均匀的墙壁等。

b)相机变焦。区分相机变焦和相机向前运动是 SFM 算法面临的一个难题,如果误将变焦当做向前运动,则会得到错误的运动估计,因此很多方案会假设相机是已标定的或者内参固定的。

c)卷帘效应。大多目前消费级别的相机(如流行智能设备)都使用 CMOS 传感器,会面临卷帘效应,这严重地影响了 SFM 算法的实现。

2)SFM 算法效率问题

效率是 SFM 算法的一个大问题,因为 SFM 通常需要进行全局非线性优化,而且大多 SFM 的实现需要多次迭代而难以流水化。当然,目前也存在少量实时或接近实时的 SFM 系统[22,23],但它们都需要假定额外的条件是相机已标定,尚未能够在消费级别设备中普及应用。

以上原因成为 3D 稳像技术应用的主要阻碍因素,尤其在当前主流智能设备计算能力相对较弱的现状下,研究人员因此试图寻求 3D 方案之外能够解决平行视差问题、稳像效果接近 3D 方案的技术。

3 传感器辅助稳像运动估计

近年来,随着微机电系统(micro-electro-mechanical system, MEMS)传感器在智能设备中的逐渐普及且精度越来越高,出现了使用 MEMS 传感器来进行相机运动估计以完成视频稳像的方案。常用的传感器设备包括陀螺仪、加速度传感器、重力传感器等,这些也是当前智能设备中都会配备的运动传感器。

3.1 运动模型

与 3D 稳像运动估计不同的是,传感器辅助运动估计都采用相机 3D 旋转运动模型,而忽略平移参量的恢复。图 4 为相机 3D 旋转运动模型。相机旋转运动模型中相机运动时中光心位置不变,在该模型下,世界坐标系中一点 X 以及齐次坐标系下表示的图像坐标 x 的映射关系为

$$x = KX \text{ 以及 } X = \lambda K^{-1}x \tag{10}$$

其中:K 为相机内参,定义如第 2 章所述;λ 为未知尺度因子,表示在该相机模型下,图像点坐标的来源被映射为一条射线。

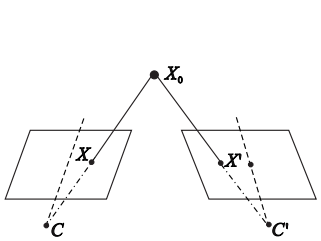

图3 SFM算法的双视角视图

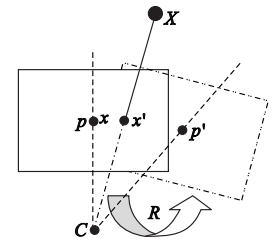

图4 相机旋转运动模型

在此旋转运动模型下,任意两个时刻下,世界坐标系中一点 X 到图像的投影点为

$$x = KRX \text{ 以及 } x' = KR'X \tag{11}$$

式中:R 与 R' 为 3×3 旋转矩阵,分别表示相机两个时刻的旋转参数。角速率传感器(一般为陀螺仪)输出数据为各轴角速率值,设备旋转角度可以直接通过角速率值得到。对任何一轴,设角速率传感器输出角速率为 ω_i,对应的采样时间为 Δt_i,则从 m 到 n 时刻设备的旋转角度 $\Delta\Theta_i$ 为

$$\Delta\Theta_i = \sum_{i=m}^{n}\omega_i\Delta t_i \tag{12}$$

用 $k=(x,y,z)$ 表示三轴的旋转向量,x、y、z 的值由式(12)计算得出;$\Theta=\sqrt{x^2+y^2+z^2}$ 表示总体旋转量,旋转矩阵 R 可以由 Rodrigues 公式来表示:

$$R = I + [R]_\times \sin\theta + [k]_\times^2(1-\cos\theta) \tag{13}$$

其中:I 为单位矩阵,$[k]_\times$ 表示 k 的叉积矩阵:

$$[k]_\times = \begin{bmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{bmatrix}$$

酉矩阵 $R\in SO(3)$,表示相机的旋转运动。

在传感器辅助方案中,采用 3D 旋转模型主要有三方面的原因:

a)传感器运动估计精度问题。角速度传感器(如陀螺仪)获取设备旋转角速率,通过一次时间积分可以估计设备旋转角度;加速度传感器需要二次积分运算来获取位移信息,然而二次积分带来的误差太大,无法得到精确的平移参量。

b)即便获取到了精确的平移参量,没有恢复深度信息的条件下平移参量无法转换为像素平移量,因而无法用于稳像应用。

c)造成视频抖动的主要原因是旋转,平移的影响很小。如图 5 所示,对于距离相机 3 m 外远的场景,1 mm 的相机平移运动造成的像素平移小于 1 个像素[24]。Whyte 等人[25]也解释说明了旋转为抖动主要因素的原因。

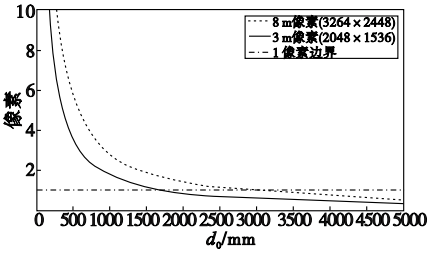

图5 不同景深条件下1 mm相机平移量导致的像素平移量
(相机焦距28 mm或35 mm均是此结果)

3.2 传感器标定

使用传感器辅助方法进行稳像运动估计时,由于运动传感器与图像传感器在空间和时间上都并不完全同步,所以通常情况下为了保证运动估计数据的准度和可靠性,需要进行传感器

标定。

a)空间标定。运动传感器的位置与光轴的位置并不吻合,存在细微的差别,因此使用传感器数据来直接表示相机的运动是有一定误差的。空间位置的标定可以使用 Kelly 等人[26]的方法。对于目前的传感器辅助方法而言,由于流行智能设备相机和传感器相对位置很近,而且用于稳像目的应用的时候,如此细微的误差并不影响稳像结果。所以当前主流传感器辅助方案都略过了传感器位置的标定步骤。

b)时间标定。运动传感器和图像传感器属于不同部件,两者数据时间戳之间存在误差,需要同步。同步的方法是使用特征匹配点作为真实值,然后根据旋转模型下匹配点对的映射关系使用传感器进行特征点的匹配估计,找到与真实值误差最小的参数值作为时间标定量。Karpenko 等人[27]使用 RANSAC 算法对 SIFT 特征点进行偏值滤除,将剩余的特征点对作为真实值,并建立重投影误差方程,将标定问题转换为最小值问题;值得注意的是,此方法同时还可以标定相机焦距。Hanning 等人[28]使用 KLT 跟踪算法跟踪连续帧间的若干像素点,然后通过"反跟踪"的方法去除偏值,将剩余的帧间匹配点对作为真实值,最后使用与 Karpenko 等人[27]类似的方法,利用传感器数据在旋转模型下对这些匹配点对进行旋转运算,并取与真实值误差最小的参数作为标定量。Bell 等人[29]使用了标定模板,这样做的目的是让标定算法能容易跟踪特征点对,其原理与 Hanning 等人[28]的方法类似,也是最小化重投影误差,并寻找使其函数值最小的参数值作为标定量。

3.3　深度摄像头辅助方法

MEMS 传感器辅助方法简化了相机旋转运动获取过程,然而它们实质上还是 2D 运动估计方法,因此它们没有恢复平移运动。应用于视频稳像的时候,由于它们忽略了景深问题,依然无法避免平行视差问题。

Sun 等人[30]引入深度摄像头来解决视频 3D 稳像问题,由于深度摄像头输出的深度信息有噪声、深度图像不完整且分辨率低,作者将深度图像与原图像结合进行相机 3D 运动估计,即在 2D 特征点基础上使用深度信息完成相机 3D 运动估计。

3.4　问题与挑战

传感器辅助方法能够简化相机运动估计,流行方案是使用 MEMS 传感器建立相机旋转运动模型,解决视频抖动的主要因素,而忽略次要因素。这种传感器辅助方法无须使用特征点匹配或跟踪技术,且传感器数据可实时获取,减少了稳像工作的计算量,对于当前流行智能设备很适用。表 3 是对当前代表性传感器辅助稳像运动估计方案的总结。

表 3　传感器辅助稳像运动估计总结

时间	方案	技术传感器	运动模型	效率
2011	文献[27]	陀螺仪	3D 旋转	实时(30 fps) iPhone4,GPU 辅助
2011	文献[28]	陀螺仪 加速度仪	3D 旋转	近实时(10 fps) iPhone4,720P
2012	文献[30]	深度摄像头	3D	线下
2014	文献[29]	陀螺仪	3D 旋转	实时,iPhone5,整合到 图像获取流水中

然而这种方法也存在着问题与挑战:

a)模型选择问题。这种方法通常只使用相机旋转模型,忽略平移运动。此模型不能恢复景深,因而并未解决平行视差问题。但对于当前视频应用来说,经常有近距离场景拍摄的情况,如室内场景或近距离移动自拍等,此时平移运动带来的相机抖动并不能简单忽略,如图 5 所示。而这种情况下只使用旋转模型建模并不能取得很好的稳像效果。

b)传感器数据可靠性问题。随着时间积累,传感器会存在累积误差和漂移,对于 MEMS 传感器,两者都不可忽略。对于长时间视频来说,使用传感器进行运动估计在消费级别的设备中是一个挑战。

另外还有一种思路,使用深度摄像头结合图像处理方法来获取完成的相机 3D 运动估计,从而降低 3D 运动估计的复杂度,并解决平行视差问题。然而这种方法需要额外的深度摄像头,不适用于实际应用。

4　新兴 2D 稳像运动估计

3D 稳像方案和传感器辅助方案各有优缺点,前者稳像效果最好,而计算复杂,不适合当前智能设备应用;后者计算简单,但没有解决平行视差问题,无法达到与 3D 方法可比的稳像效果。近年来研究人员试图使用 2D 方法来解决平行视差问题,达到接近 3D 稳像的效果,并陆续提出了效果不错的解决方案。这些方案不尽相同,但共同点是将相机 3D 运动重建过程的约束条件放松,绕过复杂的 3D 重建操作,使用更简单的方法来解决平行视差问题。本文将这类稳像方案称为新兴 2D 稳像方法,并根据其采用的策略对它们进行了分类:特征轨迹法和多路径法。

4.1　特征轨迹法

特征轨迹法将相机 3D 运动重建问题的约束放宽为处理相应 2D 图像空间特征点轨迹的问题。当然此方法的前提是稳健的特征轨迹获取问题。

KLT 跟踪算法是特征轨迹法中主流应用的算法。特征轨迹法用于稳像时,在获取特征轨迹后,有各种不同处理特征轨迹的方法来达到稳像目的。Liu 等人[31]提出一种子空间稳像法,方案的理论基础是移动相机短时间拍摄的 3D 刚体场景图像的运动轨迹矩阵可以近似表示在低维度子空间内[32],先使用 KLT 跟踪方法来建立稀疏场景点的 2D 轨迹矩阵,并对上述矩阵进行了一个移动分解变换(moving factorization)来有效地找到对输入运动的一个时变子空间近似,这个时变子空间近似将运动估计局部表示为两部分的乘积,一个称为特征轨迹的基向量,一个将特征点描述为这些特征轨迹线性组合的系数矩阵;然后对特征轨迹进行运动平滑,将平滑后的特征轨迹与原始系数矩阵重新相乘来得到平滑后的输出轨迹,最后后者可以交给渲染方案完成视频稳像。Hsu 等人[33]认为稳定视频应该满足两个条件,一是平滑的运动轨迹,二是连贯的帧间过渡。传统方法大多只解决第一个问题,它们需要合适的、特定场景的参数设置,无法通用于不同场景。针对此问题,作者使用 Harris 角点提取法来提取特征点,并使用 KLT 跟踪器来进行特征点匹配,最终提供一种基于单应一致性的算法来直接提取最佳平滑轨迹并使得帧间过渡均匀分布。Goldstein 等人[34]提出一种利用极线几何进行稳像的方案,首先使用 KLT 法来跟踪特征点(假设前提是获取的 KLT 点都位于背景),然后使用 8 点算法计算包含了相机运动信息的基础矩阵,再用 Gaussian 滤波器对轨迹进行平滑,最后使用对极点转移法[35]来确定平滑后特征点的位置。这种对极点转移法将复杂的 3D 重建过程转换为利用像素点和极线的几何关系进行运动估计。Ryu 等人[36]认为在相机全局运动估计中,传统方法多数使用累加的

方法来获取,导致累加误差随时间增加,不适用于长时间视频稳像。针对此问题作者提出一种仅依赖特征轨迹的稳像方案,首先使用KLT跟踪器来跟踪特征点轨迹,然后使用Kalman滤波器来生成平滑特征点轨迹,两者之差即为需要补偿的运动,最后使用双线性插值法,在2D仿射模型下,根据修正运动来进行图像拼接生成最终稳定的图像。之后作者又实现了此方案的实时稳像[37]。

除了KLT方法,还有使用其他方法获取稳健特征轨迹的方法。Lee等人[38]使用特征轨迹提取算法融合了SIFT流方法[7]和粒子视频法[8],前者利用空域运动的连续性减少误匹配,后者利用轨迹时域运动的相似性保证大范围特征跟踪。

上述特征轨迹法需要长时间的轨迹跟踪(通常20帧以上),Wang等人[39]提出一种时空优化法,能够使用短时间特征轨迹完成稳像。方案用贝塞尔曲线表示特征点轨迹,从而将稳像问题转换为平滑特征曲线并避免视觉畸变的时空优化问题。作者提出,贝塞尔曲线表示法能够有效平滑特征轨迹并且减少优化问题中的变量数目,提高稳像效率。

4.2 多路径法

与特征轨迹法不同,Liu等人[40]将"as-similar-as-possible"的思想[77]引入到相机运动估计中以提高运动估计的鲁棒性,方案使用更有效的2D相机运动模型来替代复杂的3D运动模型,提出一种多重的、时空可变的相机路径模型,让不同位置可以有独自的相机路径,但与简单2D方法不同,每个块的运动估计采用基于形变的运动模型。

4.3 问题与挑战

新兴2D稳像方案绕过复杂的3D重建操作,将相机3D运动重建过程的约束条件放松,目的是为了用接近2D方案的计算量来近似达到3D方案的稳像效果。表4对当前代表性新兴2D稳像运动估计方案进行了总结。

表4 新兴2D稳像运动估计方案总结

时间	方案	特征点技术	效率
2009	文献[38]	SIFT流+粒子视频	线下
2011	文献[31]	长KLT(50帧以上)	近实时(640×360,4 fps) 3.16 GHz Intel双核CPU
2012	文献[37]	长KLT	实时(640×480,39.54 fps) 2.5 GHz CPU
2012	文献[33]	长KLT	近实时(640×480,10 fps)
2012	文献[34]	长KLT(20帧以上)	近实时(1280×720, 4 fps) Intel i7 3.07 GHz CPU
2013	文献[39]	Bezier曲线表示轨迹,无须长特征轨迹跟踪	线下(作者声明若有GPU辅助可实现实时) Core i7 3.0 GHz CPU
2014	文献[40]	SURF	近实时(2.5 fps) Intel i7 3.2 GHz 4核CPU

这种方案当前主流的思路是使用对特征轨迹进行直接处理,因此其实际应用与特征轨迹获取质量相关,而特征轨迹的获取会受以下因素的限制:

a)遮挡。特征轨迹跟踪过程中如果遇到遮挡物体,那么特征跟踪会被中断。

b)相机快速运动。相机运动速度较快,可能导致被跟踪特征很快从场景消失,此时特征轨迹跟踪失败。

c)景深变化大时效果差。这是2D稳像方案的通病。

Wang等人[39]提出的时空优化法可以缓解对长特征轨迹的依赖,Liu等人[40]的多路径法不依赖特征轨迹,但它们还是要依赖背景特征点,同样受到上述限制。

5 总结与展望

5.1 总结

本文对3D法、传感器辅助法、新兴2D法等新兴数字稳像中相机运动估计技术进行了综述。与早期2D稳像运动估计技术不同,新兴运动估计技术采用更加稳健有效的运动模型,依赖更加可靠的技术。表5是对视频稳像应用中相机运动估计技术的总结。

表5 视频稳像应用中相机运动估计技术总结

方法		技术依赖	优点	缺陷
3D		a)特征点 b)SFM	解决平行视差	a)难以实时 b)依赖特征点
传感器辅助		传感器	a)可实时 b)不依赖特征点	a)未解决平行视差 b)传感器漂移
新兴2D	特征轨迹	特征轨迹	a)缓解平行视差 b)近实时	依赖长特征轨迹
	多路径	特征点		依赖特征点

在早期2D视频稳像技术之后,这三类新兴数字视频稳像技术依次出现。3D方法的出现是为了解决2D方法无法解决的平行视差问题以提供更佳的稳像效果,然而3D方法计算量大,对于当前主流智能设备计算能力来说只能进行线下处理,难以实时应用。传感器辅助方法能够有效解决计算速度问题,且不依赖特征点技术,运动估计可以实时完成,但由于传感器估计平移参量误差过大,所以当前主流传感器辅助方案都采用了旋转运动模型,忽略平移影响。这一方面无法解决平行视差问题,另一方面简单忽略平移分量对于近距离场景稳像效果不佳;此外当前流行消费级别的MEMS传感器普遍存在漂移问题,不适合长时间视频稳像。最近几年出现的新方法直接处理特征轨迹来完成稳像,特征轨迹代表了相机运动状况,这种方法能够缓解平行视差问题,且处理速度能够接近实时;缺点在于通常依赖长特征轨迹,在遮挡、相机快速运动或者景深变化较大时无法提供稳健的效果。当前的数字视频稳像中相机运动估计技术围绕着平行视差问题,在运算效率和稳像效果这一对矛盾中发展。

5.2 展望

综合分析视频稳像中相机运动估计技术的发展可见,各种方案在运算效率和稳像效果这对矛盾中逐渐更新,发展到当前新兴2D稳像技术,实现了速度与效果某种程度上的折中。本文展望将来视频稳像中相机运动估计技术的相关研究热点和难点如下。

1)实时应用方面

a)新型传感器的研究。出现适用于智能设备的获取平移参量的传感器,这样可以极大简化相机3D运动重建过程,加快3D稳像速度。现有的平移参量获取传感器如加速度传感器等受限于其在进行位移估计时对噪声影响的脆弱性,无法满足视频稳像应用中3D运动估计这样高精度需求应用的要求。如果将来能出现精度更高的加速度传感器或距离传感器,或者其他新型传感器,实时3D稳像应用在智能设备的应用将更加现实。

b)实时SFM算法的研究。目前虽然已经存在一些实时SFM系统,但并不适用于当前主流智能设备。3D运动重建的关键步骤是SFM算法,研究适合当前主流智能设备的实时SFM算法,使得实时3D运动重建能够更加实用。

2)运动估计精度方面

a)传感器漂移消除。当前传感器辅助方法一般假设视频

时间较短,传感器漂移影响不大。针对视频稳像应用,如何消除传感器长时间使用时的误差和漂移,对于使用传感器辅助方法有重要意义。

b)背景特征提取。无论特征点还是特征轨迹,当前方法都没有很好地区分背景和前景。一些方案使用经验法大致划分背景和前景或者用 RANSAC 法来剔除偏值,但都没有从根本上解决背景特征的提取。优秀背景特征提取算法能够为依赖特征的运动估计提供精确的特征,提高运动估计精度。

参考文献:

[1] Rawat P, Singhai J. Review of motion estimation and video stabilization techniques for hand held mobile video[J]. International Journal of Signal & Image Processing,2011,2(2):159-168.

[2] Amisha P. A survey on video stabilization techniques[J]. International Journal of Engineering Sciences & Research Technology,2015,4(2):338-342.

[3] Lowe D. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision,2004,60(2):91-110.

[4] Bay H, Tuytelaars T, Gool L V. SURF: speeded up robust features [J]. Computer Vision & Image Understanding,2006,110(3):404-417.

[5] Lucas B, Kanade T. An iterative image registration technique with an application to stereo vision[C]//Proc of International Joint Conference on Artificial Intelligence. 1981.

[6] Shi Jianbo, Tomasi C. Good features to track[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1994:593-600.

[7] Liu Ce, Yuen J, Torralba A, *et al*. SIFT flow: dense correspondence across different scenes[C]//Proc of the 10th European Conference on Computer Vision. 2008:28-42.

[8] Sand P, Teller S. Particle video: long-range motion estimation using point trajectories[J]. International Journal of Computer Vision, 2006,2(1):2195-2202.

[9] Buehler C, Bosse M, McMillan L. Non-metric image-based rendering for video stabilization[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2001:609-614.

[10] Hartley R, Zisserman A. Multiple view geometry in computer vision [M]. 2nd ed. Cambridge:Cambridge University Press, 2003.

[11] Faugeras O. Three dimensional computer vision: a geometric viewpoint[M]. Boston:MIT Press, 1993.

[12] Hartley R. In defense of the eight-point algorithm[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,1997,19(6):580-593.

[13] Stewenius H, Nister D, Kahl F, *et al*. A minimal solution for relative pose with unknown focal length[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2005:789-794.

[14] Kukelova Z, Bujnak M, Pajdla T. Polynomial eigenvalue solutions to minimal problems in computer vision[J]. IEEE Trans on Software Engineering,2012,34(7):1381-1393.

[15] Nister D. An efficient solution to the five-point relative pose problem [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004,26(6):756-770.

[16] Li Hongdong, Hartley R. Five-point motion estimation made easy [C]//Proc of the 18th International Conference on Pattern Recognition. 2006:630-633.

[17] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM,1981,24(6):381-395.

[18] Brückner M, Bajramovic F, Denzler J. Experimental evaluation of relative pose estimation algorithms[C]//Proc of the 3rd International Conference on Computer Vision Theory and Applications. 2008:431-438.

[19] Liu Feng, Gleicher M, Jin Hailin, *et al*. Content-preserving warps for 3D video stabilization[J]. ACM Trans on Graphics,2009,28(3):341-352.

[20] Zhang Guofeng, Hua Wei, Qin Xueying, *et al*. Video stabilization based on a 3D perspective camera model[J]. Visual Computer, 2009,25(11):997-1008.

[21] Wang J M, Chou H P, Chen S W, *et al*. Video stabilization for a hand-held camera based on 3D motion model[C]//Proc of the 16th IEEE International Conference on Image Processing. 2009: 3477-3480.

[22] Nistér D, Naroditsky O, Bergen J. Visual odometry[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004:652-659.

[23] Davison A, Reid I, Molton D, *et al*. MonoSLAM: real-time single camera SLAM[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2004,26(6):1052-1067.

[24] Rajagopalan A N, Chellappa R. Motion deblurring: algorithms and systems[M]. Cambridge:Cambridge University Press, 2014.

[25] Whyte O, Sivic J, Zisserman A, *et al*. Efficient, blind, spatially-variant deblurring for shaken images[M]. [S.l.]:Cambridge University Press, 2014.

[26] Kelly J, Sukhatme G S. Fast relative pose calibration for visual and inertial sensors[J]. Springer Tracts in Advanced Robotics,2009, 54(1):515-524.

[27] Karpenko A, Jacobs D, Baek J, *et al*. Digital video stabilization and rolling shutter correction using gyroscopes,CTSR 2011-03 [R]. Stanford:Stanford University,2011.

[28] Hanning G, Forslow N, Forssen P E, *et al*. Stabilizing cell phone video using inertial measurement sensors[C]//Proc of IEEE International Conference on Computer Vision. 2011:1-8.

[29] Bell S, Troccoli A, Pulli K. A non-linear filter for gyroscope-based video stabilization[C]//Proc of the 13th European Conference on Computer Vision. [S.l.]:Springer,2014:294-308.

[30] Sun Jian. Video stabilization with a depth camera[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2012:89-95.

[31] Liu Feng, Gleicher M, Wang Jue, *et al*. Subspace video stabilization [J]. ACM Trans on Graphics,2011,30(1):623-636.

[32] Tomasi C, Kanade T. Shape and motion from image streams under orthography: a factorization method [J]. International Journal of Computer Vision,1992,9(2):137-154.

[33] Hsu Y F, Chou C C, Shih M Y. Moving camera video stabilization using homography consistency[C]//Proc of the 19th IEEE International Conference on Image Processing. 2012:2761-2764.

[34] Goldstein A, Fattal R. Video stabilization using epipolar geometry [J]. ACM Trans on Graphics,2012,31(5):573-587.

[35] Laveau S, Faugeras O. 3D scene representation as a collection of images[C]//Proc of the 12th International Conference on Pattern Recognition. 1994:689-691.

[36] Ryu Y G, Roh H C, Chung M J. Long-time video stabilization using point-feature trajectory smoothing[C]//Proc of IEEE International Conference on Consumer Electronics. 2011:189-190.

[37] Ryu Y G, Chung M J. Robust online digital image stabilization based on point-feature trajectory without accumulative global motion estimation[J]. IEEE Signal Processing Letters,2012,19(4):223-226.

[38] Lee K Y, Chuang Y Y, Chen Bingyu, *et al*. Video stabilization using robust feature trajectories[C]//Proc of IEEE International Conference on Computer Vision. 2009:1397-1404.

[39] Wang Yushuen, Liu Feng, Hsu P S, *et al*. Spatially and temporally optimized video stabilization [J]. IEEE Trans on Visualization & Computer Graphics,2013,19(8):1354-1361.

[40] Liu Shuaicheng, Yuan Lu, Tan Ping, *et al*. Bundled camera paths for video stabilization[J]. ACM Trans on Graphics,2013,32(4).