基于多个手持摄像机的动态场景时空一致性深度恢复

姜翰青[1)]，章国锋[1)*]，谭 平[2)]，鲍虎军[1)]

[1)](浙江大学 CAD & CG 国家重点实验室 杭州 310058)

[2)](Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576)

(zhangguofeng@cad.zju.edu.cn)

摘 要：针对现有的动态场景深度恢复方法普遍需要较多数目的同步摄像机才能获得理想深度估计的问题，提出一个能够从 2～3 个手持摄像机所拍摄的同步视频序列中自动地恢复出高质量的深度图序列的鲁棒、便捷的动态场景稠密深度恢复方法. 首先对不同序列同一时刻的图像帧进行匹配以完成每帧的深度初始化，然后采用一种新的双层分割方法在手持摄像机自由移动的情况下将像素进行静动态分类，并对静态和动态像素点采用不同的方式进行时空一致性的深度优化. 特别地，文中采用了一个基于多帧统计信息、迭代式的优化框架，使得深度优化与双层分割在该优化框架之下交替迭代地进行，最终实现高质量的动态场景的分割和深度恢复. 最后通过各种动态场景实例证明了文中方法的鲁棒性和有效性.

关键词：动态场景；时空一致性深度恢复；双层分割

中图法分类号：TP391

Spatio-Temporal Depth Recovery of Dynamic Scenes with Multiple Handheld Cameras

Jiang Hanqing[1)], Zhang Guofeng[1)*], Tan Ping[2)], and Bao Hujun[1)]

[1)](*State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou* 310058)

[2)](*Department of Electrical and Computer Engineering, National University of Singapore, Singapore* 117576)

Abstract: Most previous dynamic depth recovery methods require many synchronous cameras to achieve good depth estimation, which is inflexible in practice. In this paper, a novel dynamic dense depth recovery method is proposed to automatically recover high-quality dense depth information from the synchronized videos taken by two or three handheld cameras. Initial depth maps are computed by matching the synchronized frames in the same time instance. Then a novel bilayer segmentation method for freely moving cameras is employed to classify the pixels of each frame into static and dynamic ones, so that their depths can be more effectively optimized with different spatio-temporal coherence constraints. Especially, an iterative optimization framework based on multiple frames is proposed, which iteratively performs depth optimization and bilayer segmentation to finally achieve a set of temporally consistent segmentation and depth maps. A variety of dynamic scene examples demonstrate the effectiveness and robustness of the proposed method.

Key words: dynamic scene; spatio-temporal depth recovery; bilayer segmentation

---

修改稿收到日期：2012-09-04. 基金项目：国家科技支撑计划(2012BAH35B02)；国家自然科学基金(61103104)；中央高校基本科研业务费专项资金. 姜翰青(1984－)，男，博士研究生，主要研究方向为基于视频的三维重建与分割；章国锋(1981－)，男，博士，副教授，论文通讯作者，主要研究方向为摄像机跟踪、三维重建、视频分割与编辑、增强现实等；谭 平(1980－)，男，博士，助理教授，主要研究方向为计算机视觉、计算机图形学；鲍虎军(1966－)，男，博士，教授，博士生导师，CCF 理事，主要研究方向为计算机图形学、三维视觉、虚拟现实、增强现实等.

三维重建作为计算机视觉领域的经典问题，经过几十年的发展，已经涌现出了很多优秀的 3D 重建算法[1-3]，但其中大部分方法主要针对静态场景. 由于现实复杂场景还有很多动态元素，因此从拍摄的场景视频中准确地恢复动态物体的稠密深度信息是一个非常具有挑战性的问题. 直接将静态场景的 3D 重建方法应用到动态场景中难以得到理想的结果，特别是时空一致性难以保证，因此，最近几年一些针对动态场景的 3D 重建方法相继被提出来.

大部分现有的动态场景重建方法[4-7]需要基线较窄的固定摄像机阵列，并且其中许多方法[4,6-7]都要求较多数目的摄像机来保证重建质量. 对于数据捕获设备和环境的较高要求使得这些方法仅适用于实验室环境下严格拍摄的数据. 为此，本文提出了一种利用多目手持摄像机来实现动态场景的稠密深度恢复方法，不但允许每个摄像机可以独立自由地移动，而且仅需 2～3 个摄像机就能获得高质量的深度恢复. 与传统基于多个固定摄像机的深度恢复方法相比，本文方法不但改善了数据捕获的便携性和适用范围，还有效地提高了深度恢复的质量.

对于含有动态物体的场景，由于静态和动态像素在时域上具有不同的特性，因此有必要先对像素进行静动态分类，然后对静态和动态像素进行各自不同的时空一致性深度优化. 本文方法的主要思想是先根据颜色和几何一致性将静动态像素分离出来，然后对于静态和动态像素利用不同时空一致性约束进行深度优化. 特别需要指出的是，本文提出的双层分割方法即使在手持摄像机自由运动的情况下，也可有效地将视频中的动态物体分割出来. 而大多数现有的方法并没有进行静动态分割，而是对于所有像素采用相同的策略进行优化. 虽然一些方法(如文献[8-9])也进行了双层分割，但往往需要已知背景信息，而且要求摄像机的位置固定，局限性较大. 相比而言，本文的双层分割方法利用了多帧上的对应关系，在摄像机自由运动的情况下亦能可靠地区分出前景和背景. 由于优化后的深度图可以用来进一步改善静动态分割结果，本文还提出了一种基于多帧统计的迭代优化框架，通过对分割结果和深度图进行交替迭代的优化，最终获得精确的深度图和静动态分割结果.

1 相关工作

现有的许多方法都是利用多个固定的摄像机来重建动态场景. 下面简要地介绍其中具有代表性的方法：文献[10]方法提出同时恢复场景流和 3D 结构，从而为每个图像区域拟合一个 3D 仿射模型；Zitnick 等[4]针对视角插值问题提出一种基于视频分割的视角相关 3D 重建方法；Gong[5]引入深度流的概念，在实时立体重建过程中加强时域上的深度一致性；Lei 等[7]利用一种基于区域树的立体匹配方法对固定的摄像机阵列拍摄的多目视频恢复时空一致性深度图. 还有一些方法通过在时域上平滑对应像素点的深度从而对动态像素点的 3D 位置进行优化；Tao 等[11]用 3D 面片模拟每个图像分割区域，并通过匹配相邻时序帧上的投影来重建 3D 面片；Larsen 等[6]利用一种改进的置信度传递算法加强多视频流中连续帧之间的时序深度一致性. 然而，简单的时域平滑容易造成过平滑等瑕疵，而且对深度噪声和对应点跟踪错误很敏感. Yang 等[12]在集束优化[2]的框架上做了扩展，使之能够处理动态场景，该方法比简单的时域平滑法效果要好. 最近，Yang 等[13]提出一个针对三目摄像机的时空一致性深度恢复方法，但要求摄像机之间的相对姿态不变，而且基线比较窄. 总的来说，这些方法普遍要求摄像机之间的基线较窄或摄像机数目较多，以鲁棒地处理遮挡.

还有一些方法在 3D 重建过程中利用了与本文方法类似的双层分割. Goldlücke 等[8]的方法以及 Guillemaut 等[9]的方法均在假设已知背景颜色和深度的条件下同时求解深度估计和层次分割，这 2 种方法都需要固定的摄像机以方便预先采集或者估计背景的颜色和深度，然后通过背景差异法来实现双层分割. 相比而言，本文实现了在摄像机移动条件下的双层分割. Zhang 等[14]提出了一种鲁棒的基于稠密深度和运动估计的双层分割方法，能够较好地处理摄像机移动条件下的双层分割，但是该方法需要预先通过人工标记的方式学习前景颜色概率模型. Liu 等[15]对于前景的颜色及位置信息进行学习，能够实现运动幅度较小的物体的双层分割；然而此方法须假设视频序列的背景区域保持仿射变换，以便实现关键帧的初始分割. Zhong 等[16]提出了 FLKDE 模型来同时模拟局部颜色分布和时域一致性约束，并以此实现渐进式的双层分割；此方法仅允许摄像机缓慢连续地移动，而且要求第一帧提供正确的分割结果. 最近，Jiang 等[17]提出一种多摄像机自由移动条件下的动态场景深度恢复方法，将深度估计和双层分割利用统一的能量优化函数进行联合求解，其中双层分割只是为了更好地辅助深度优化，并不

能精确地区分静态和动态像素点，这种联合优化方法的确能够处理较复杂、通用的情况（例如前景运动幅度比较小的情况）；然而由于其分割项没有显示地结合颜色分割和时域光流等信息，因此静动态分割结果会在不连续边界上产生少量的噪声，从而影响深度恢复的质量. 相比而言，本文提出的显示双层分割方法结合了 Mean-shift 分割和时序对应信息来加强边界分割的准确性和时空一致性，因此生成的深度图结果能够保持更加清晰的边界.

2　算法目标和系统概述

给定一组由 M 个摄像机拍摄的同步视频序列 $\mathcal{I}=\{I_m\,|\,m=1,\cdots,M\}$，其中每个序列 I_m 包含 n 帧，记为 $I_m=\{I_m^t\,|\,t=1,\cdots,n\}$，其对应的深度图序列为 $\mathcal{Z}=\{Z_m^t\,|\,t=1,\cdots,n\}$. $I_m^t(\boldsymbol{x})$ 和 $Z_m^t(\boldsymbol{x})$ 用来表示在序列 m 第 t 帧中像素点 $\boldsymbol{x}$ 的颜色和深度. $D_m^t(\boldsymbol{x})=1/Z_m^t(\boldsymbol{x})$ 定义为 $\boldsymbol{x}$ 的视差，待恢复的视差序列图 $\mathcal{D}=\{D_m^t\,|\,m=1,\cdots,M;t=1,\cdots,n\}$. 为了简化表示，本文用 $\boldsymbol{x}_m^t$ 来表示 $\boldsymbol{x}$.

假设所有序列每一帧的摄像机参数都已知. 本文实验中先利用文献[18]的方法来跟踪视频中的特征点，然后使用文献[19]的方法恢复相机参数. 为保证算法的鲁棒性，摄像机内部参数都是预先标定并在拍摄数据过程中固定不变. 本文方法的整体框架如图 1 所示，首先利用 t 时刻 M 个序列的同步帧 $\mathcal{I}(t)=\{I_m^t\,|\,m=1,\cdots,M\}$ 来初始化视差图 $\{D_m^t\,|\,m=1,\cdots,M\}$；有了初始深度图之后，先进行双层分割，然后运用时空一致性优化方法来对深度图进一步求精，对于静态和动态像素点利用不同的方法进行优化. 采用本文方法对深度和分割结果进行迭代地优化，最终实现高质量的 3D 重建.

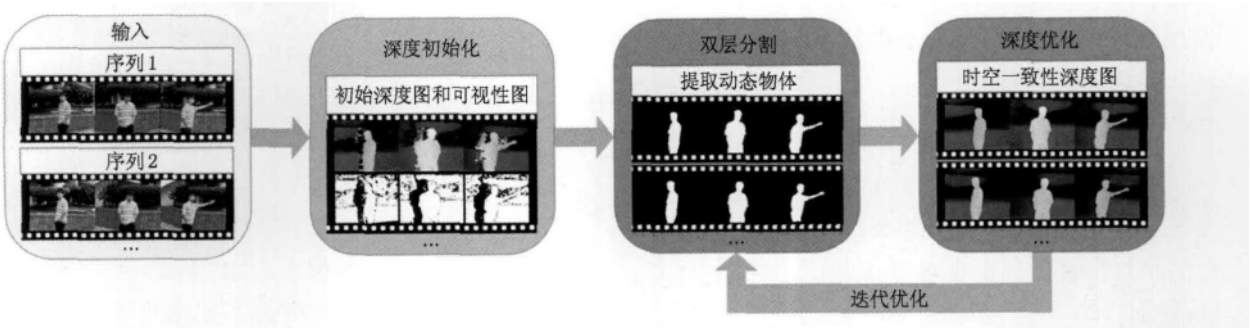


图 1　本文方法整体框架

3　深度初始化

假设视差值的范围为 $[d_{\min},d_{\max}]$，将其等分为 k 个候选视差层，第 i 层 $d_i=(k-i)/(k-1)\cdot d_{\min}+(i-1)/(k-1)\cdot d_{\max}$，其中 $i=1,\cdots,k$.

本文的深度初始化环节与文献[17]类似，即利用每个 t 时刻不同序列的同步视频帧来估计初始深度图，其能量优化函数定义为

$$E_{\mathrm{D}}(D_m^t;\mathcal{I}(t))=E_{\mathrm{D}}^1(D_m^t;\mathcal{I}(t))+E_{\mathrm{D}}^2(D_m^t)\qquad(1)$$

其中，E_{D}^1 为数据项，E_{D}^2 为平滑项. E_{D}^2 定义为 $E_{\mathrm{D}}^2(D_m^t)=\lambda\sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}\in N(\boldsymbol{x})}\min\{|D_m^t(\boldsymbol{x})-D_m^t(\boldsymbol{y})|,\eta\}$. 其中，$N(\boldsymbol{x})$ 表示像素点 $\boldsymbol{x}$ 的相邻点集；λ 为平滑权重，实验中取为 $0.8/(d_{\max}-d_{\min})$；η 为不连续视差截断值，通常取为 $0.03(d_{\max}-d_{\min})$.

与文献[17]类似，本文在数据项 E_{D}^1 中使用 DAISY 描述符[20]实现宽基线摄像机之间的鲁棒立体匹配. 基于 DAISY 描述符相似度的深度度量函数定义为

$$L_{\mathrm{d}}(\boldsymbol{x}_m^t,d_i;I_m^t,I_{m'}^t)=\|\mathcal{D}(\boldsymbol{x}_m^t)-\mathcal{D}(\boldsymbol{x}_{m'}^t)\|_2,$$

其中，$\mathcal{D}(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 的 DAISY 描述符，$\boldsymbol{x}_{m'}^t$ 是 $\boldsymbol{x}_m^t$ 利用其视差 d_i 和摄相机参数投影得到的对应点. 由此定义数据项

$$E_{\mathrm{D}}^1(D_m^t;\mathcal{I}(t))=\sum_{\boldsymbol{x}_m^t}\frac{\sum_{m'\neq m}L_{\mathrm{d}}(\boldsymbol{x}_m^t,D_m^t(\boldsymbol{x}_m^t);I_m^t,I_{m'}^t)}{M-1}\qquad(2)$$

式(1)的求解采用松弛置信度传递算法[21]. 图 2 c 所示为深度图初始化结果，可以看出，在遮挡区域仍然存在一些明显问题. 为解决此问题，本文通过估计同步帧之间的可见性图来处理遮挡.

不可见点的深度推断

定义序列 m 关于 m' 的可视性图为

$$V_{m\to m'}^t(\boldsymbol{x}_m^t)=|D_{m\to m'}^t(\boldsymbol{x}_m^t)-D_{m'}^t(\boldsymbol{x}_{m'}^t)|\leqslant\delta_{\mathrm{d}};$$

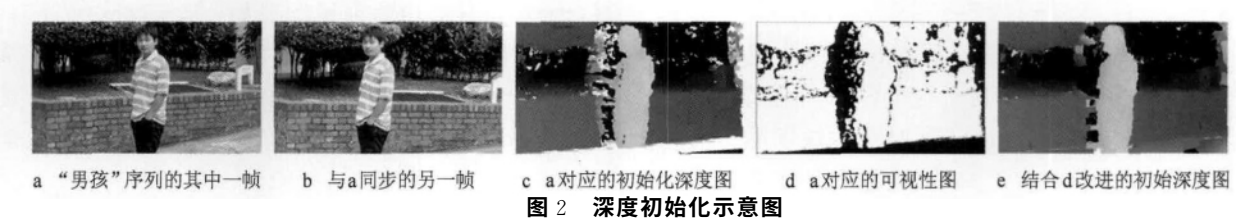
其中，$V_{m\to m'}^t(\boldsymbol{x}_m^t)$ 表示像素点 $\boldsymbol{x}_m^t$ 在 $I_{m'}^t$ 中是否可见（1 表示可见，0 表示不可见），δ_{d} 代表视差一致性阈值. $Z_{m\to m'}^t(\boldsymbol{x}_m^t)$ 表示 $\boldsymbol{x}_m^t$ 根据其深度反投影至 3D 空间然

后投影至 m' 中的深度. $\mathcal{V}_m^t(\boldsymbol{x}_m^t)$ 定义为 $\boldsymbol{x}_m^t$ 的"总体可见性",如果 $\boldsymbol{x}_m^t$ 在所有其余同步帧中均不可见,$\mathcal{V}_m^t(\boldsymbol{x}_m^t)=0$;否则,$\mathcal{V}_m^t(\boldsymbol{x}_m^t)=1$. 图 2 d 所示为图 2 a 关于图 2 b 的可视性图,可以将可视性图引入式(2)以改进初始化深度图,仅对那些总体可见性 $\mathcal{V}_m^t(\boldsymbol{x}_m^t)=1$ 的像素点计算数据惩罚值. 对于 $\mathcal{V}_m^t(\boldsymbol{x}_m^t)=0$ 的像素点,利用相邻点的深度来进行填补,而不是利用式(2)计算数据惩罚值.

首先利用 Mean-shift[22] 对每帧进行过分割,并为每个分割区域利用其中可见的像素点($\mathcal{V}_m^t(\boldsymbol{x})=1$)来拟合参数为$[a,b,c]$的 3D 平面,其余不可见点的视差计算方法为 $D(\boldsymbol{x})=ax+by+c$;然后用

$$E_{\mathrm{D}}^1(\boldsymbol{x}_m^t,D_m^t)=\sum_{\boldsymbol{x}_m^t}\frac{\sigma_{\mathrm{d}}}{\sigma_{\mathrm{d}}+|D(\boldsymbol{x}_m^t)-D_m^t(\boldsymbol{x}_m^t)|}$$

计算不可见点的数据惩罚值,其中 σ_{d} 控制数据项对于拟合平面视差差异的敏感度. 可以引入不可见点的数据惩罚值进行能量优化,从而改进初始化深度图,并利用改进的深度图来重新计算可见性图. 因此可对深度图和可见性图迭代地进行计算,在实验中通常经过 2 次迭代就足以获得较可靠的初始化结果,图 2 e 展示了改进后的初始化深度图.



a "男孩"序列的其中一帧 b 与a同步的另一帧 c a对应的初始化深度图 d a对应的可视性图 e 结合d改进的初始深度图

图 2 深度初始化示意图

4 双层分割

深度初始化之后,本文将每帧的像素点分为静态和动态 2 类,使得不同类像素点的深度可以用不同的时空一致性约束来进行优化,以实现高质量的 3D 重建.

4.1 初始化分割

利用初始深度估计可以将像素点投影至相邻帧验证颜色和深度一致性. 利用

$$P_{\mathrm{d}}(\boldsymbol{x}_m^t)=\frac{\sum_{(m',t')\in N(m,t)}\mathcal{C}_{m\to m'}^{t\to t'}(\boldsymbol{x}_m^t)=\mathrm{dynamic}}{|N(m,t)|} \quad (3)$$

计算每个像素点为动态的概率值. 其中,$N(m,t)$ 为 (m,t) 在所有 M 个序列中的相邻帧集;$\mathcal{C}_{m\to m'}^{t\to t'}(\boldsymbol{x}_m^t)$ 为一个启发式函数,根据 $\boldsymbol{x}_m^t$ 在 (m',t') 帧中投影点 $\boldsymbol{x}_{m'}^{t'}$ 的颜色和深度来决定 $\boldsymbol{x}_m^t$ 是否为动态. 比较 $\boldsymbol{x}_m^t$ 和 $\boldsymbol{x}_{m'}^{t'}$ 的深度:如果 $Z_{m\to m'}^{t\to t'}(\boldsymbol{x}_m^t)$ 远大于 $Z(\boldsymbol{x}_{m'}^{t'})$(即 $\boldsymbol{x}_m^t$ 投影至 $\boldsymbol{x}_{m'}^{t'}$ 后面),则很可能 $\boldsymbol{x}_m^t$ 属于静态背景区域并在 (m',t') 帧中被遮挡;如果 $Z_{m\to m'}^{t\to t'}(\boldsymbol{x}_m^t)$ 远小于 $Z(\boldsymbol{x}_{m'}^{t'})$($\boldsymbol{x}_m^t$ 投影至 $\boldsymbol{x}_{m'}^{t'}$ 前方),则 $\boldsymbol{x}_m^t$ 必定是动态前景点并且遮挡住其他像素点. 如果两者深度相近,则比较其颜色. 如果 $I_m^t(\boldsymbol{x}_m^t)$ 和 $I_{m'}^{t'}(\boldsymbol{x}_{m'}^{t'})$ 亦相似,则认为 $\boldsymbol{x}_m^t$ 为静态,否则为动态.

对于不可见的像素点($\mathcal{V}_m^t(\boldsymbol{x}_m^t)=0$),由于通过平面拟合初始化的深度并不完全可靠,因此不能用于投影验证. 对于这些像素点,可利用类似文献[14]的方法统计局部窗口内从其余帧映射得到的背景颜色并创建高斯混合模型,然后统计最大的高斯分布颜色概率作为被遮挡像素的动态概率值.

利用式(3)可以为每帧计算动态概率图. 图 3 a 所示为图 2 a 对应的动态概率图,其中包含了很多噪声. 为改进分割结果,本文用 Mean-shift 分割对动态概率图进行规整化. 对于每个分割区域统计动态概率大于 η_{p}(实验中取 0.4)的像素点所占百分比,如果此百分比大于 η_{s}(实验中取 0.5),则认为整个分割区域均为动态;否则为静态. 经过规整化后可以获得初始的静动态分割,如图 3 b 所示.

a 图2a动态概率图 b Mean-shift规整化 c 时序优化结果 d GrabCut优化结果 e 与a~d对应的局部放大图

图 3 静动态分割示意图

4.2 时序优化

双层分割初始化完毕后,通过统计多个相邻帧上的分割信息来优化当前帧分割. 对于每个序列,利用文献[23]算法计算向前帧和向后帧的光流,并通

过反向验证摒弃不准确的光流信息. 假设 x_m^t 利用光流 $O_m^{t\to t+1}(x_m^t)$ 跟踪至对应点 $\hat{x}_m^{t+1}$，将 $\hat{x}_m^{t+1}$ 利用反向光流 $O_m^{t+1\to t}(\hat{x}_m^{t+1})$ 跟踪至第 t 帧，并用

$$\varepsilon(x_m^t) = \| O_m^{t\to t+1}(x_m^t) + O_m^{t+1\to t}(\hat{x}_m^{t+1}) \|$$

计算光流误差. 如果 $\varepsilon(x_m^t) > \eta_o$（实验中 η_o 取 2 个像素单位），则认为光流不准确并且停止跟踪. 对于每个像素点验证其在相邻时序帧上对应点的分割标记，由此计算当前点的时序动态概率

$$P'_d(x_m^t) = \frac{\sum_{t'\in N(t)} S_m^{t'}(x_m^t + O_m^{t\to t'}(x_m^t)) = \text{dynamic}}{|N(t)|};$$

其中，$N(t)$ 是 t 的相邻时刻集，$S_m^t(x)$ 表示 (m,t) 帧中 x 的静动态标记. 简言之，$P'_d(\cdot)$ 统计时序对应点集中标记为动态像素点的比例.

下面定义能量函数

$$E_S(S_m^t; P'_d, I_m^t) = E_S^1(S_m^t; P'_d) + E_S^2(S_m^t; I_m^t) \quad (4)$$

求解分割问题. 其中 E_S^1 为数据项，E_S^2 为平滑项. E_S^1 定义为

$$E_S^1(S_m^t; P'_d) = \sum_{x_m^t} e_S(S_m^t(x_m^t)),$$

其中，

$$e_S(S_m^t(x_m^t)) = \begin{cases} -\log(1 - P'_d(x_m^t)), & S_m^t(x_m^t) = \text{static} \\ -\log(P'_d(x_m^t)), & S_m^t(x_m^t) = \text{dynamic} \end{cases}.$$

E_S^2 使得分割边界与颜色变化更为一致，定义为

$$E_S^2(S_m^t; I_m^t) = \lambda \sum_x \sum_{y\in N(x)} \frac{|S_m^t(x) - S_m^t(y)|}{1 + \| I_m^t(x) - I_m^t(y) \|_2}.$$

利用图切割算法[24]求解式(4)从而得到更为一致的静动态分割. 图 3 c 所示为优化后的分割结果，其中少数几处分割边界仍存在噪声. 为了更进一步改进分割边界，对目前的分割结果进行腐蚀和膨胀（实验中腐蚀和膨胀半径均为 2）从而创建 trimap，并利用 GrabCut 算法[25]优化分割结果. 经 GrabCut 优化后边界噪声被进一步消除，如图 3 d 所示.

5　时空一致性深度优化

给定静态和动态像素点的分割之后，对于静动态像素点利用不同的方法进行深度优化. 对于静态像素点，可以利用 Bundle Optimization 技术[2]有效地进行深度优化. 对于动态像素点，采用类似文献[12,17]的方法，即利用多个相邻时刻的视频帧来加强颜色/几何一致性统计的鲁棒性.

假设候选视差为 d_i，参考相机为 m'，可以在 t 时刻将 x_m^t 从相机 m 投影至 m'，投影点记为 $x_{m'}^t$. 为提高动态像素点深度求解的鲁棒性，估计 x_m^t 和 $x_{m'}^t$ 在相邻帧上对应点的颜色和几何一致性. 如图 4 所示，利用光流将 x_m^t 和 $x_{m'}^t$ 跟踪至 t' 时刻得到对应点 $\hat{x}_m^{t'}$ 和 $\hat{x}_{m'}^{t'}$. 如果光流跟踪准确，则计算 $\hat{x}_m^{t'}$ 和 $\hat{x}_{m'}^{t'}$ 的颜色和几何一致性[17]

$$L_g(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'}) = p_c(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'}) p_g(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'}).$$

其中，$p_c(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'})$ 用于衡量 $\hat{x}_m^{t'}$ 和 $\hat{x}_{m'}^{t'}$ 的颜色一致性，定义为

$$p_c(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'}) = \frac{\sigma_c}{\sigma_c + \| I_m^t(\hat{x}_m^{t'}) - I_{m'}^{t'}(\hat{x}_{m'}^{t'}) \|_1},$$

σ_c 控制颜色差异的敏感度；$p_g(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'})$ 为几何一致性，定义为

$$p_g(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'}) = \frac{\sigma_g}{\sigma_g + d_g(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'}; D_m^{t'}, D_{m'}^{t'})},$$

d_g 为 $\hat{x}_m^{t'}$ 和 $\hat{x}_{m'}^{t'}$ 之间的对称投影误差. 将 $\hat{x}_m^{t'}$ 投影至 (m',t') 帧并计算投影点与 $\hat{x}_{m'}^{t'}$ 的距离，同样将 $\hat{x}_{m'}^{t'}$ 投影至 (m,t') 帧并计算与 $\hat{x}_m^{t'}$ 的距离，d_g 为两者的平均距离.

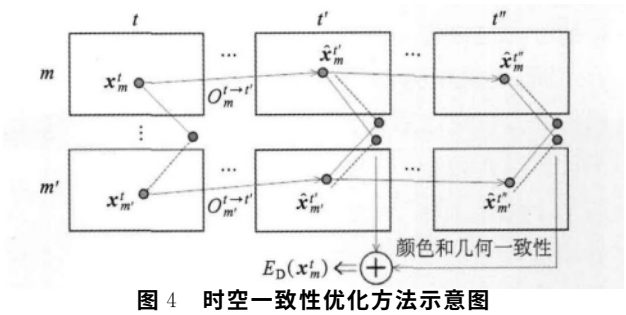


图 4　时空一致性优化方法示意图

假设相邻时刻帧的深度是正确的，如果 d_i 是真实深度，则 $L_g(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'})$ 会很大；否则便会很小. 如同文献[17]，本文累积相邻时刻帧（最邻近 10 帧）的颜色和几何一致性，由此重新定义动态像素点的数据项，定义为

$$E_D^1(D_m^t; \mathcal{I}, \mathcal{D}) = \sum_{x_m^t} \left(1 - \frac{\sum_{t'\in N(t)} \sum_{m'\neq m} L_g(\hat{x}_m^{t'}, \hat{x}_{m'}^{t'})}{(M-1)|N(t)|}\right) \quad (5)$$

其中 $N(t)$ 表示最邻近 10 帧的集合. 利用多个相邻时刻帧统计的颜色和几何一致性能够很好地推断正确的深度值，从而大大提高了本文优化方法的鲁棒性. 将式(5)代入式(1)并重新求解以优化每帧的深度图. 值得注意的是，在优化当前帧深度图时保持其余帧的深度不变. 经过优化之后，静态和动态深度值的准确性将会大大提高，而且在时序上会更加一致.

许多现有的方法[6,11]只是简单地对相邻时刻帧上对应像素的深度值采用线性插值或曲线拟合的方式进行平滑，以此加强深度的时空一致性；然而此类优化方法并不鲁棒（如图 5 d 所示），因为简单的平滑

处理并不能从本质上推断真实的深度值,而且对于错误的初始深度和不准确的光流估计比较敏感.需要指出的是,文献[12]由于计算复杂性时只添加了前一时刻的一致性约束,因而对于较少数目宽基线的摄像机序列不能够保证鲁棒性,并且不可见区域的深度无法利用较少数目的摄像机进行有效估计,如图 5 f 所示.图 5 所示为本文方法的一些中间结果,以及与这些方法的结果比较.

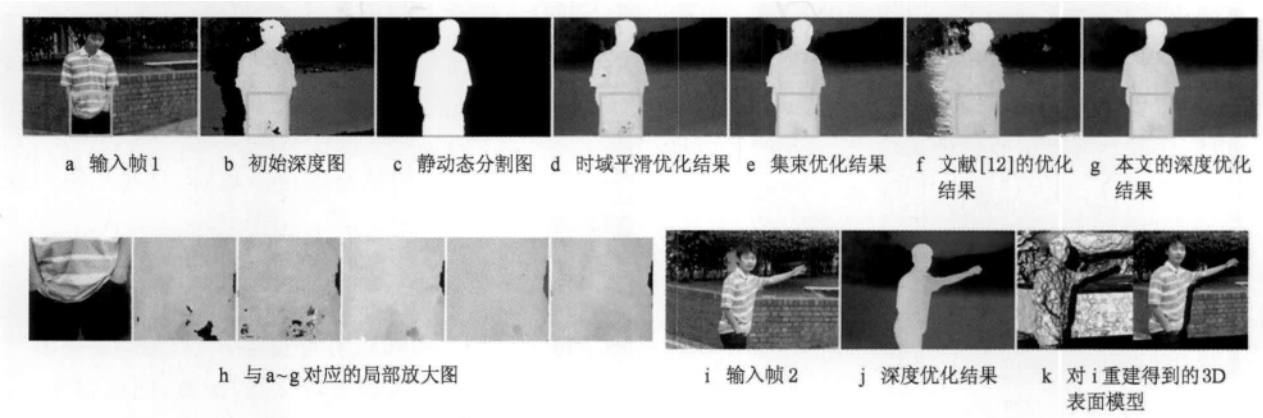


图 5 时空一致性深度优化结果

利用时空一致性优化后的深度值结合第 4 节所述的方法可以进一步优化静动态分割.与第 4 节略有不同,此时在优化分割的过程中不考虑遮挡问题,因为经过优化之后被遮挡像素点的深度值亦可以用来计算动态概率.本文交替迭代地执行如图 1 所示的 2 个优化步骤,从而获得最终优化后的深度图和静动态分割结果.在实验中通常经过 2 次迭代就足以达到收敛.

6 实验及结果分析

本文对由 2～3 个手持摄像机拍摄的视频序列进行了实验.拍摄开始时在视域范围内放置一个闪烁光源,并在实验数据处理时利用闪光时刻作为标记来同步不同序列之间的视频帧.所有实验都是在主频为 2.83 GHz 的 4 核 CPU 上运行的.对于分辨率为 960×540 的序列,静动态分割平均每帧需要 6 s,初始化每帧耗时 175 s,深度优化每帧耗时 40 s,比现有的许多方法[7,12]更为高效.图 6 所示为一组捕获一只爬行的棕熊的实例,并通过与其他方法的比较展现本文时空一致性深度优化方法的鲁棒性.如图 6 b,6 e 所示,本文方法能够有效地纠正棕熊背部的错误初始深度,并精确地重建出棕熊的复杂运动.图 7 所示为"女孩"序列的结果,通过时空一致性约束迭代优化之后,能够获得高质量的动态深度图结果以及准确的静动态分割.如图 7 b 所示,本文方法有效地还原了女孩的身体运动和裙子摆荡的动态深度细节.本文的方法还能够处理多个动态物体,如图 8 所示的一个很有挑战性的例子中总共有 2 个视频序列,场景中含有 3 位行人,可以看出,本文方法同时恢复出了静态区域和每个动态行人的准确深度值.图 9 所示为一组 3 个摄像机拍摄的序列,可以看出,本文方法在运动较快的条件下亦能够产生高质量的重建结果.

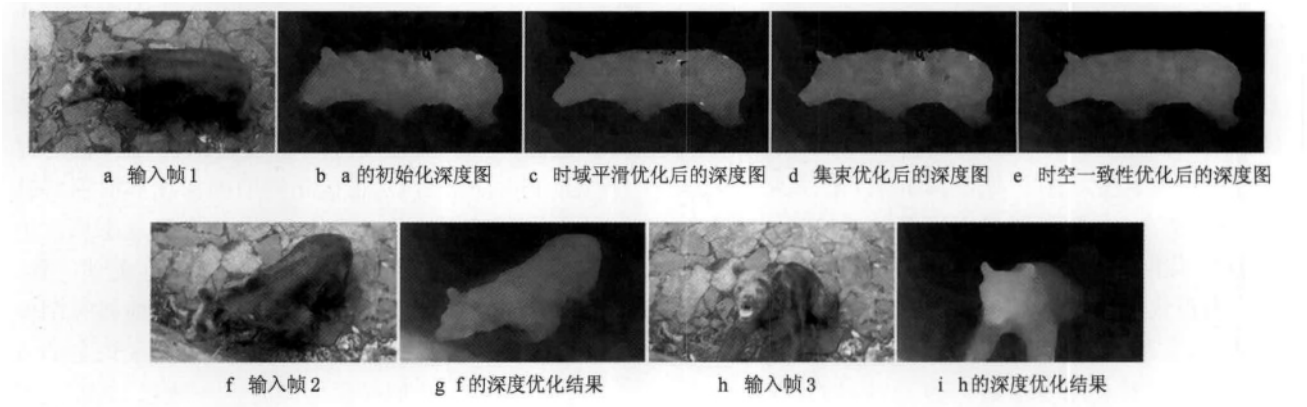


图 6 "棕熊"序列的时空一致性深度优化结果

a　2个视频序列的其中3组同步帧以及序列1的静动态分割和深度优化结果

b　与a方框区域对应的局部放大图

图 7　“女孩”序列的动态 3D 重建结果

a　原始图像1　　b　a的深度恢复结果　　c　原始图像2　　d　c的深度恢复结果　　e　对a深度重建得到的场景模型

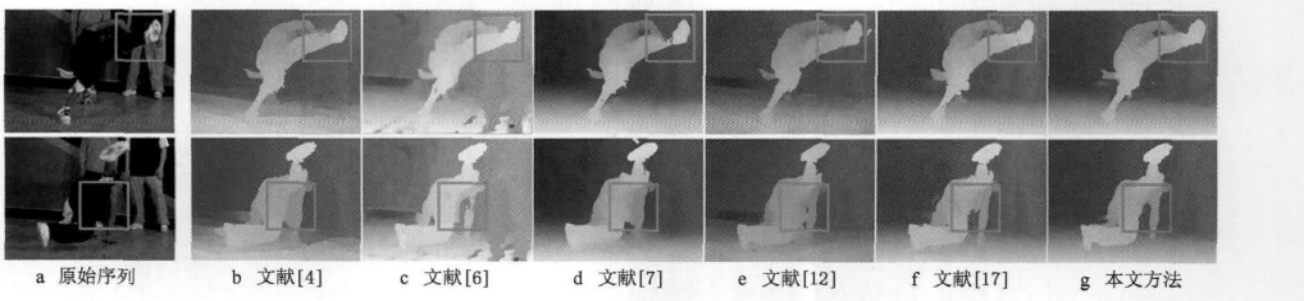
图 8　“3 个行人”实例的动态 3D 重建结果

a　3个输入视频序列的其中一组同步帧　　b　与a中序列2对应的深度图

c　另外2帧的深度恢复结果

图 9　“格斗”序列的深度恢复结果

本文还对微软研究院的街舞数据[4]进行了实验. 图 10 所示为本文方法与文献[4,6-7,12,17]的比较结果，如方框标记的区域所示，文献[4,6]重建的地面和文献[7,12]重建的舞蹈者身体部分均有明显的



a　原始序列　　b　文献[4]　　c　文献[6]　　d　文献[7]　　e　文献[12]　　f　文献[17]　　g　本文方法

图 10　6 种方法的结果比较

问题，文献[17]的深度结果在不连续边界附近存在少量噪声，相比之下，本文的深度恢复结果在不连续边界附近更加准确.

7 结 语

本文提出一种新的基于多目手持摄像机的动态场景稠密深度恢复方法，首先利用多序列同步帧计算初始深度，然后根据多帧上的颜色和几何一致性统计将像素点分为静态和动态 2 类，并进行各自不同的深度优化；分割结果和深度优化迭代地进行，最终实现高质量的深度恢复.

如果运动物体和背景的颜色非常相似，本文的双层分割方法可能无法准确地区分出静态和动态像素，这也是双层分割方法普遍存在的问题；如果动态物体出现严重的自遮挡情况，本文的双层分割对于自遮挡区域像素点的静动态判断并不可靠. 如何消除这些歧义性以获得更准确的双层分割结果，还需要进一步研究和改进；此外，如何更有效地避免错误光流的干扰，并把光流结合深度和分割进行整体求解，也是我们未来的一个研究方向.

参考文献(References)：

[1] Seitz S M, Curless B, Diebel J, *et al*. A comparison and evaluation of multi-view stereo reconstruction algorithms [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2006, 1: 519-528

[2] Zhang G F, Jia J Y, Wong T T, *et al*. Consistent depth maps recovery from a video sequence [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(6): 974-988

[3] Kong Xiangli, Zhang Guofeng, Hua Wei. Detail preserving hierarchical multi-view stereo matching via global optimization [J]. Journal of Computer-Aided Design & Computer Graphics, 2011, 23(1): 177-184 (in Chinese)
(孔相澧，章国锋，华 炜. 基于全局优化的保细节分层多视图立体匹配[J]. 计算机辅助设计与图形学学报, 2011, 23(1): 177-184)

[4] Zitnick C L, Kang S B, Uyttendaele M, *et al*. High-quality video view interpolation using a layered representation [J]. ACM Transactions on Graphics, 2004, 23(3): 600-608

[5] Gong M L. Enforcing temporal consistency in real-time stereo estimation [M] //Lecture Notes in Computer Science. Heidelberg: Springer, 2006, 3953: 564-577

[6] Larsen E S, Mordohai P, Pollefeys M, *et al*. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation [C] //Proceedings of the 11th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2007: 1-8

[7] Lei C, Chen X D, Yang Y H. A new multiview spacetime-consistent depth recovery framework for free viewpoint video rendering [C] //Proceedings of the 12th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2009: 1570-1577

[8] Goldlücke B, Magnor M A. Joint 3D-reconstruction and background separation in multiple views using graph cuts [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2003,1: 683-688

[9] Guillemaut J Y, Kilner J, Hilton A. Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes [C] //Proceedings of the 12th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2009: 809-816

[10] Zhang Y, Kambhamettu C. Integrated 3D scene flow and structure recovery from multiview image sequences [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2000, 2: 674-681

[11] Tao H, Sawhney H S, Kumar R. Dynamic depth recovery from multiple synchronized video streams [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2001, 1: 118-124

[12] Yang M J, Cao X, Dai Q H. Multiview video depth estimation with spatialtemporal consistency [C] //Proceedings of the British Machine Vision Conference. Manchester: BMVA Press, 2010: 67.1-67.11

[13] Yang W Z, Zhang G F, Bao H J, *et al*. Consistent depth maps recovery from a trinocular video sequence [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 1466-1473

[14] Zhang G F, Jia J Y, Hua W, *et al*. Robust bilayer segmentation and motion/depth estimation with a handheld camera [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 603-617

[15] Liu F, Gleicher M. Learning color and locality cues for moving object detection and segmentation [C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2009: 320-327

[16] Zhong F, Qin X Y, Peng Q S. Transductive segmentation of live video with non-stationary background [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2010: 2189-2196

[17] Jiang H Q, Liu H M, Tan P, *et al*. 3D reconstruction of dynamic scenes with multiple handheld cameras [C] // Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2012: 601-615

[18] Zhang G F, Dong Z L, Jia J Y, *et al*. Efficient non-consecutive feature tracking for structure-from-motion[C] // Proceedings of the 11th European Conference on Computer Vision: Part V. Heidelberg: Springer, 2010: 422-435

[19] Zhang G F, Qin X Y, Hua W, *et al*. Robust metric reconstruction from challenging video sequences [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2007: 1-8

[20] Tola E, Lepetit V, Fua P. Daisy: an efficient dense descriptor applied to wide-baseline stereo [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(5): 815-830

[21] Felzenszwalb P F, Huttenlocher D P. Efficient belief propagation for early vision [J]. International Journal of Computer Vision, 2006, 70(1): 41-54

[22] Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619

[23] Liu C. Beyond pixels: exploring new representations and applications for motion analysis [D]. Cambridge: Massachusetts Institute of Technology, 2009

[24] Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(11): 1222-1239

[25] Rother C, Kolmogorov V, Blake A. "GrabCut": interactive foreground extraction using iterated graph cuts [J]. ACM Transactions on Graphics, 2004, 23(3): 309-314