基于运动目标三维轨迹重建的视频序列同步算法

王 雪 [1] SHI Jian-Bo[2] PARK Hyun-Soo[2] 王 庆 [1]

摘 要 提出一种利用运动目标三维轨迹重建的视频时域同步算法. 待同步的视频序列由不同相机在同一场景中同时拍摄得到, 对场景及相机运动不做限制性约束. 假设每帧图像的相机投影矩阵已知, 首先基于离散余弦变换基重建运动目标的三维轨迹. 然后提出一种基于轨迹基系数矩阵的秩约束, 用于衡量不同序列子段间的空间时间对准程度. 最后构建代价矩阵, 并利用基于图的方法实现视频间的非线性时域同步. 我们不依赖已知的点对应关系, 不同视频中的跟踪点甚至可以对应不同的三维点, 只要它们之间满足以下假设: 观测序列中跟踪点对应的三维点, 其空间位置可以用参考序列中所有跟踪点对应的三维点集的子集的线性组合描述, 且该线性关系维持不变. 与多数现有方法要求特征点跟踪持续整个图像序列不同, 本文方法可以利用长短不一的图像点轨迹. 本文在仿真数据和真实数据集上验证了提出方法的鲁棒性和性能.

关键词 视频同步, 独立运动相机, 运动恢复非刚性结构, 轨迹基, 秩约束

引用格式 王雪, Shi Jian-Bo, Park Hyun-Soo, 王庆. 基于运动目标三维轨迹重建的视频序列同步算法. 自动化学报, 2017, **43**(10): 1759−1772

DOI 10.16383/j.aas.2017.c160584

Synchronization of Video Sequences Through 3D Trajectory Reconstruction

WANG Xue[1] SHI Jian-Bo[2] PARK Hyun-Soo[2] WANG Qing[1]

Abstract We present an algorithm for synchronization of an arbitrary number of videos captured by cameras independently moving in a dynamic 3D scene. Assuming the 3D spatial poses of the cameras are known for each frame, we first reconstruct the 3D trajectory of a moving point using the trajectory basis-based method. The trajectory coefficients are computed for each sequence separately. Point correspondences across sequences are not required, or even it is possible to track different points in different sequences, only if every 3D point tracked in the second sequence is a linear combination of subsets of the 3D points tracked in the first sequence. Then we propose use a robust rank constraint of the coefficient matrices to measure the spatio-temporal alignment quality for every feasible pair of video fragments. Finally, the optimal temporal mapping is found using a graph-based approach. Our algorithm can use both short and long feature trajectories, and it is robust to mild outliers. We verify the robustness and performance of the proposed approach on synthetic data as well as on challenging real video sequences.

Key words Video synchronization, independently-moving cameras, non-rigid structure from motion, trajectory basis, rank constraint

Citation Wang Xue, Shi Jian-Bo, Park Hyun-Soo, Wang Qing. Synchronization of video sequences through 3D trajectory reconstruction. *Acta Automatica Sinica*, 2017, **43**(10): 1759−1772

视频同步, 又称视频对准, 是计算机视觉领域中一个重要的基础问题. 根据同步方式不同, 现有的视频同步方法可分为基于外触发脉冲的同步和基于视频图像序列中视觉特征的同步. 其中, 基于外触发脉冲的同步技术作用在相机端, 多用于控制多相机同步实现高速图像采集存储, 硬件成本较高; 基于视觉特征的同步算法通过分析图像序列中的同步线索实现多个视频间的时域对齐, 可用于行为识别、基于内容的视频检索及非刚性结构三维重建等视觉任务. 本文主要讨论基于视觉特征的视频同步方法, 其常规思路是联合优化图像序列间的空间和时间对准. 空间对准多指在待同步帧对的二维图像或三维相机坐标系下计算某种几何变换, 因此依赖精确的特征提取和匹配. 时域对准通过估算图像序列间的线性或非线性时域映射以获得最优的空间对准.

为了降低问题求解的复杂度, 研究者们提出各种假设来减少待估计参数的数量. 假设静止相机或联合运动相机, 则空间变换关系恒定不变. 现有方法多在二维图像坐标系中估算几何变换, 如单应[1−2]、

收稿日期 2016-08-10 录用日期 2017-03-02
Manuscript received August 10, 2016; accepted March 2, 2017
国家自然科学基金 (61531014) 资助
Supported by National Natural Science Foundation of China (61531014)
本文责任编委 黄庆明
Recommended by Associate Editor HUANG Qing-Ming
1. 西北工业大学计算机学院 西安 710072 中国 2. 宾夕法尼亚大学工程与应用科学学院 费城 19104 美国
1. School of Computer Science, Northwestern Polytechinical University, Xi′an 710072, China 2. School of Engineering and Applied Science, University of Pennsylvania, Philadelphia PA 19104, USA

仿射变换[3]、射影变换[2−6] 等, 并利用重投影误差来度量空间对准的程度. 由于求解基础矩阵对噪声敏感, Rao 等[7] 针对透视模型提出一种基于对极几何的秩约束. 进一步地, Tresadern 等[8] 综合单应、仿射变换和射影变换三种几何模型下的亏秩条件, 提出了统一的算法框架. 这类方法多用于窄基线条件下图像点轨迹及对应关系已知的视频对准. 为克服宽基线条件下特征匹配难的问题, 文献 [9−10] 提出一种弱假设, 即观测序列中像点的空间位置可以用参考序列中像点子集空间位置的线性组合描述, 且该线性关系维持不变. 这样, 算法不再依赖已知的像点对应关系, 各序列中的像点甚至可以对应不同的空间点. 缺点是该方法仅适用于固定仿射相机间的常量偏移时域同步. 假设相机沿相似轨迹运动[11−15], 则对应帧的相机坐标系可近似认为原点重合, 仅对应坐标轴间存在较小的旋转角度. 因此, 内容上越相似的两幅图像帧, 其时域同步的可能性越高. 基于这种思想, Wang 等[16] 提出了一种基于 SIFT 特征点匹配的视频同步算法, 并提供友好的交互界面允许用户手动设置入点、出点或剪辑标记来同步多机位序列. 值得一提的是, 这种交互方式也是众多视频编辑工具实现多机位序列同步的方式, 此外还包括使用基于音频的同步来准确对齐剪辑, 例如 Edius、Premiere 等. 假设时域映射关系为线性, 例如常量偏移模型[4, 9−10], 或者一维仿射模型[1−2, 5, 7], 则时域映射关系可以用一个简单的参数化模型 $t_r = \rho t_o + \Delta$ 来描述, 其中 t_r 和 t_o 分别表示参考序列和观测序列中的图像帧索引, ρ 为两序列的帧率比, Δ 为帧索引偏移常量.

联合空间和时间对准能够提高系统的鲁棒性, 但这类方法面临两个主要的挑战. 1) 对于独立运动相机和包含多个运动目标的三维动态场景 (图 1) 来说, 基于几何变换的空间对准是十分困难的. 2) 考虑到丢帧、时域连续性等问题, 线性时域映射不再满足需求, 而非线性时域映射的估算会增加现有算法求解的复杂度.

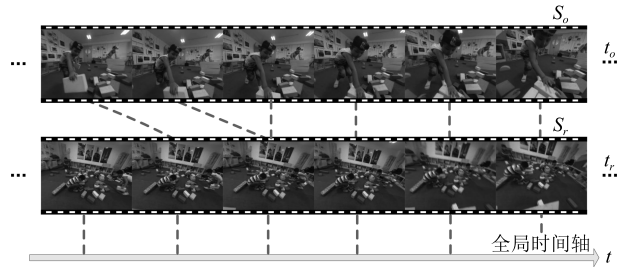


图 1 待同步的第一人称视角图像序列

Fig. 1 Video sequences captured by first-person cameras

针对独立运动相机, Tuytelaars 等[17] 提出一种基于反向投影的同步方法, 用于弱透视投影模型下视频间的偏移常量时域同步, 通过将图像点反向投影到空间中的一条线, 从而将时域对准问题转换为在空间中寻找相交或距离最短的直线问题. Lei 等[18] 基于三视图几何约束建立时间轴图, 用于多个图像序列间的偏移常量时域同步. 这些方法都假设线性时域映射关系, 并且依赖精确的特征点跟踪和匹配, 因此在实际应用中受到限制. Dexter 等[19] 利用图像序列的自相似矩阵为每帧图像计算时域自适应的特征描述, 通过时域运动特征匹配实现图像序列对的非线性时域同步. 该算法利用场景中静止的背景点估计全局运动实现相机自运动补偿, 仅适用于平面场景或远视角.

本文提出一种针对独立运动相机的视频时域对准算法, 其主要想法是利用空间中运动目标的轨迹 (时间空间特征) 来同步图像序列. 取代传统的帧对空间对准, 我们在时间轴上从单帧的图像点延长到持续整个子序列段的图像点轨迹, 通过三维运动轨迹重建, 提出一个基于轨迹基系数的秩约束用于度量任意子序列对的空间对准程度. 随后, 本文提出一种基于图的非线性时域对准算法, 用于搜索最优时域映射关系. 进一步地, 我们将该算法从双序列对准扩展到多序列的情况. 最后, 本文在仿真数据和真实第一视角数据集上进行验证.

1 双序列时域对准

1.1 基于三维轨迹重建的线性时域对准

当两个或多个相机相对静止, 或者沿相似轨迹运动, 空间中同一个运动点在不同相机拍摄的图像序列中的二维轨迹是相似的. 然而, 若相机自由运动, 则上述结论不再成立. 为了消除相机自运动对目标运动分析的影响, 我们可以在二维图像空间中尝试相机自运动补偿, 或者将二维观测反投影回三维空间中进行运动分析. 后者的优点是受场景及相机运动限制较小. 本文采取后一种方法, 首先利用文献 [20−21] 提出运动目标三维轨迹重建算法恢复运动目标的三维轨迹.

令摄像机在第 t 帧的投影矩阵为 $P^{(t)} \in \mathbf{R}^{3\times 4}$, 按透视射影变换将空间中一点 $X^{(t)} = [X^{(t)}\ Y^{(t)}\ Z^{(t)}]^{\mathrm{T}}$ 投影到二维像点 $\boldsymbol{x}^{(t)} = [x^{(t)}\ y^{(t)}]^{\mathrm{T}}$, 根据相机成像模型有

$$\begin{bmatrix}\boldsymbol{x}^{(t)}\\1\end{bmatrix} \simeq P^{(t)}\begin{bmatrix}X^{(t)}\\1\end{bmatrix} \quad \text{或}$$
$$\begin{bmatrix}\boldsymbol{x}^{(t)}\\1\end{bmatrix}_{\times} P^{(t)}\begin{bmatrix}X^{(t)}\\1\end{bmatrix} = \mathbf{0} \tag{1}$$

其中, $[\cdot]_\times$ 是向量叉乘的反对称矩阵表示. 令 $\vec{\boldsymbol{x}}^{(t)} = [\boldsymbol{x}^{(t)\mathrm{T}}\ 1]^{\mathrm{T}}$ 为二维像点 $\boldsymbol{x}^{(t)}$ 的齐次坐标, 可以推导出 $Q^{(t)} = ([\vec{\boldsymbol{x}}^{(t)}]_\times P^{(t)}_{1:3})_{1:2}$, $\boldsymbol{q}^{(t)} = (-[\vec{\boldsymbol{x}}^{(t)}]_\times P^{(t)}_4)_{1:2}$, 其中, $P^{(t)}_{1:3}$ 为 $P^{(t)}$ 的前三列, $P^{(t)}_4$ 为第四列, $(\cdot)_{1:2}$ 为 $(\cdot)$ 的前两行. 累计 F 帧连续图像序列上的观测值可以得到如下方程[21]

$$\begin{bmatrix} Q^{(1)} & & \\ & \ddots & \\ & & Q^{(F)} \end{bmatrix} \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(F)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{q}^{(1)} \\ \vdots \\ \boldsymbol{q}^{(F)} \end{bmatrix} \quad \text{或}$$

$$QX = \boldsymbol{q} \tag{2}$$

其中, $X = [X^{(1)\mathrm{T}}, \cdots, X^{(F)\mathrm{T}}]^{\mathrm{T}}$ 为重建的运动目标三维轨迹. 假设轨迹可以采用离散余弦变换 (Discrete cosine transform, DCT) 基以较少的低频分量线性表示.

$$X = [X^{(1)\mathrm{T}}\ \cdots\ X^{(F)\mathrm{T}}]^{\mathrm{T}} \approx \boldsymbol{\Theta}_1\beta_1 + \cdots + \boldsymbol{\Theta}_{3K}\beta_{3K} = \boldsymbol{\Theta}\boldsymbol{\beta} \tag{3}$$

其中, $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1 \cdots \boldsymbol{\Theta}_{3K}] \in \mathbf{R}^{3F\times 3K}$ 为轨迹基矩阵, $\boldsymbol{\beta} = [\beta_1\ \cdots\ \beta_{3K}]^{\mathrm{T}} \in \mathbf{R}^{3K}$ 为组合系数, K 为每个维度上基的数量. 将式 (3) 导入式 (2), 当 $2F \geq 3K$ 时, 可以得到具有最小二乘解的超定系统

$$Q\boldsymbol{\Theta}\boldsymbol{\beta} = \boldsymbol{q} \tag{4}$$

如果该目标的三维轨迹同时被另外一台摄像机捕捉到, 类似地, 我们可以得到

$$\hat{Q}X = \hat{Q}\boldsymbol{\Theta}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{q}} \tag{5}$$

为了提高公式的易读性, 本文用相同符号加角号表示与第二个图像序列相关. 由于 $\boldsymbol{\Theta}$ 是正交矩阵, 基系数 $\boldsymbol{\beta}$ 和 $\hat{\boldsymbol{\beta}}$ 理论上应相同. 若空间中有 P 个点同时被两个摄像机看到, 当满足不等式 $3K \geq 2P$ 时, 基系数矩阵 $M = [\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_P\ \hat{\boldsymbol{\beta}}_1 \cdots \hat{\boldsymbol{\beta}}_P] \in \mathbf{R}^{3K\times 2P}$ 的秩最大不超过 P. 若两序列同步, M 的秩减小, 相反, 若两序列不同步, 则 M 的秩增加. 因此, 我们可以通过比较不同偏移量下 M 的秩, 来估算时域映射关系. 值得注意的是, P 不是一个上确界, 这取决于 P 个点间的刚性约束关系. 无论如何, 基系数矩阵 M 的秩在同步时的下降量不低于非同步时的下降量.

令 $S_r = \{I_r(1), I_r(2), \cdots, I_r(N_r)\}$ 和 $S_o = \{I_o(1),\ I_o(2), \cdots, I_o(N_o)\}$ 分别表示由独立运动相机拍摄的参考图像序列和观测图像序列, 其中 N_r 和 N_o 分别为两个序列的帧数. 可检验的整数时间偏移量 Δ 的取值范围是 $R = [-N_o + F, N_r - F]$.

在上述关于秩约束的推导中, 我们用到了三个假设: 1) 视频间的图像点对应已知; 2) 图像点跟踪持续整个图像序列; 3) 视频间的时域关系为常量偏移模型. 本节先讨论第一个假设, 其余两个假设在下一节中进行论述.

若视频间的点对应关系未知, 我们可以使用一个弱假设[9−10] 令秩约束仍然成立: 观测序列中跟踪点对应的三维点 $\hat{X}^{(t)}_i$ $(1 \leq i \leq P_o)$, 其空间位置可以用参考序列中所有跟踪点对应的三维点集的子集的线性组合描述, 即满足下列关系

$$\begin{bmatrix} \hat{X}^{(1)}_1 & \cdots & \hat{X}^{(1)}_{P_o} \\ \vdots & \ddots & \vdots \\ \hat{X}^{(F)}_1 & \cdots & \hat{X}^{(F)}_{P_o} \end{bmatrix} = \begin{bmatrix} X^{(1)}_1 & \cdots & X^{(1)}_{P_r} \\ \vdots & \ddots & \vdots \\ X^{(F)}_1 & \cdots & X^{(F)}_{P_r} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^{\mathrm{T}}_1 \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}_{P_o} \end{bmatrix}^{\mathrm{T}} \tag{6}$$

其中, $\boldsymbol{\alpha}_i \in \mathbf{R}^{P_r}$ 为线性关系系数向量, 满足 $\sum_j \boldsymbol{\alpha}_i(j) = 1$, 且在 F 帧图像序列中维持恒定, P_r 为参考序列中跟踪图像点的数量. 该假设的成立基于以下事实: 给定同一刚体上的三维点集, 其中至少有四个点处于不同平面, 则该集合中其他点的运动可以由这四个非平面点的线性组合描述. 当对该点集进行非奇异线性变换时, 如平移、旋转、尺度变换等, 该线性关系维持不变[22]. 人体属于铰接式物体, 其整体运动虽然是非刚性运动, 但各个部位的运动可以看作刚性运动. 结合式 (3) 和式 (6), 并在等号两边同乘上 $\boldsymbol{\Theta}^{\mathrm{T}}$, 可以得到

$$[\hat{\boldsymbol{\beta}}_1\ \cdots\ \hat{\boldsymbol{\beta}}_{P_o}] = [\boldsymbol{\beta}_1\ \cdots\ \boldsymbol{\beta}_{P_r}][\boldsymbol{\alpha}_1\ \cdots\ \boldsymbol{\alpha}_{P_o}] \tag{7}$$

同理, 当满足不等式 $3K \geq P_r + P_o$ 时, 新基系数矩阵 $\overline{M} = [\boldsymbol{\beta}_1\ \cdots\ \boldsymbol{\beta}_{P_r}\ \hat{\boldsymbol{\beta}}_1\ \cdots\ \hat{\boldsymbol{\beta}}_{P_o}]$ 的秩最大不超过 P_r. 引入这一弱假设的好处是, 在省去了估计视频间图像点对应的同时, 还使得该算法能够处理宽基线条件下的视频同步, 即被两个相机同时看到的三维点数量有限或者为零.

在实际应用中, 考虑到噪声的存在, $\overline{M}$ 几乎都是满秩的. 即便没有噪声, 由于轨迹基系数是一个超定系统的最小二乘解, $\boldsymbol{\beta}$ 和 $\hat{\boldsymbol{\beta}}$ 不可能完全相同. 为此, 我们采用矩阵的有效秩 $\hat{n}$[10]. 令 $s_1, \cdots, s_h$ 表示 $\overline{M}$ 的奇异值, 对给定的阈值 $\theta = \lambda\sum_{k=1}^{h} s_k$, 有效秩 $\hat{n} = \arg\min_j\{\sum_{k=1}^{j} s_k > \theta\}$, 其中 $0 < \lambda < 1$ 是预先确定的阈值. 我们定义下列距离函数 dst, 用来度量两个图像序列的时域对齐程度.

$$dst = \sum_{k=\hat{n}+1}^{h} s_k \tag{8}$$

图 2 是基系数矩阵 $\overline{M}$ 的奇异值在测试序列对同步和不同步两种情况下的一个示例. 这里我们令 $P_r = P_o = 20$. 与测试序列对不同步时的灰色曲线相比, 同步时的黑色曲线具有更快的下降速度, 代表其对应的基系数矩阵具有较小的奇异值.

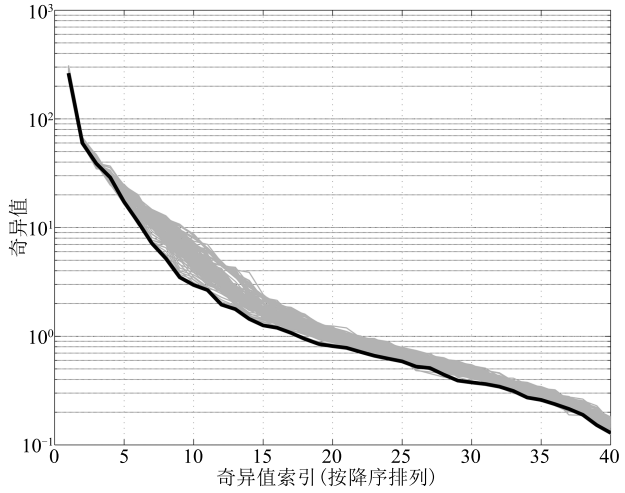


图 2 测试序列对同步和不同步时基系数矩阵 $\overline{M}$ 的奇异值

Fig. 2 An example of the singular values of $\overline{M}$ in synchronized case and non-synchronized cases

进一步地, 我们将距离函数 dst 转换为归一化的代价函数 c.

$$c(\overline{M}_{\Delta}) = 1 - \exp\left(-\frac{dst(\overline{M}_{\Delta})}{\sigma^2}\right) \tag{9}$$

其中, $\overline{\boldsymbol{M}}_{\Delta}$ 表示对应整数偏移量 Δ 的基系数矩阵. 从而, 求解 Δ 的最优化函数可以表示为

$$\Delta^{\star} = \arg\min_{\Delta} c(\overline{M}_{\Delta}) \tag{10}$$

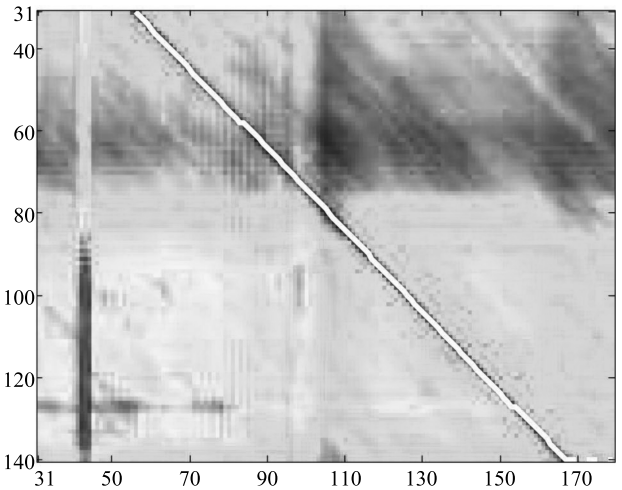
1.2 非线性时域对准

上述算法虽然不依赖视频间已知的像点对应, 但仍然假设图像点跟踪持续整个图像序列. 实际应用中, 由于遮挡、光线局部变化等原因, 多数跟踪算法很难维持长时间的精确跟踪. 此外, 假设视频间时域关系为一维常量偏移模型, 这也限制了同步算法的适用范围. 为此, 本文提出一种能够利用不同长度图像点轨迹的非线性时域对准算法.

我们将 S_r 和 S_o 分别划分为若干 F 帧长的子序列段, 令每个子序列段的中间帧为参考帧. 然后针对候选子序列对 $(f_r(j), f_o(k))$, 选择跟踪持续 $f_r(j)$ 的 P_r 个图像点和跟踪持续 $f_o(k)$ 的 P_o 个图像点, 重建这些点的三维轨迹并计算基系数矩阵 $\overline{M}$. 其中, $f_r(j)$ 表示 S_r 中参考帧为 $I_r(j)$ 的子序列, $f_o(k)$ 同理. 最后利用代价函数 c 估算 $(f_r(j), f_o(k))$ 的对齐程度, 记为 c_{jk}. 这样, 我们得到代价矩阵 $C \in \mathbf{R}^{(N_o-2*\lfloor F/2\rfloor)\times(N_r-2*\lfloor F/2\rfloor)}$, 其中 $\lfloor\cdot\rfloor$ 是地板函数.

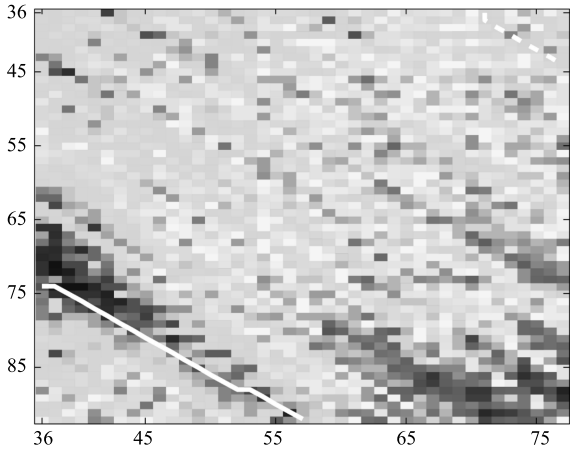
P_r 和 P_o 的确定依据以下两个规则: 1) 以参考图像序列为例, 将持续跟踪子序列段 $f_r(j)$ 的图像点个数记为 $P_r(j)$, 则 $P_r = \min\{P_r(j)\}$, $\lfloor F/2\rfloor + 1 \le j \le N_r - \lfloor F/2\rfloor$. 对 P_o 同理. 2) 满足两个不等式, $2F \ge 3K$ 和 $3K \ge P_r + P_o$. 前者为了确保运动目标轨迹重建时的超定系统, 后者则保证了 $\overline{M}$ 的秩最大不超过 P_r.

式 (6) 的成立基于常量时间偏移模型的假设, 即 $t_r = t_o + \Delta$. 当图像序列对的帧率不同但相近, 或者存在轻微的丢帧现象时, 该等式仍然近似成立. 由此, 相较于不同步的子序列对, 由时域同步或者最相近的子序列对构造的基系数矩阵仍然具有较小的秩. 非线性时域关系可以用一个离散映射函数 $\omega(t_o) = t_r$, $t_o = 1, \cdots, N$ 表示, 其中 $N \le N_o$, 表示观测序列中有 N 帧图像在参考序列中有时域对准的图像. 该函数在基于图的方法中为经过代价矩阵的一条路径. 为了应对局部时域重叠, 受文献 [16] 的最优路径搜索算法启发, 首先, 我们基于 Dijkstra 算法计算候选路径集, 每一条候选路径可以开始和结束于参考序列或观测序列的任何帧, 该路径的代价为其经过代价矩阵中各节点值加和的平均值. 然后, 基于候选路径集选择最优路径. 为了避免选择长度过短的路径, 根据路径结束于 S_r 或者 S_o, 我们将候选路径集划分为两个池, 分别在每个池中选择最小代价路径. 多数情况下, 一条路径被完全包含在另一条路径中 (图 3 (a)), 这时我们选择较短那条作为最优路径. 若两条候选路径不重叠 (图 3 (b)), 则选择较长那条



(a) 一条路径被完全包含在另一条路径中

(a) One path is contained entirely in the other



(b) 两条最小代价路径不重叠

(b) Two separate and non-overlapping paths

图 3　代价矩阵和最优路径 (白实线)

Fig. 3　Cost matrix and optimal path (white solid curve)

作为最优路径. 图中代价矩阵的横轴和纵轴分别表示 S_r 和 S_o 的帧索引, 其元素的颜色越深, 对应值越小.

双序列时域对准算法的具体流程如图 4 所示. 其中跟踪二维点轨迹和重建三维轨迹都是针对单个图像序列独立执行的, 唯一需要联合双序列的步骤是估算代价矩阵和最优路径. 注意, 我们分别选择 $f_r(j)$ 中的 P_r 个点和 $f_o(k)$ 中的 P_o 个点进行三维轨迹重建, 然后计算 $(f_r(j), f_o(k))$ 的对齐代价, 这一步骤需要重复 T 次, 最后取中值作为最终的 c_{jk}. 通常地, 当 P_r 和 P_o 值一定时, 重复次数越多, 算法鲁棒性越好, 相应地, 时间复杂度越高. 非线性时域对准算法的时间复杂度为 $\mathrm{O}(N_r \times N_o \times T)$. 试验中, 我们根据跟踪结果的精度和跟踪点数量决定 T. 一般地, 跟踪结果越准确, 跟踪点数越少, T 值越小. 若已知时域映射为线性或常量偏移模型, 可以利用线性时域对准算法使时间复杂度降为 $\mathrm{O}(N_r + N_o)$.

2　多序列时域对准

理论上, 我们可以将上述双序列时域对准算法简单地扩展到多序列的情况, 即增加代价矩阵的维度, 并搜索最优映射 $p : \mathbf{R} \rightarrow \mathbf{R}^D$, 其中 D 为待同步图像序列的数量. 然而在实际应用中, 这种方法是不可取的. 假设有五个待同步的序列, 每个序列以 30 fps 的帧率持续 10 秒, 即有 300 帧图像. 那么代价矩阵的元素数量达到 300^5. 若采用 32 位浮点数存储这个代价矩阵, 需要约 8.8 TB 的内存. 这显然已经超出了现有的硬件支持能力. 本文采用文献 [16] 中基于最小生成树 (Minimum spanning tree, MST) 的方法寻找双序列对准集合, 将多对多 (All-to-all) 的多序列时域对准问题简化为只利用最优的双序列对准获取全局的时域映射变换.

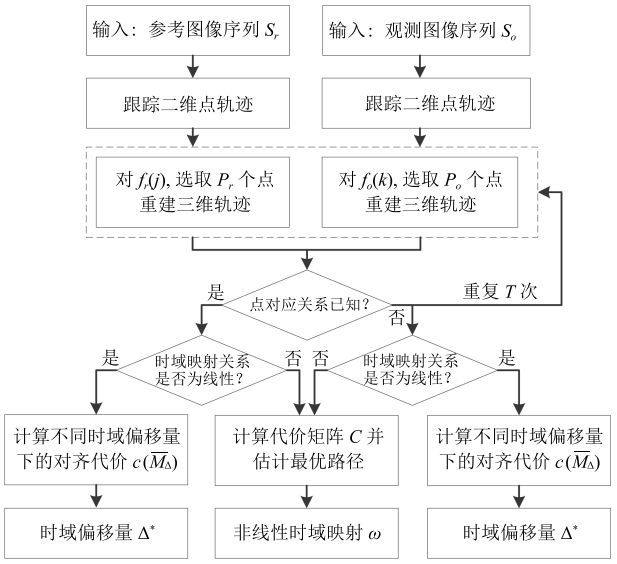


图 4　双序列时域对准算法流程图

Fig. 4　The flow chart of pairwise alignment

3　仿真实验

本文基于文献 [21] 提供的人体运动捕捉数据 (倒地、站立和步行) 生成仿真数据. 通过随机合成相机投影矩阵, 将 13 个人体关节点的三维运动轨迹投影到两组不断变化的图像平面上. 重建后的三维运动轨迹及其真实值如图 5 所示. 我们将其中一个图像序列作为参考序列, 将另外一个图像序列时域偏移 Δ 帧后, 再随机去掉若干帧 (丢帧率不高于 5 %), 作为观测序列. 每组实验重复 10 次, 每次采用不同的随机相机运动轨迹. 该实验中所有二维点轨迹均持续完整的图像序列, 计算代价矩阵时令采样次数 $T = 1$. 本文采用原始 DCT 基重建三维轨迹, 令每个维度上基的数量 $K = 30$.

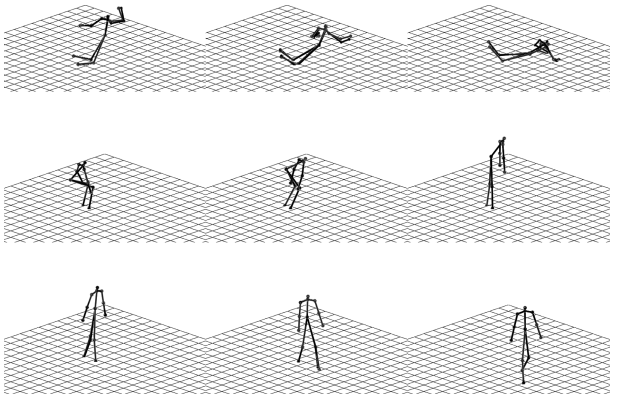


图 5　仿真数据重建结果 (黑) 和真实值 (灰)

Fig. 5　Reconstruction (black) and ground truth (gray) of simulated data

令真实时域映射函数为 $\{\hat{\omega}(t_o), t_o\}_{t_o=1,\cdots,N}$, 本文采用如下定义的归一化时域对准误差 ε 作为算法精度的衡量标准

$$\varepsilon = \frac{1}{N}\sum_{t_o=1}^{N}|\hat{\omega}(t_o)-\omega(t_o)| \tag{11}$$

3.1 鲁棒性

为了验证跟踪误差、数据丢失和图像点数量对新算法精度的影响, 我们设置了如下三组实验. 如不做特殊说明, 认为无数据丢失. 1) 不同跟踪误差和子序列段长度与时域对准误差间的关系, 结果如图 6 (a) 所示. 从图中可以看出, 子序列段越长, 算法受跟踪误差影响越小, 其同步精度越高. 但是, 延长子序列段会缩小代价矩阵, 相应地, 时域映射关系中自变量的取值范围变小, 表现在图中就是最优路径变短. 注意, 像素单位的跟踪偏移误差是通过用跟踪误差级 α 乘上一个服从标准正态分布的伪随机数得到. 2) 造成跟踪过程中目标数据丢失的原因有遮挡、自遮挡、度量失败等. 图 6 (b) 为不同程度的数据丢失 (0 %, 5 %, 10 %) 与时域对准误差间的关系. 只要跟踪到足够多帧数的观测值能确保三维轨迹重建时的超定系统, 新算法的精度基本不受数据丢失的影响. 3) 不同图像点数量与时域对准误差间的关系, 结果如图 6 (c) 所示. 理论上, 如果一个三维点的运动足够快并且随机, 它被两个独立运动的相机同时捕获到, 那么仅用这一个点就可以同步两个相机. 实际应用中考虑到单个点重复性运动的情况, 加入空间相对位置关系的约束, 综合多个位于不同刚性物体上的点能大大提高时域对准的精度.

3.2 准确性

本文在仿真数据基础上对比了本文方法与现有方法的时域同步精度, 包括文献 [17] 中基于反向投影的方法 BPM 和文献 [8] 中基于透视模型对极几何的方法 ECM. 其中, BPM 用到 3 组不同的对应点集, 每组点集包含 5 个图像点. 这两种方法均假设线性时域映射关系, 为了对比公平, 在它们计算代价矩阵的基础上, 利用本文提出的基于图的最优路径搜索算法, 寻找非线性时域映射函数.

除上述两种方法外, 我们还提出以下对比基准. 基于不同序列重建对应点的三维运动轨迹, 当观测序列和参考序列精确同步时, 同步帧索引的空间点重合; 当观测序列和参考序列为子帧级别同步时, 即帧和帧之间的时域偏移量为非整数, 则同步帧索引的空间点距离最小. 因此, 我们将新算法中基于秩约束的时域对齐度量准则替换为基于三维重建点距离的度量准则, 记为 PDM.

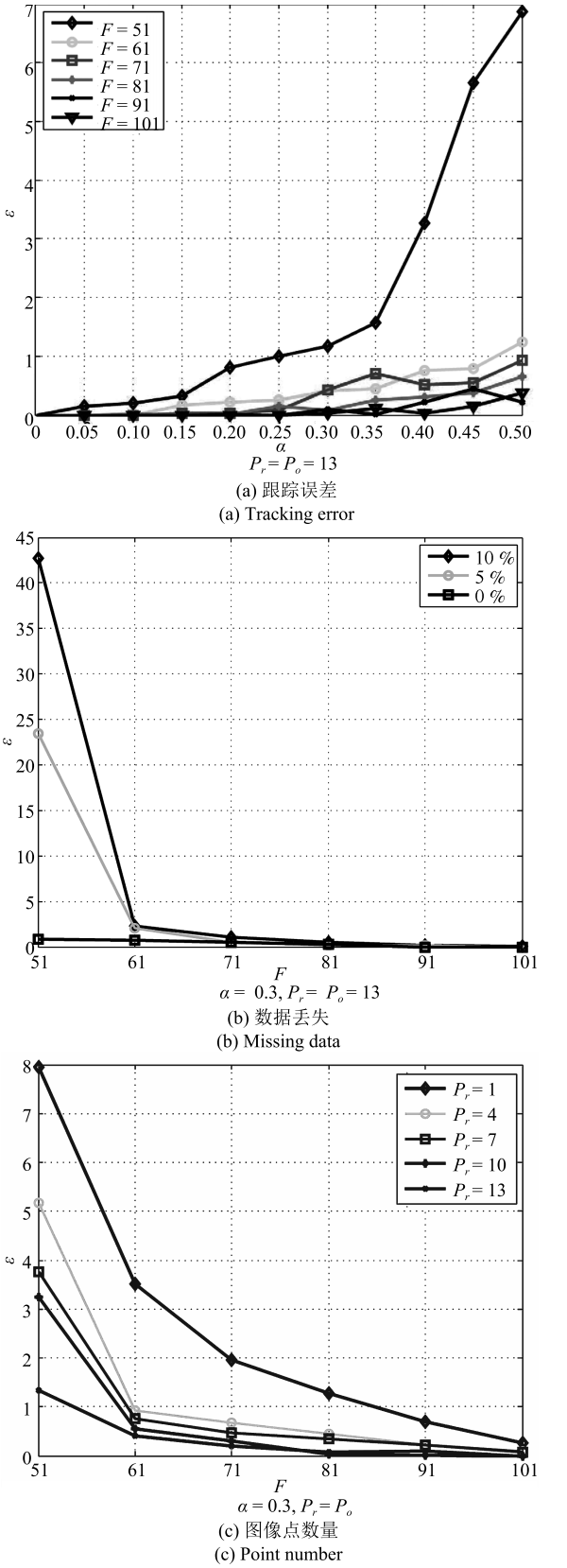


图 6 跟踪误差、数据丢失和图像点数量对同步结果的影响

Fig. 6 Comparisons of robustness with regard to tracking error, missing data and point number

图 7 (a) 为步行数据集上各算法受跟踪误差影响的时域对准精度曲线图, 是仿真数据集上各算法试验结果对比. 随着跟踪误差逐渐变大, 本文算法表现出更好的鲁棒性. ECM 方法的时域对准误差与跟踪误差不成正比, 这可能归结于试验中的非线性时域映射, 相比原始方法的线性时域映射假设, 非参数模型大大增加了解空间的维度, 导致当输入有噪声时该算法的准确度降低. 图 7 (b) ∼ 7 (i) 是针对倒地数据集中一个测试序列对 (常量偏移量 $\Delta = 35$) 各算法的同步结果. 图 7 (b) ∼ 7 (e) 为没有跟踪误差时, 各算法计算的代价矩阵和最优路径. 图 7 (f) ∼ 7 (i) 为当跟踪误差级 $\alpha = 0.30$ 时, 各算法计算的代价矩阵和最优路径. 注意, 与本文算法基于子序列对计算代价矩阵不同, 三种对比方法均计算任意帧对的时域对齐程度, 生成代价矩阵的维度为 $N_o \times N_r$.

4　第一人称视角数据

为了验证新算法在实际应用中的性能, 我们提出一个基于第一人称视角的社交场景视频数据集, 包括积木、健身毯、篮球和玩具火车四个场景. 其中, 前两个场景记录了 4 个 5 ∼ 6 岁儿童的交互式行为, 篮球场景记录了两组成年球员之间的 5 vs 5 对抗性比赛, 玩具火车场景是简单的刚体运动. 对象在

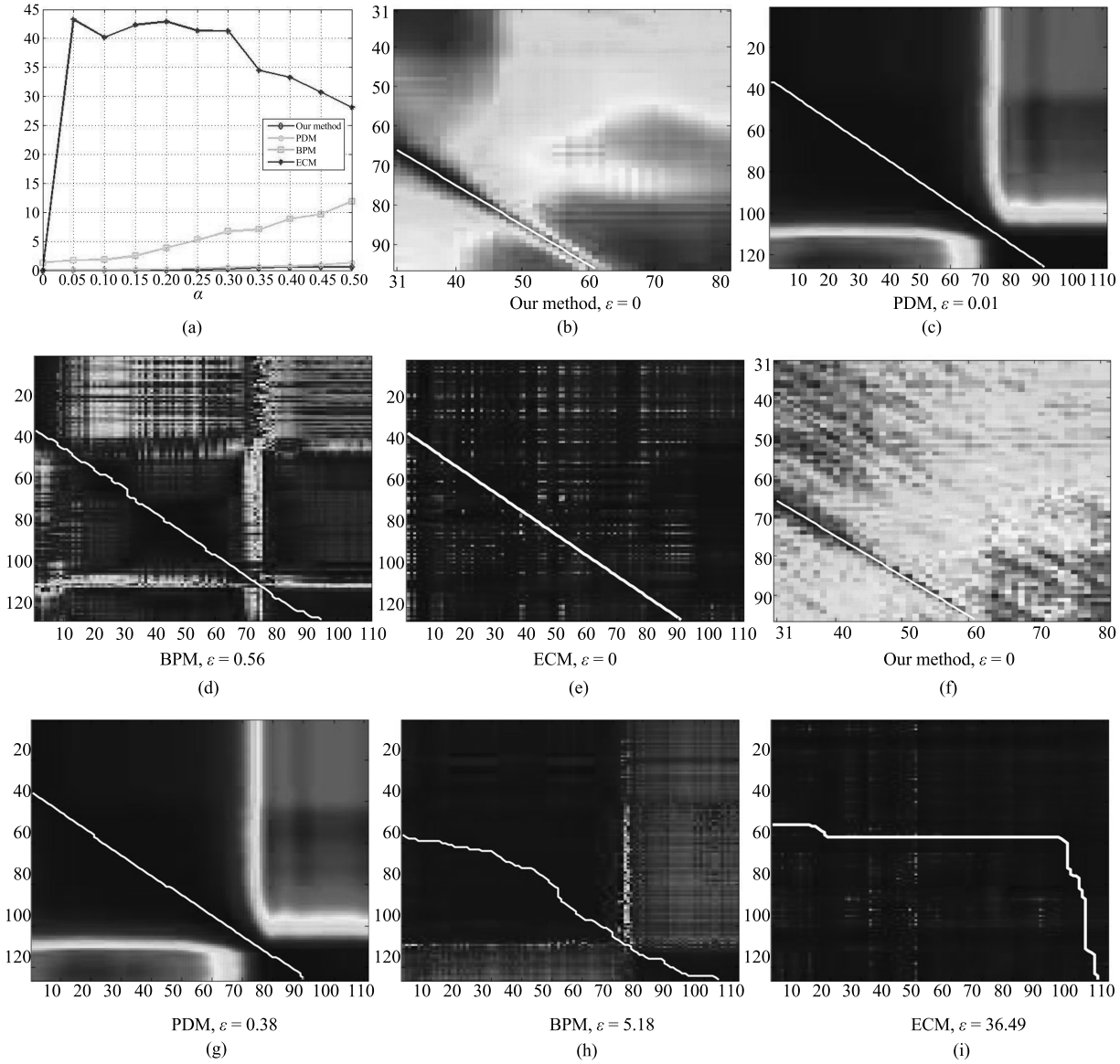


图 7　仿真数据集上各算法在不同跟踪误差下的实验结果对比以及估算的代价矩阵示例

Fig. 7　Comparisons of alignment accuracy using different methods regarding tracking noise level and representative cost matrices with estimated optimal paths superimposed

场内可以自由运动, 固定在其头部的 GoPro 相机以第一人称视角拍摄整个动态场景, 相机运动可近似认为相互独立. 区别于一般监控相机通常采取固定的位置和视角, 以第三人称视角从场景外部向场景内部进行拍摄, 第一人称视角相机是从场景内部佩戴者的视角出发 "观察" 场景, 通过不断变换位置或姿势获得对感兴趣目标的持续最佳观测视角. 每个场景的数据包括多个图像序列, 单个图像序列持续时间约 5 ∼ 10 秒, 伴随着相机平移运动约 3 ∼ 12 米, 绕光轴旋转运动约 20 ∼ 60 度.

数据采集时, 所有相机被设置到相同的拍摄模式, 例如图像尺寸、帧率等. 前期我们利用 FFmpeg 工具包从同一场景的不同视频源文件中提取图像帧作为测试序列. 该试验中, 参考序列和观测序列的生成帧率分别为 48 fps 和 46 fps. 我们利用外置闪光灯在数据采集开始时、采集中 (多次) 和采集结束时标记若干同步帧, 并手动同步余下帧, 以此作为视频序列同步的真实值.

对于包含关节人体的场景, 本文采用双粒度跟踪算法[23] 获取二维点轨迹, 其优势在于能够跟踪到大量位于人体躯干及四肢靠近上端部位的点, 从而确保其空间位置线性相关假设的成立. 另外, 该算法提供了一个控制空间采样率的参数, 可以避免像点分布过于集中. 缺点是基于稠密光流估计的点轨迹计算开销较大. 在玩具火车场景中, 我们利用 KLT 算法[24] 跟踪特征点轨迹. 基于跟踪算法的输出结果, 我们需要选择位于运动目标上的点轨迹, 并去掉长度过短以及明显错误的轨迹. 除上述自动跟踪点轨迹外, 我们还手动标记运动目标上若干特征点的二维运动轨迹. 由于遮挡导致某特征点不可见时, 我们根据前后相继帧以及辅助视角推测当前帧中该特征点的位置. 由于超出视角范围导致特征点不可见, 我们则不做标记.

本文利用运动恢复结构算法[25−27] 估计每一帧相机的空间姿态. 图 8 展示了对各场景的三维重建结果, 包括相机轨迹、静态场景和部分运动点轨迹. 在积木和健身毯场景中, 我们试图同步三个图像序列. 其中, 相较于 2 号图像序列, 3 号图像序列和参考图像序列的相机视角差别更大, 相机朝向几乎相反的方向. 本文用 #1 和 #2 分别表示各场景中的两组测试序列对.

在重建空间点的运动轨迹时, 如果相机运动缓慢, 其运动轨迹也可以用 DCT 基的线性组合表示, 这会导致轨迹重建的精度降低[21]. 由于不同时刻的图像集合可以模拟相机的快速随机运动, 为了提高轨迹重建的质量, 我们引入非测试用图像序列辅助轨迹重建, 并人工标注对应点.

除了仿真实验中提到的三种方法, 这里还额外对比了两种基于二维特征的方法: 基于二维运动特征的方法 MFM[16] 和基于 SIFT 特征匹配的方法 SMM[19]. 表 1 列出了各算法在真实数据集上的归一化时域对准误差 ε (式 (11)), 除玩具火车场景外, 本文算法在各测试序列对上的同步误差最小. 当自动跟踪点数量较少或者不满足空间位置线性相关假设时, 本文算法的同步精度下降. 这时, 可以通过添加手动标注图像点轨迹的方法提高同步质量. 由于 SMM 假设同步帧在图像内容上最相似, 从而不适用于宽基线条件下的相机同步, 在积木和健身毯场景中的同步误差较大. 图 9 ∼ 13 展示了不同场景中各算法的帧同步结果, 各算法的输入图像点叠加显示在对应图像帧上, 空白表示观测序列中不存在同步帧. 由于空间有限, 这里仅给出了本文算法在自动跟踪点轨迹输入下的同步结果. 试验中令 $K = 30$, $F = 81$, $\lambda = 0.99$. 关于有效秩定义中阈值 λ 的取值, 图 14 给出了本文算法在积木 #1 上的一组对比结果. 图 14 (a) 为不同有效秩对同步结果的影响, 图 14 (b) ∼ 14 (e) 是不同有效秩对应的代价矩阵. 当 $\lambda < 0.99$ 时, 同步结果的精度出现明显下降. 而当 λ 越接近于 1 时, 同步结果的精度越好.

由于本文算法只适用于帧率相同或者相近的图像序列对, 帧率相差越大, 秩约束越弱. 图 15 对比了不同帧率比时本文算法的同步误差. 图 15 (a) 为不同帧率比对同步结果的影响. 图 15 (b) ∼ 15 (d) 是当观测序列帧率分别为 46 fps、40 fps 和 24 fps 时

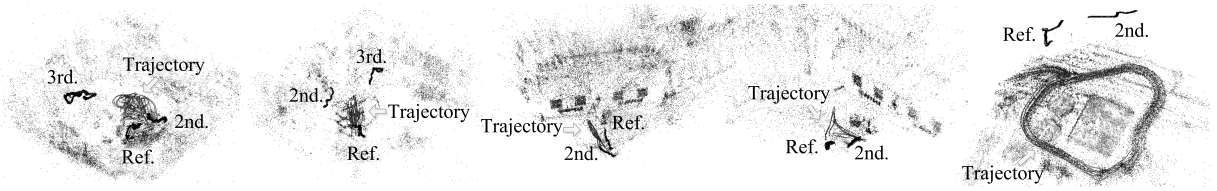


图 8 三维重建结果 (从左到右对应场景依次为: 积木, 健身毯, 篮球 #1, 篮球 #2 和玩具火车)

Fig. 8 The 3D reconstruction results (From left to right: block building, exercise mat, basketball (#1), basketball (#2) and toy train.)

表 1　真实数据集上各算法的归一化时域对准误差对比 (帧)

Table 1　Quantitative comparisons of alignment error on real scenes (frame)

	积木 #1	积木 #2	健身毯 #1	健身毯 #2	篮球 #1	篮球 #2	玩具火车
BPM (手动标记点轨迹)	39.61	9.16	12.05	15.63	16.81	12.42	56.80
ECM (手动标记点轨迹)	25.15	32.37	57.48	62.60	50.44	29.83	24.86
MFM (自动跟踪点轨迹)	11.81	21.70	22.17	9.44	17.68	22.78	70.04
SMM (SIFT)	155.75	196.56	132.08	202.50	9.71	31.74	130.83
PDM (手动标记点轨迹)	0.85	2.53	2.96	4.60	4.29	1.49	**1.28**
本文算法 (手动标记点轨迹)	**0.45**	**1.27**	2.52	2.76	3.07	1.12	1.33
本文算法 (自动跟踪点轨迹)	0.52	1.74	**1.35**	**1.48**	**2.84**	1.54	3.18
本文算法 (手动标记和自动跟踪)	0.56	1.40	2.07	1.99	3.75	**0.92**	2.01

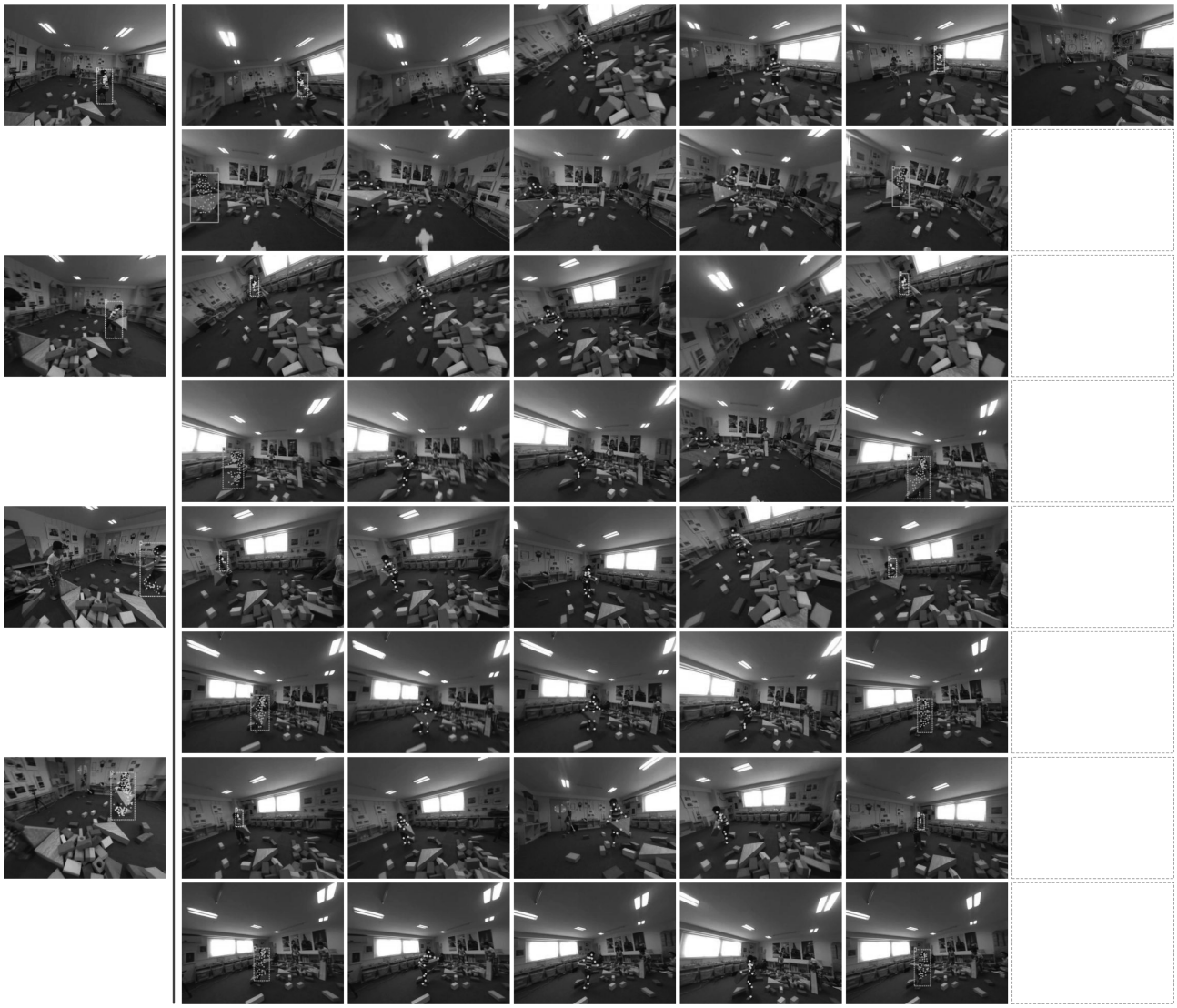


图 9　积木场景中各算法的时域对准结果对比 (从左到右依次为: 参考序列中的图像帧、本文算法、PDM、BPM、ECM、MFM 和 SMM 找到的第二个序列中的对应帧 (上) 及第三个序列中的对应帧 (下))

Fig. 9　Synchronization results on the blocks scene (From left to right: sample frames from the reference sequence, corresponding frames from the second sequence (top) and the third sequence (bottom) by our method, PDM, BPM, ECM, MFM and SMM, respectively.)

图 10 健身毯场景中各算法的时域对准结果对比 (同图 9)

Fig. 10 Synchronization results on the exercise mat scene idem as Fig. 9

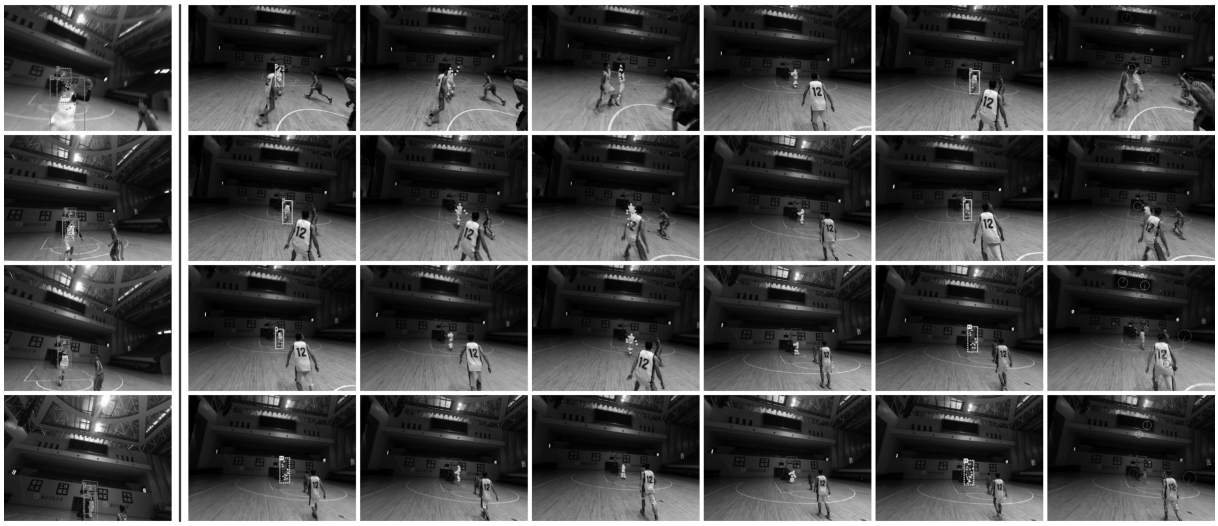


图 11 篮球 #1 场景中各算法的时域对准结果对比 (从左到右依次为: 参考序列中的图像帧、本文算法、PDM、BPM、ECM、MFM 和 SMM 找到的第二个序列中的对应帧)

Fig. 11 Synchronization results on the basketball scene (#1) (From left to right: sample frames from the reference sequence, corresponding frames from the second sequence by our method, PDM, BPM, ECM, MFM and SMM, respectively.)

图 12　篮球 #2 场景中各算法的时域对准结果对比 (同图 11)

Fig. 12　Synchronization results on the basketball scene (#2) idem as Fig. 11

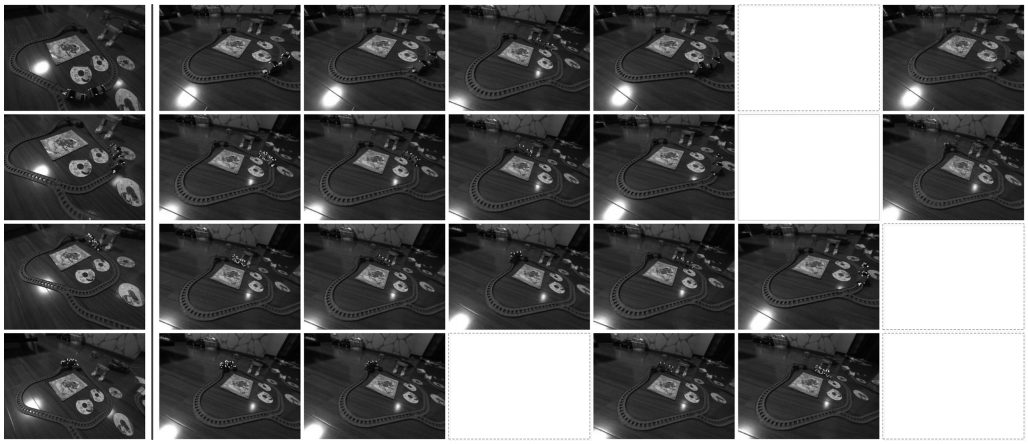


图 13　玩具火车场景中各算法的时域对准结果对比 (同图 11)

Fig. 13　Synchronization results on the toy train scene idem as Fig. 11

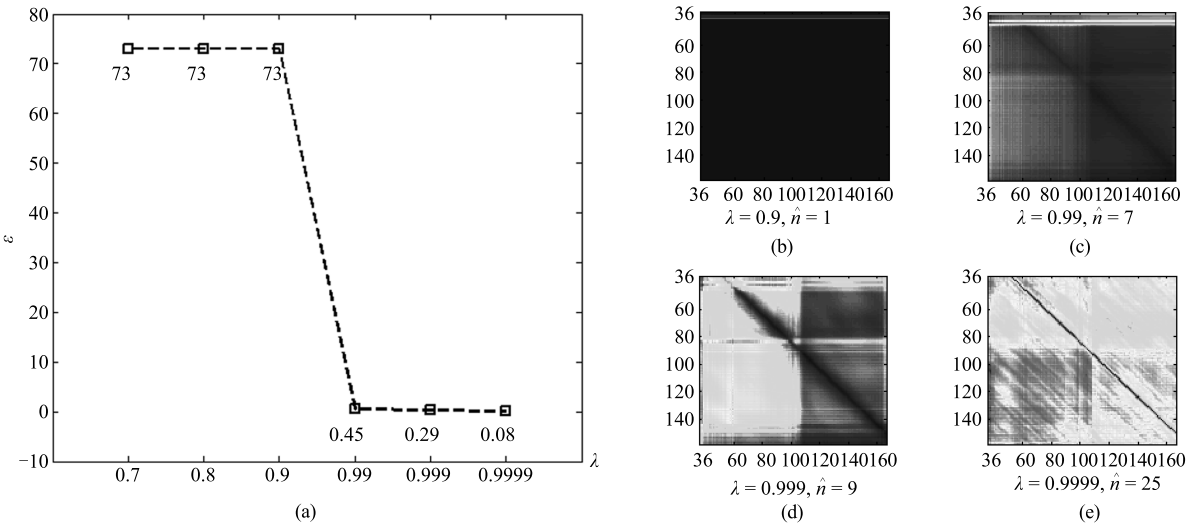


图 14　不同有效秩对同步结果的影响及不同有效秩对应的代价矩阵

Fig. 14　Comparisons of alignment accuracy with different λ values for efficient rank and cost matrices computed with different λ values

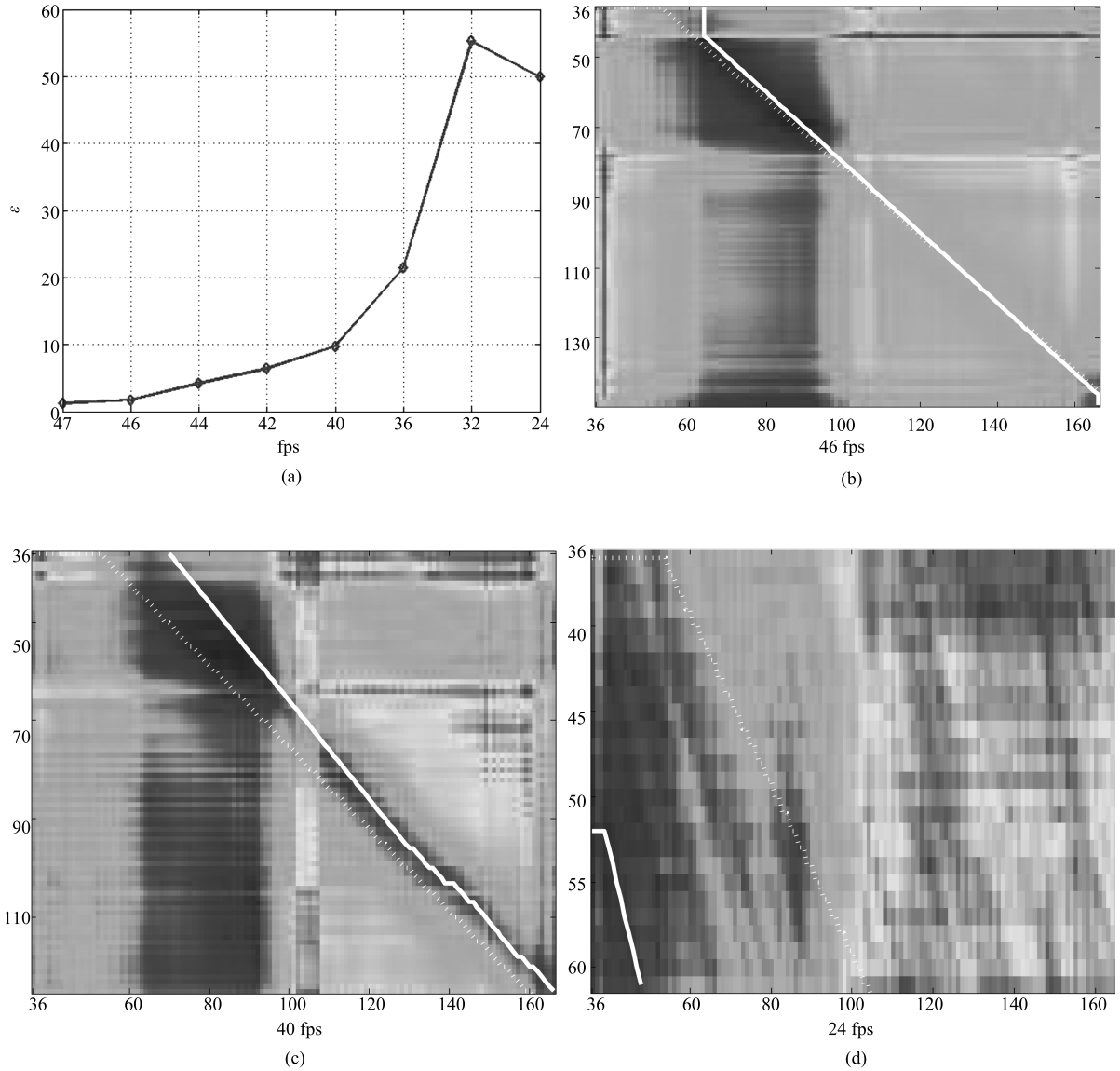


图 15 不同帧率比对同步结果的影响及观测序列帧率为 46 fps、40 fps 和 24 fps 时的代价矩阵

Fig. 15 Comparisons of alignment accuracy with different frame rate ratios and cost matrices computed when the frame rate of the observed sequence is 46, 40 and 24, respectively

的代价矩阵, 图中最优路径的估算值和真实值分别用实线和虚线标识. 试验中我们令参考图像序列的生成帧率为 48 fps, 仅改变观测图像序列的帧率. 当帧率比接近于 2 时, 序列同步时基系数矩阵具有较小秩的特征几乎不明显.

实验选用了如下测试环境: CPU 为 Inter i5-4570 4-Core 3.20 GHz, 8 GB 内存, MATLAB R2010a 编程环境. 对分辨率为 640 像素 × 480 像素的一帧图像来说, 预处理阶段平均花费时间为 204 s, 其中 195 s 用于图像点轨迹跟踪, 6 s 用于相机空间姿态估计, 完整同步算法的平均运行时间为 453 ms, 其中 429 ms 用于三维点轨迹重建. 如果序列间点对应关系已知, 将算法中基于秩约束的度量准则替换为基于三维重建点距离的度量准则, 可以大大提高算法效率, 算法平均运行时间缩短到每帧 2.8 ms. 原因在于, 对每个空间点, 后者仅需要执行一次三维轨迹重建即可, 而在基于秩约束的方法中, 计算每组子序列对的对齐代价时都要执行一次三维轨迹重建, 从而保证参考序列和观测序列具有相同的轨迹基.

5 结论

本文提出一种针对独立运动相机和动态场景的视频时域同步算法. 对于给定的轨迹基, 利用不同图像序列重建的空间点运动轨迹的系数能够用于同步

这些图像序列. 我们提出一种基于轨迹基系数的秩约束, 结合基于图的最优路径搜索算法, 实现视频间的非线性时域对准. 本文提出方法不要求图像点轨迹持续整个序列, 也不依赖已知的视频间点对应关系, 从而能够处理动态场景下由独立运动相机拍摄的视频间的时域同步.

本文方法仅限于若干相机同时拍摄同一场景的情形, 类似问题例如人体动作识别或视频检索, 是若干相机在不同时刻拍摄相似的场景. 本文作者在接下来的工作中会继续研究这类问题的视频同步方法.

致谢

感谢西北工业大学体育部和附属幼儿园在本文数据采集工作中给予的协助.

References

1 Caspi Y, Irani M. Spatio-temporal alignment of sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**(11): 1409−1424

2 Caspi Y, Simakov D, Irani M. Feature-based sequence-to-sequence matching. *International Journal of Computer Vision*, 2006, **68**(1): 53−64

3 Lu C, Mandal M. A robust technique for motion-based video sequences temporal alignment. *IEEE Transactions on Multimedia*, 2013, **15**(1): 70−82

4 Pundik D, Moses Y. Video synchronization using temporal signals from epipolar lines. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer Berlin Heidelberg, 2010. 15−28

5 Pádua F, Carceroni F, Santos G, Kutulakos K. Linear sequence-to-sequence alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(2): 304−320

6 Yilmaz A, Shah M. Matching actions in presence of camera motion. *Computer Vision and Image Understanding*, 2006, **104**(2−3): 221−231

7 Rao C, Gritai A, Shah M, Syeda-Mahmood T. View-invariant alignment and matching of video sequences. In: Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE, 2003. 939−945

8 Tresadern P A, Reid I D. Video synchronization from human motion using rank constraints. *Computer Vision and Image Understanding*, 2009, **113**(8): 891−906

9 Wolf L, Zomet A. Correspondence-free synchronization and reconstruction in a non-rigid scene. In: Proceedings of the 7th European Conference on Computer Vision, Workshop on Vision and Modelling of Dynamic Scenes. Copenhagen, Denmark: Springer Berlin Heidelberg, 2002.

10 Wolf L, Zomet A. Wide baseline matching between unsynchronized video sequences. *International Journal of Computer Vision*, 2006, **68**(1): 43−52

11 Sand P, Teller S. Video matching. *ACM Transactions on Graphics*, 2004, **23**(3): 592−599

12 Evangelidis G D, Bauckhage C. Efficient subframe video alignment using short descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(10): 2371−2386

13 Serrat J, Diego F, Lumbreras F, Álvarez J M. Synchronization of video sequences from free-moving camreas. In: Proceedings of the 3rd Iberian Conference on Pattern Recognition and Image Analysis, Part II. Girona, Spain: Springer Berlin Heidelberg, 2007. 620−627

14 Diego F, Ponsa D, Serrat J, López A M. Video alignment for change detection. *IEEE Transactions on Image Processing*, 2011, **20**(7): 1858−1869

15 Diego F, Serrat J, López A M. Joint spatio-temporal alignment of sequences. *IEEE Transactions on Multimedia*, 2013, **15**(6): 1377−1387

16 Wang O, Schroers C, Zimmer H, Gross M, Sorkine-Hornung A. VideoSnapping: interactive synchronization of multiple videos. *ACM Transactions on Graphics*, 2014, **33**(4): 77: 1−77: 10

17 Tuytelaars T, van Gool L. Synchronizing video sequences. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington D C, USA: IEEE, 2004. 762−768

18 Lei C, Yang Y. Trifocal tensor-based multiple video synchronization with subframe optimization. *IEEE Transactions on Image Processing*, 2006, **15**(9): 2473−2480

19 Dexter E, Pérez P, Laptev I. Multi-view synchronization of human actions and dynamic scenes. In: Proceedings of the 2009 British Machine Vision Conference. London, UK: BMVA Press, 2009. 122: 1−122: 11

20 Akhter I, Sheikh Y, Khan S, Kanade T. Nonrigid strcture from motion in trajectory space. In: Proceedings of the 2008 Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS, 2008. 41−48

21 Park H S, Shiratori T, Matthews I, Sheikh Y. 3D reconstruction of a moving point from a series of 2D projections. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010. 158−171

22 Kutulakos K N, Vallino J. Affine object representations for calibration-free augmented reality. In: Proceedings of the 1996 IEEE Virtual Reality Annual International Symposium. Washington DC, USA: IEEE, 1996. 25−36

23 Fragkiadaki K, Zhang W J, Zhang G, Shi J B. Two-granularity tracking: mediating trajectory and detection graphs for tracking under occlusions. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 552−565

24 Lucas B D, Kanade T. An interative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence. Vancouver, Canada: Morgan Kaufmann Publishers Inc., 1981. 674−679

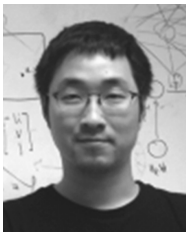
25 Snavely N, Seitz S M, Szeliski R. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics*, 2006, **25**(3): 835−846

26 Hartley R I, Zisserman A. *Multiple View Geometry in Computer Vision* (2nd edition). Cambridge: Cambridge University Press, 2004.

27 Park H S, Jain E, Sheikh Y. 3D gaze concurrences from head-mounted cameras. In: Proceedings of the 2012 Advances in Neural Information Processing Systems. Nevada, USA: NIPS, 2012. 422−430

王 雪 西北工业大学计算机学院博士研究生. 主要研究方向为目标跟踪, 人体行为分析.
E-mail: xwang@mail.nwpu.edu.cn
(**WANG Xue** Ph. D. candidate at the School of Computer Science and Engineering, Northwestern Polytechnical University. Her research interest covers object tracking and human behavior analysis.)

SHI Jian-Bo 宾夕法尼亚大学工程与应用科学学院教授. 主要研究方向为人体行为分析, 图像识别分割.
E-mail: jshi@seas.upenn.edu
(**SHI Jian-Bo** Professor at the School of Engineering and Applied Science, University of Pennsylvania, USA. His research interest covers human behavior analysis and image recognition-segmentation.)



PARK Hyun-Soo 宾夕法尼亚大学工程与应用科学学院博士后. 主要研究方向为基于视觉社交信号的人体交互行为分析, 如注意力运动、面部表情和身体姿势等. E-mail: hypar@seas.upenn.edu
(**PARK Hyun-Soo** Postdoctoral fellow at the School of Engineering and Applied Science, University of Pennsylvania, USA. His research interest covers human interact with one another by sending visible social signals, such as gaze movements, facial expressions, and body gestures.)

王 庆 西北工业大学计算机学院教授. 主要研究方向为计算机视觉, 图像与视频处理, 光场成像, 虚拟现实. 本文通信作者. E-mail: qwang@nwpu.edu.cn
(**WANG Qing** Professor at the School of Computer Science and Engineering, Northwestern Polytechnical University. His research interest covers computer vision, image and video signal processing, light field, and virtual reality. Corresponding author of this paper.)