浙江大學

硕 士 研 究 生 读 书 报 告

题目 Tracking Everything Everywhere All at Once

作者姓名 马少杰

作者学号 22351146

指导教师 李启雷

学科专业 电子信息

所在学院 软件学院

提交日期 二〇二四年一月八日

Tracking Everything Everywhere All at Once

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Electronic Information

Advisor: Li Qilei

By

Shaojie Ma

Zhejiang University, PR China

2024.01.08

摘要

这篇文章提出了一种新的方法，用于估计视频序列中的密集和长程运动。先前的光流或粒子视频跟踪算法通常在有限的时间窗口内操作，很难通过遮挡进行跟踪并保持估计的运动轨迹的全局一致性。本文提出了一种完整且全局一致的运动表示，称为 OmniMotion，允许准确地对视频中每个像素进行完整长度的运动估计。OmniMotion 使用准 3D 标准体来表示视频，并通过局部空间和标准空间之间的双射进行像素级的跟踪。这种表示方式使我们能够确保全局一致性，通过遮挡进行跟踪，并对相机和物体运动的任何组合进行建模。结果显示我们的方法在定量和定性上均大幅优于先前的最新方法。

关键词：运动估计； 辐射场； 3D 表示

Abstract

We present a new test-time optimization method for estimating dense and long-range motion from a video sequence. Prior optical flow or particle video tracking algorithms typically operate within limited temporal windows, struggling to track through occlusions and maintain global consistency of estimated motion trajectories. We propose a complete and globally consistent motion representation, dubbed OmniMotion, that allows for accurate, full-length motion estimation of every pixel in a video. OmniMotion represents a video using a quasi-3D canonical volume and performs pixel-wise tracking via bijections between local and canonical space. This representation allows us to ensure global consistency, track through occlusions, and model any combination of camera and object motion. Extensive evaluations on the TAP-Vid benchmark and real-world footage show that our approach outperforms prior state-of-the-art methods by a large margin both quantitatively and qualitatively.

Keywords： Motion estimation; Radiation fields; 3D representation

1. Introduction

运动估计方法传统上遵循两种主要方法：稀疏特征追踪和密集光流。尽管每种方法都在各自的应用中证明了有效性，但两种表示都没有完全模拟视频的运动：成对光流无法捕捉长时间窗口内的运动轨迹，而稀疏追踪则不能模拟所有像素的运动。

一些方法试图填补这一差距，即在视频中估计密集和长程像素轨迹。这些方法范围从简单地链接两帧光流场到最近的一些方法，直接预测跨多帧的每个像素的轨迹。然而，这些方法在估计运动时都使用了有限的上下文，忽视了在时间或空间上远离的信息。生成密集和长程轨迹仍然是该领域的一个未解决问题，其中存在三个关键挑战：（1）在长序列中保持准确的轨迹，（2）在遮挡情况下跟踪点，以及（3）在空间和时间上保持一致性。

在这项工作中，我们提出了一种综合方法来估计视频运动，使用视频中的所有信息来联合估计每个像素的完整长度的运动轨迹。我们的方法被称为 OmniMotion，使用了准三维表示，其中一个规范的 3D 体积通过一组局部规范双射映射到每帧本地体积。这些双射作为动态多视图几何的灵活松弛，模拟了相机和场景运动的组合。我们的表示保证了循环一致性，并且可以跟踪所有像素，即使在遮挡时也可以实现。我们通过视频优化我们的表示来联合解决整个视频的运动。优化后，我们的表示可以在视频中的任何连续坐标处进行查询，以获取跨越整个视频的运动轨迹。

总之，我们提出了一种方法：1）为整个视频中的所有点产生全局一致的完整长度的运动轨迹，2）可以在遮挡时跟踪点，3）可以处理具有任意组合的相机和场景运动的野外视频。我们在 TAP 视频跟踪基准测试上定量地展示了这些优势，通过显著的优势超越了所有先前的方法，实现了业内领先的性能。

2. Overview

我们提出一种优化方法，用于从视频序列中估计密集而长程的运动。我们的方法以一组帧和成对的有噪运动估计（例如光流场）作为输入，并利用这些来求解整个视频的完整、全局一致的运动表示。一旦优化完成，我们的表示可以查询任何帧中的任何像素，以产生整个视频中平滑准确的运动轨迹。我们的方

法能够识别出点是否被遮挡，并且甚至能够在遮挡情况下跟踪点。

3. OmniMotion representation

传统的运动表示，如成对光流，在物体被遮挡时会丢失目标的轨迹，并且在多个帧上叠加对应关系时会产生不一致性。为了通过遮挡获取准确、一致的轨迹，因此我们需要一种“全局”运动表示，即一种编码场景中所有点轨迹的数据结构。其中一种全局表示是将场景分解为一组离散的、深度分离的层次结构。然而，大多数实际场景不能被表示为一组固定的、有序的层次结构，例如在物体在三维空间中旋转的简单情况。另一种极端情况是完整的三维重建，即将三维场景几何、相机位姿和场景运动区分开。

OmniMotion 将视频中的场景表示为一个规范的三维体积，通过局部-规范双射将其映射到每帧的局部体积中。局部-规范双射的参数化采用神经网络，并捕捉摄像机和场景的运动而不对两者进行解耦。因此，该视频可以被视为从一个固定的、静态摄像机渲染出来的局部体积。由于 OmniMotion 没有明确地解耦相机和场景的运动，所以得到的表示不是一个物理上准确的三维场景重建。相反，通过在每个局部帧和一个规范帧之间建立双射，OmniMotion 保证所有局部帧之间的全局循环一致的三维映射，这模拟了真实世界的度量三维参考帧之间的一对一对应关系。其次，OmniMotion 保留了被投影到每个像素上的所有场景点的信息以及它们的相对深度顺序，使得即使这些点暂时从视野中被遮挡，也能够跟踪它们。

3.1. Canonical 3D volume

OmniMotion 使用规范体积 G 来表示视频的内容，它作为观察场景的三维地图。与 NeRF 中一样，我们在 G 上定义了一个基于坐标的网络 F_θ，将每个规范的三维坐标 u $\in$ G 映射到密度 σ 和颜色 c。在 G 中存储的密度非常关键，因为它告诉我们在规范空间中的表面位置。配合三维双射，这使得我们能够跟踪多个帧上的表面并推断遮挡关系。在 G 中存储的颜色能够在优化过程中计算光度损失。

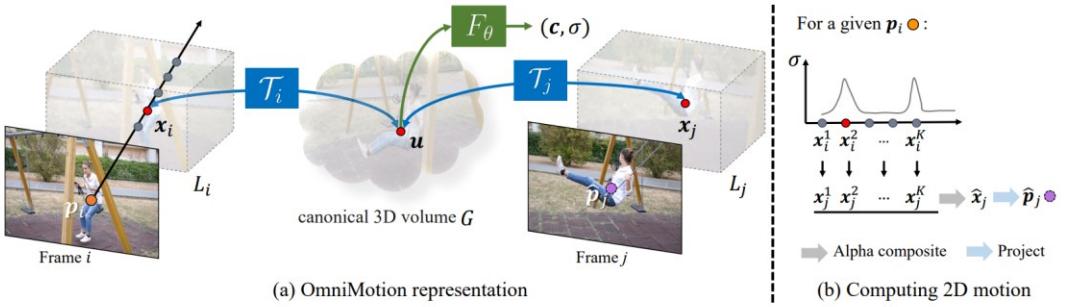

图 2: Method overview. (a) Our OmniMotion representation is comprised of a canonical 3D volume G and a set of bijections $\mathcal{T}_i$ that map between each frame's local volume L_i and the canonical volume G. Any local 3D location $\boldsymbol{x}_i$ in frame i can be mapped to its corresponding canonical location $\boldsymbol{u}$ through $\mathcal{T}_i$, and then mapped back to another frame j as $\boldsymbol{x}_j$ through the inverse mapping $\mathcal{T}_j^{-1}$. Each location $\boldsymbol{u}$ in G is associated with a color $\boldsymbol{c}$ and density σ, computed using a coordinate-based MLP $F_{\boldsymbol{\theta}}$. (b) To compute the corresponding 2D location for a given query point $\boldsymbol{p}_i$ mapped from frame i to j, we shoot a ray into L_i and sample a set of points $\{\boldsymbol{x}_i^k\}_{k=1}^K$, which are then mapped first to the canonical space to obtain their densities, and then to frame j to compute their corresponding local 3D locations $\{\boldsymbol{x}_j^k\}_{k=1}^K$. These points $\{\boldsymbol{x}_j^k\}_{k=1}^K$ are then alpha-composited and projected to obtain the 2D corresponding location $\hat{\boldsymbol{p}}_j$.

3.2. 3D bijections

我们定义了一个连续双射映射 T_i，将每个局部坐标系 L_i 中的 3D 点 x_i 映射到规范的 3D 坐标系中，即 $\mathrm{u} = \mathrm{T}_i(\mathrm{x}_i)$，其中 i 是坐标系的索引。规范坐标 u 是与时间无关的，可以被视为特定场景点或 3D 轨迹的全局一致的“索引”。通过组合这些双射映射及其反函数，我们可以将一个 3D 点从一个局部 3D 坐标系 (L_i) 映射到另一个 (L_j)：

$$\mathrm{x}_j = \mathrm{T}_j^{-1} \circ \mathrm{T}_i(\mathrm{x}_i). \tag{1}$$

双射映射确保个别帧中的 3D 点之间的对应关系具有循环一致性，因为它们源自同一标准点。为了允许捕捉真实世界运动，我们将这些双射参数化为可逆神经网络。Real-NVP 通过合成称为仿射耦合层的简单双射变换来构建双射映射。

3.3. Computing frame-to-frame motion

基于这种表示法，我们现在描述如何计算帧 i 中任何查询像素 p_i 的 2D 运动。直观地说，我们通过对光线上的点进行采样，将查询像素提升到 3D，使用双射 Ti 和 Tj 将这些 3D 点映射到目标帧 j，通过 alpha 混合从不同采样点渲染这些映射的 3D 点，最后投影回 2D 以获得推测的对应关系。

具体而言，由于我们假设相机运动由局部规范双射 Ti 包含，所以我们只使用一个固定的正交相机。然后，p_i 处的光线可以定义为 $\mathrm{r}_i(z) = \mathrm{o}_i + z\mathrm{d}$，其

中 $o_i = [p_i, 0]$，$d = [0,0,1]$。我们在光线上采样 K 个样本点 $\{x^k_i\}$，这相当于将一组深度值$\{z^k_i\}_{k=1}^K$附加到 p_i 上。尽管它不是真正的相机光线，但它捕捉了每个像素的多个表面的概念，并足以处理遮挡。

接下来，我们通过将样本点映射到规范空间，并查询密度网络 F_θ 来获得这些样本的密度和颜色。以第 k 个样本 x^k_i 为例，其密度和颜色可以写为 $(\sigma_k, c_k) = F_\theta(M_\theta(x^k_i; \psi_i))$。我们还可以将沿着光线的每个样本映射到帧 j 中相应的 3D 位置 x^k_j（公式 1）。

现在，我们可以聚合所有样本的对应关系 x^k_j 来生成单个对应关系 $\hat{x}_j$。我们使用 alpha 混合，其中第 k 个样本的 alpha 值为 $\alpha_k = 1 - \exp(-\sigma_k)$。然后，我们计算 $\hat{x}_j$ 如下：

$$T_k = \prod_{l=1}^{k-1}(1-\alpha_l) \tag{2}$$

通过类似的过程，我们使用复合函数将 c^k 得到图像空间中的颜色 $\hat{C}_i$，用于查询像素 p_i。然后，我们使用静止正交摄像机模型对 $\hat{x}_j$ 进行投影，得到查询位置 p_i 的预测二维对应位置 $\hat{p}_j$。

3.4. Loss functions

我们的主要损失函数是流量损失。我们最小化经过优化的表示产生的预测流量 $\hat{f}_{i\to j} = \hat{p}_j - p_i$ 与通过光流运行导出的监督输入流量 $f_{i\to j}$ 之间的平均绝对误差（MAE）。

$$\mathcal{L}_{\text{flo}} = \sum_{\boldsymbol{f}_{i\to j}\in\Omega_f} ||\hat{\boldsymbol{f}}_{i\to j} - \boldsymbol{f}_{i\to j}||_1 \tag{3}$$

其中，Ω_f 是所有被滤波的成对流的集合。此外，我们最小化一个光度损失，定义为预测颜色 $\hat{C}_i$ 与源视频帧中观察到的颜色 C_i 之间的均方误差（MSE）。

$$\mathcal{L}_{\text{pho}} = \sum_{(i,\boldsymbol{p})\in\Omega_p} ||\hat{\boldsymbol{C}}_i(\boldsymbol{p}) - \boldsymbol{C}_i(\boldsymbol{p})||_2^2 \tag{4}$$

其中，Ω_p 是所有帧上所有像素位置的集合。最后，为了确保由 M_θ 估计的 3D 运动的时间平滑性，我们应用了一个正则化项，惩罚大加速度。给定帧 i 中的

采样的 3D 位置 x_i，我们使用公式 1 将其映射到帧 $i-1$ 和帧 $i+1$，得到分别为 x_i-1 和 x_{i+1} 的 3D 点，并且像中那样最小化 3D 加速度。

$$\mathcal{L}_{\text{reg}} = \sum_{(i,\boldsymbol{x})\in\Omega_x} ||\boldsymbol{x}_{i+1} + \boldsymbol{x}_{i-1} - 2\boldsymbol{x}_i||_1 \tag{5}$$

其中，Ω_x 是所有帧的局部 3D 空间的并集。我们的最终联合损失可以写成：

$$L = Lflo + \lambda phoLpho + \lambda regLreg \tag{6}$$

权重 λ 控制每个项的相对重要性。

这种优化的直观理念是利用单一的规范体积 G 的双射、光照一致性以及基于坐标的网络 M_0 和 F_0 所提供的自然时空平滑性，从而调节不一致的成对流量并填补对应图中的缺失内容。

4. Evaluation

4.1. Benchmarks

我们在 TAP-Vid 基准测试上评估了我们的方法，该基准测试旨在评估长视频剪辑中的点跟踪性能。TAP-Vid 包括具有准确人工注释的真实世界视频的点轨迹和具有完美地面真值的合成视频。每个点轨迹都在整个视频中进行注释，当不可见时标记为遮挡。

数据集。 我们在以下来自 TAP-Vid 的数据集上进行评估：

1) DAVIS，来自 DAVIS 2017 验证集的 30 个真实数据集，剪辑的帧数范围为 34-104，每个视频平均有 21.7 个点轨迹注释。

2) Kinetics，来自 Kinetics-700-2020 验证集的 1,189 个真实数据集，每个视频有 250 帧，每个视频平均有 26.3 个点轨迹注释。为了使测试时间优化方法（如我们的方法）的评估变得可行，我们随机抽取了 100 个视频的子集，并在该子集上评估所有方法。

3) RGB-Stacking，一个合成数据集，包含 50 个每个有 250 帧的视频和 30 个轨迹。

4.2. Comparisons

Method	Kinetics				DAVIS				RGB-Stacking			
	AJ ↑	$< \delta^x_{avg}$ ↑	OA ↑	TC ↓	AJ ↑	$< \delta^x_{avg}$ ↑	OA ↑	TC ↓	AJ ↑	$< \delta^x_{avg}$ ↑	OA ↑	TC ↓
RAFT-C [66]	31.7	51.7	84.3	0.82	30.7	46.6	80.2	0.93	42.0	56.4	91.5	0.18
RAFT-D [66]	50.6	66.9	85.5	3.00	34.1	48.9	76.1	9.83	72.1	85.1	92.1	1.04
TAP-Net [15]	48.5	61.7	86.6	6.65	38.4	53.4	81.4	10.82	61.3	73.7	91.5	1.52
PIPs [23]	39.1	55.3	82.9	1.30	39.9	56.0	81.3	1.78	37.3	50.6	89.7	0.84
Flow-Walk-C [5]	40.9	55.5	84.5	0.77	35.2	51.4	80.6	0.90	41.3	55.7	92.2	0.13
Flow-Walk-D [5]	46.9	65.9	81.8	3.04	24.4	40.9	76.5	10.41	66.3	82.7	91.2	0.47
Deformable-Sprites [81]	25.6	39.5	71.4	1.70	20.6	32.9	69.7	2.07	45.0	58.3	84.0	0.99
Ours (TAP-Net)	53.8	68.3	88.8	0.77	50.9	66.7	85.7	0.86	73.4	84.1	92.2	0.11
Ours (RAFT)	55.1	69.6	89.6	0.76	51.7	67.5	85.3	0.74	77.5	87.0	93.5	0.13

定量比较。我们的方法在不同数据集上始终实现位置准确性、遮挡准确性和时间连贯性最好的效果。我们的方法可以良好地处理来自 RAFT 和 TAP-Net 的不同输入成对对应关系，并在这两种基准方法上持续改善。

定性比较。我们强调了我们在长时间遮挡情况下识别和跟踪的能力，同时为遮挡期间的点提供了合理的位置，并处理了大的相机运动视差。请参见补充视频

以获取动态比较。

5. Limitations

与许多动作估计方法一样，本文的方法在快速非刚性运动和细小结构方面存在困难。在这些场景中，成对对应方法可能无法提供足够可靠的对应关系，以便我们的方法计算准确的全局运动。

此外，由于底层优化问题的高度非凸性，我们观察到我们的优化过程对于某些困难视频的初始化非常敏感。这可能导致次优的局部最小值，例如，错误的表面排序或在规范空间中重复的对象，有时很难通过优化进行修正。

6. Conclusion

这篇文章提出了一种新的方法，用于估计整个视频的完整和全局一致的运动。这篇文章提出了一种名为 OmniMotion 的新视频运动表示，其中包括一个准三维规范体和逐帧的局部规范双射。OmniMotion 可以处理具有不同相机设置和场景动态的一般视频，并通过遮挡产生准确且平滑的运动。这个方法在定性和定量上都显著优于先前的最先进方法。

参考文献

[1] Wang Q, Chang Y Y, Cai R, et al. Tracking Everything Everywhere All at Once[J]. arXiv preprint arXiv:2306.05422, 2023.