E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)11-3634-17

论文引用格式:Zhang L D, Li W, Wei D Y, Ma C W, Li Z Y and Shen G. 2025. Key frame extraction method for video 3D reconstruction based on mutual check weighted optical flow. Journal of Image and Graphics, 30(11):3634-3650(张泷丹, 李雯, 魏东岩, 马朝伟, 李政祎, 申戈. 2025. 基于互校验加权光流的视频三维重建关键帧提取方法. 中国图象图形学报, 30(11):3634-3650)[DOI:10.11834/jig.250009]

基于互校验加权光流的视频三维重建关键帧提取方法

张泷丹[1,2],李雯[1*],魏东岩[1],马朝伟[1],李政祎[1,2],申戈[1]

1. 中国科学院空天信息创新研究院,北京　100094; 2. 中国科学院大学电子电气与通信工程学院,北京　100049

摘　要:**目的**　视觉匹配导航需要预先构建场景三维点云信息,相较于传统软件和专业仪器测图建模,基于消费级终端的视频流数据视觉建模具有成本低、数据更新方便和空间覆盖广等优势,但视频帧因数量庞大存在图像冗余,造成三维模型重建计算代价高、累计误差较大甚至重建失败的问题,因此本文提出一种基于互校验加权光流的三维重建关键帧提取方法。**方法**　首先,利用传感器陀螺仪数据对视频流中的图像进行场景预分类;然后,采用SIFT(scale invariant feature transform)算法检测图像特征点和描述符,并结合FLANN(fast library for approximate nearest neighbors)匹配和金字塔LK(Lucas-Kanade)光流算法,捕捉相邻帧的动态变化,提取两种算法同时检测成功的特征点并计算欧氏距离,筛选出相邻帧强匹配点对;最后,基于场景预分类结果,对图像消失点附近的强匹配点对,在直线道路采取高斯加权,在转弯道路采取均匀加权,计算帧间光流场总运动从而获取相似度,最终实现视频关键帧提取。**结果**　实验利用消费级终端自采集4组不同场景数据,将本文算法与传统关键帧提取算法进行对比,统计提取关键帧数量并利用结构相似性指数计算高相似度帧数量,将直线和转弯道路提取结果与原视频帧分别进行对比,最后进行三维模型重建实验从而评估提取效果。实验结果表明,本文算法可以将视频帧总数量降低到10%左右,其中高相似度帧数量明显少于其他算法;相较于直线道路,在转弯处关键帧数量占比较大,符合三维重建预期需求;最终模型重建完整度在4组数据上分别为100%、100%、97.46%和96.54%,优于其他算法。**结论**　本文提出基于互校验加权光流的三维重建关键帧提取方法能有效降低视频帧数量,筛选的关键帧能够提高相邻帧匹配精度和稳定性,增强在多样化场景下三维重建的鲁棒性。

关键词:视频流;关键帧;图像相似度;互校验加权光流;三维重建

Key frame extraction method for video 3D reconstruction based on mutual check weighted optical flow

Zhang Longdan[1,2], Li Wen[1*], Wei Dongyan[1], Ma Chaowei[1], Li Zhengyi[1,2], Shen Ge[1]

1. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing　100094, China;

2. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing　100049, China

Abstract: **Objective** Nowadays, navigation and positioning technology has become an indispensable part of people's daily life, and the satellite positioning system has been successfully built and widely used. However, in the environment where buildings are dense or satellites are blocked indoors, satellite positioning is inaccurate due to signal interference. There-

收稿日期:2025-01-08;**修回日期**:2025-03-06;**预印本日期**:2025-03-13

*通信作者:李雯　wen.li@aircas.ac.cn

基金项目:国家自然科学基金项目(42204048);中国科学院航天信息研究所科学与颠覆性技术研究基金项目(2024-AIRCAS-SDTP-09)

Supported by: National Natural Science Foundation of China(42204048); Science and Disruptive Technology Research Fund Program of Aerospace Information Research Institute, Chinese Academy of Sciences(2024-AIRCAS-SDTP-09)

fore, visual navigation and positioning technology has been developed to overcome this difficulty. This technology determines the position and attitude of the camera in 3D space through image processing and computer vision technology. It is an important means to solve the navigation and positioning problem in the satellite navigation rejection scene. Usually, visual matching navigation requires the pre-construction of 3D point cloud information of the scene. The acquisition of 3D point cloud models can be mainly divided into three types: manual model construction by mathematical modeling software, mapping by professional instruments, and crowdsourcing mapping by consumer terminals. These models are time consuming and laborious in constructing large-scale feature point cloud databases. Meanwhile, visual modeling of video stream data based on consumer terminals has advantages such as low cost, convenient data update, and wide spatial coverage. However, due to the large number of video frames and image redundancy, the 3D model reconstruction calculation cost is high, the cumulative error is large, and even the reconstruction failure is caused. Thus, this study proposes a 3D reconstruction key frame extraction method based on mutual check weighted optical flow. **Method** First, the image scene is pre-classified. In the process of video shooting, the camera passes through multiple scenes, and the video frame changes in different scenes. The pixel changes of the straight road are mainly distributed at the edge of the image. At the same time, the pixels of the entire image have changed at the turning point, such that the video needs to be pre-classified. The system receives the self-collected video stream data. Then, it combines the gyroscope data obtained by the mobile terminal to divide the scene into two types: straight road and turning road, which provides a basis for subsequent targeted optical flow aggregation adjustment and adjacent frame similarity calculation. Then, cross-check adjacent frame matching is conducted, followed by the use of the Scale Invariant Feature Transform algorithm to detect feature points and their descriptors in the previous frame image. The matching points in the second frame are calculated by fast library for approximate nearest neighbors matching and pyramid Lucas-Kanade (LK) optical flow method, and the feature points successfully matched by both methods are detected. The incorrect matching points are eliminated by calculating the 2D Euclidian distance. Strong matching point pairs are obtained to capture the dynamic changes between frames, which ensures the accuracy and effectiveness of the subsequent optical flow calculation. Finally, the total optical flow field is aggregated, and the similarity of adjacent frames is calculated. Considering the differences in the video frames of the straight road and the turning road, the contribution of matching feature points of adjacent frames also varies in the similarity calculation. Therefore, optical flow aggregation needs to be conducted by weighting. After detecting the vanishing point of the image and taking it as the center, different weights are assigned to the strong matching point pairs near the vanishing point according to the scene classification. Next, the aggregate information of the total optical flow field of adjacent frames is weighted to obtain the optical flow changes between frames. Then, the images with significant motion changes are judged as key frames according to the set threshold, and the key frames of video stream are finally extracted. **Result** In the experiment, the self-developed video acquisition app of the research group in the consumer terminal is used to conduct data self-acquisition of the target environment. In different scenes with different lighting and ground feature distribution, four groups of data are obtained using different traveling routes and speeds, and each group records the scene video information and gyroscope data synchronously. The proposed algorithm is compared with the traditional key frame extraction algorithm, and different algorithms are used to screen video key frames and count the number of extracted key frames. The structural similarity index is used to calculate the number of highly similar frames in key frames. High-similarity frames are the images with high visual similarity between the two frames in the extracted video key frames. A greater number of high-similarity frames correspond to higher redundancy in the extracted key frames. Then, the results of key frame extraction in the straight line and turning road are compared with the those for the original video frame, respectively. The proportion of the number of key frames extracted in the original video frame is calculated in different scenes to evaluate the scene adaptability of the algorithm. Finally, the key frames extracted by the algorithm are used for 3D model reconstruction experiment, and the road map is drawn with global navigation satellite system (GNSS) data to evaluate the integrity of the reconstructed model. Experimental results show that the proposed algorithm can reduce the total number of video frames to approximately 10%, and the minimum can reach 4.56%. Meanwhile, the proportion of high-similarity frames in key frames is less than 3%, and the minimum is 1.91%, which is significantly less than in other algorithms. In addition, the number of key frames extracted by the algorithm in this study is much larger than that in the straight road. This result is better than those of other algorithms and meets the

expected demand for the number of images in different scenes under the application of 3D reconstruction. The integrity percentages of the final model reconstruction are 100%, 100%, 97.46%, and 96.54% on the four groups of data. Therefore, it is obviously better than other algorithms. **Conclusion** This study proposes a key frame extraction method for 3D reconstruction based on mutual check weighted optical flow. This method can effectively reduce the number of video frames and improve the quality of key frame screening in diversified scenes. At the same time, the extracted key frames can improve the matching accuracy and stability of adjacent frames and enhance the robustness of 3D reconstruction.

Key words: video streaming; key frame; image similarity; mutual check weighted optical flow; three-dimensional reconstruction

0 引 言

随着科学技术的发展,导航定位技术已经渗透到人们的日常生活,准确位置信息对驾车和步行至关重要。传统位置获取方法依赖全球卫星导航系统(global navigation satellite system, GNSS),但在信号干扰较大的环境中面临困难,因此,视觉导航定位技术已成为当下的研究热点。视觉匹配定位(张铭磊,2020)能够通过摄像头采集的图像信息进行识别和分析,解决卫星导航失效问题,提供更精准可靠的定位,从而实现路线规划和实时导航。

高精度的视觉匹配定位依赖预先建立的视觉特征三维点云数据库,通过将当前图像与模型进行匹配来实现用户定位(张霄,2022)。获取三维点云模型的方法主要分为利用数学软件建模、专业仪器测图和消费级终端众包测图。其中利用软件手动建模操作复杂,工作量大;利用专业仪器测图成本高且设备体积大,一般情况下不适用,以上两种方法在构建大规模特征点云数据库的采集制作时费时费力。而利用消费级终端的众包测图获取三维点云具有成本低、采集方便、数据更新容易和空间覆盖广的优势,其中行车记录仪、手机增强现实(augmented reality, AR)导航等用户端在使用过程中可采集道路沿途视频信息,基于这类视频数据可众包构建视觉特征点云数据库,支撑广域的视觉导航定位。然而视频中存在大量冗余图像(Paul等,2018),从而造成三维模型重建计算代价高、累计误差较大甚至重建失败等问题,因此需要对视频序列进行筛选,提取出满足三维重建需求的关键帧,减小三维重建计算量并提高其效率和鲁棒性(方子赟,2022)。

目前获取视频关键帧的方法主要可以分为基于时间间隔提取、基于帧间差异提取和基于光流提取(Deshpande等,2018)。

基于时间间隔提取的方法不依赖于具体的图像内容或特征,而是利用时间间隔信息均匀选取帧。Zarco-Tejada等人(2014)提出一种基于区间分析的时间间隔提取算法,对无人机关键帧图像进行筛选并用于后续重建。Yang等人(2015)采用等时间间隔检测方法,对图像帧按时序分组并进行空间检索,保证关键帧在空间上满足重叠率需求。张航等人(2020)结合拍摄相机参数,基于时间间隔加权和阈值约束,获取视频帧的动态时差并对关键帧进行筛选。

基于帧间差异提取的方法通过比较相邻帧相似度来选择关键帧。Zhao等人(2019)将HSV(hue saturation value)直方图作为每个帧的颜色特征以减少数据量,利用轮廓系数进行图像聚类提取关键帧。Yan和Woźniak(2022)根据RGB空间进行视频色彩分类,利用非均匀化HSV空间和特征向量多层核心聚合算法进行视频分割和关键帧提取。Pandian和Maheswari(2024)基于马尔可夫链聚类对视频帧进行分析分组,然后利用相邻矩阵聚类算法提取图像运动信息,从而筛选关键帧。

基于光流提取的方法通过求解像素点在相邻帧矢量变化情况,分析视频运动的向量场大小和方向提取关键帧(Dong, 2023)。Bao等人(2020)提出一种光流—互信息熵相结合的方法,提取出邻域具有光流差极值的视频帧,并以最小互信息熵作为阈值筛选关键帧。Yuan等人(2022)将视频进行预先划分降低冗余,然后计算图像全局光流变化筛选出候选关键帧,最后根据时空一致性和分层聚类的方法提取关键帧。Li(2024)利用光流法进行关键帧滤波,并基于自适应K均值聚类算法提取图像纹理特征,通过计算欧氏距离进行优化完成关键帧筛选。

然而,应用于三维重建的关键帧选择不仅需要

考虑图像质量和信息丰富度，还需要结合场景动态变化，以保证重建的准确性和稳定性（郑义桀 等，2023），不论基于上述哪一种方法对关键帧进行提取，均面临一系列挑战，主要归纳为以下 3 个方面：1）视频中的图像序列场景特征各异，相邻帧相似度计算存在不准确性（Yan，2023）。在视频采集过程中，场景变化呈动态特征，沿道路直行时，相机光轴平行于道路，视频中场景变化主要分布于图像边缘位置，图像视觉中心附近的像素变化较小；转弯时，相机光轴与道路朝向的关系发生改变，表现于视频中的相邻帧画面整体像素存在移动。在以上两种情况中，相邻帧图像变化特点存在明显差异，可能导致图像相似度估计时出现偏差，影响关键帧的提取效果。2）相邻帧相似度差异计算仍存在缺陷。基于时间间隔的提取方法在不同运动速度下可能导致图像的缺失或冗余；基于直方图、特征聚合场等方式进行帧间差异判断提取时，无法捕捉图像的局部细节，难以描述像素偏移量、偏移方向等精确运动差异；基于光流提取方法可以描述像素运动差异，但是针对相机运动场景的光流计算鲁棒性和准确性不高。3）在筛选视频关键帧时，图像像素点在相邻帧相似度计算中的贡献值根据场景不同而存在差异。在运动变化较大的场景中，相邻帧之间某些区域的像素点差异显著增加，从而导致其具有较高贡献值；而在相对静态或简单的场景中，画面变化较小，许多像素的贡献值相对较低。采用单一光流计算方法难以区分场景变化差异，因此在多样化运动场景中，需要考虑在不同场景下像素点光流对于整幅图像光流聚合的贡献差异，针对局部像素点进行适应性光流聚合。

为解决上述问题，本文提出一种基于互校验加权光流的视频三维重建关键帧提取方法，能够将视频场景预分类，有效进行场景划分；基于尺度不变特征变换（scale invariant feature transform，SIFT）特征点检测，结合快速近似最近邻搜索库（fast library for approximate nearest neighbors，FLANN）匹配和金字塔光流计算，采用互校验检测机制进行相邻帧强匹配点对筛选；考虑图像消失点及其周围点的贡献差异进行加权聚合光流计算，从而构建稳健的相邻帧相似度判别方法，实现视频流关键帧筛选，最终得到一组符合三维重建需求的图像集合。从实验结果可以看出，与其他算法相比，本文算法可以有效降低视频帧的数量，提高关键帧筛选时的匹配精度和稳定性，最终提取一组适用于三维重建的图像序列。

1　系统整体框架

在视觉场景重建领域，关键帧筛选技术对提取目标地物信息和提高重建效率有着重要影响。本文研究了一种基于视频流数据的关键帧提取方法，如图 1 所示，主要步骤包括图像场景预分类、互校验相邻帧匹配以及相邻帧相似度判别 3 个过程。首先，对图像场景进行预分类处理。系统接收自采集视频流信息，利用移动终端获取的陀螺仪数据，将场景分为直线道路和转弯道路两种类型。由于不同场景中的相机光轴与道路方向存在差异，相邻帧像素点变化不同，此步预先完成场景划分，为后续针对性进行光流聚合和帧间相似度计算提供基础。其次，进行互校验相邻帧匹配。对上一帧图像使用 SIFT 算法进行特征点检测，并得到相应的描述符；然后分别利用 FLANN 和金字塔 LK（Lucas-Kanade）光流算法计算在第 2 帧中的匹配点；接着检测两种方法同时匹配成功的特征点，通过计算二维欧氏距离剔除错误匹配点，获取强匹配点对捕捉帧间动态变化，从而确保后续光流计算的准确性和有效性。最后，进行总光流场聚合并计算相邻帧相似度。检测图像消失点并以其为中心，根据场景分类情况，对消失点附近的强匹配点对赋予不同的权值，加权计算相邻帧总光流场聚合信息，得到帧间光流变化情况，然后根据设置的阈值将运动变化显著的图像判断为关键帧，最终完成视频流关键帧提取。

2　自适应关键帧提取方法研究

2.1　视频场景预分类

在视频拍摄过程中，摄像头经过多个场景，不同场景下的视频帧变化各异（Mushan 和 Vidap，2020）。本文重点关注转弯和直线道路的图像变化，如图 2 所示，在直线道路上图像边缘像素发生变化，在转弯时整幅画面都在改变。为了有效地获取相邻帧相似度，需要对这些差异分类计算，因此首先进行场景预分类。此步根据消费级终端采集到的视频流，利用陀螺仪数据计算进行有效的图像场景分类。

在用户转动设备的过程中，陀螺仪能够记录手

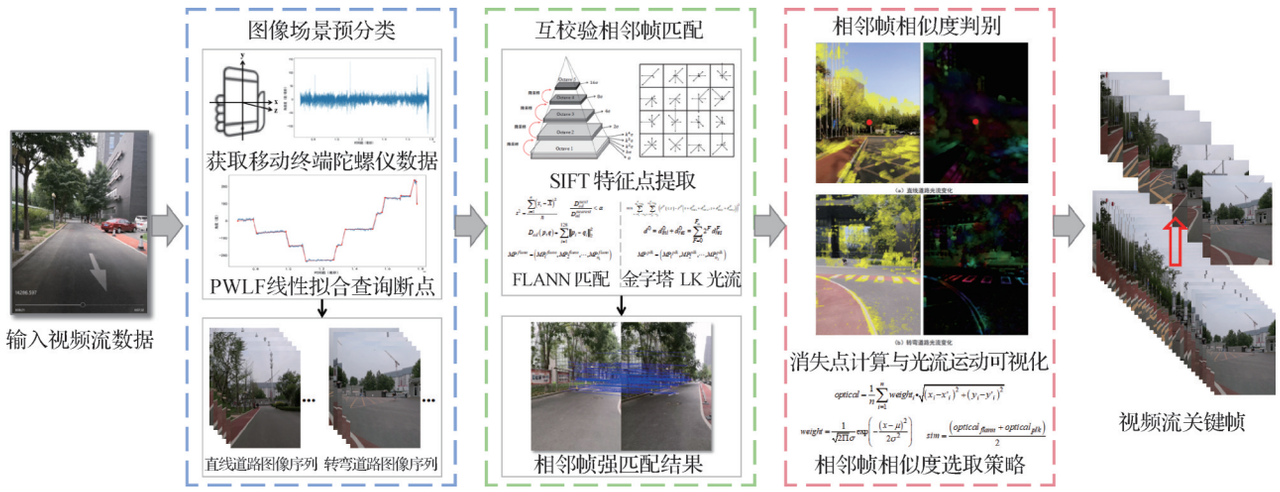
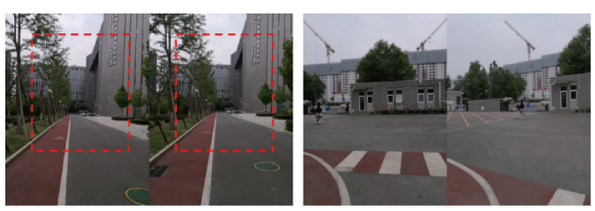


图1　关键帧提取系统整体框架

Fig. 1　Key frame extraction system overall framework



(a) 直线道路相邻帧对比图　(b) 转弯道路相邻帧对比图

图2　不同场景相邻帧变化示意图

Fig. 2　Change diagram of adjacent frames in different scenes ((a) comparison of adjacent frames on straight road; (b) comparison of adjacent frames on turning road)

持设备在各个坐标轴上的角速度数据。手持设备陀螺仪坐标轴方向如图3所示，当转弯时，手持设备主要绕y轴转动，陀螺仪在此轴向上实时记录的变化情况如图4(a)所示。

结合y方向角速度和系统时间戳数据，进行积分运算获取当前时间戳下的方位角，具体为

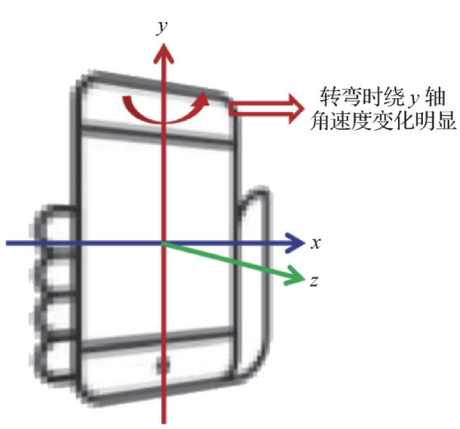


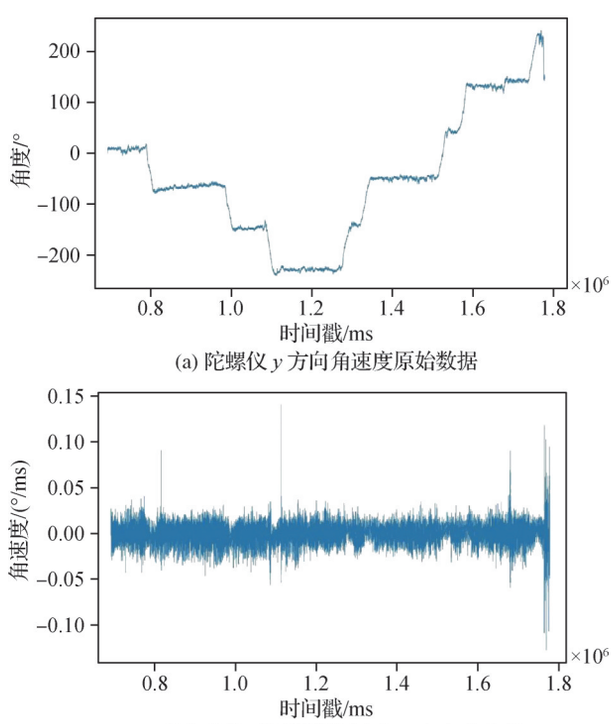
图3　传感器坐标系统

Fig. 3　Sensor coordinate system

$$angle_y = \sum gyro_y \times \left(time_{\text{now}} - time_{\text{last}}\right) \tag{1}$$

式中，$angle_y$为y方向在当前时间戳的方位角，$gyro_y$为传感器记录的y方向角速度，$time_{\text{now}}$和$time_{\text{last}}$分别为当前和前一时刻的系统时间戳。令初始方位角为0，通过式(1)计算得到不同时间戳下各视频帧的具体方位角如图4(b)所示。

然后，利用分段线性拟合(piecewise linear fit,



(a) 陀螺仪y方向角速度原始数据

(b) 积分后各时间戳视频帧在y方向方位角

图4　传感器原始数据与积分数据

Fig. 4　Sensor raw data and integral data ((a) raw data of gyro y angular velocity; (b) azimuth angle of each time-stamped video frame in y direction after integration)

PWLF）算法（Jekel 和 Venter，2019）对 y 方向方位角进行分析，将数据划分成多个区间，每个区间使用一条线性函数进行拟合，通过迭代搜索各区间起始点和终止点，在拟合误差最小时得到最优分段，从而根据划分区间对图像直线和转弯场景进行分类。

假设在 y 方向的 n 组方位角数据最终拟合成一组具有 n_b-1 个区间的分段线性函数，每个区间线性函数的斜率和截距都依赖于之前的函数值，将拟合的分段线性函数表示为

$$y(x)=\begin{cases}\beta_1+\beta_2(x-b_1) & b_1\leqslant x\leqslant b_2\\ \beta_1+\beta_2(x-b_1)+\beta_3(x-b_2) & b_2<x\leqslant b_3\\ \quad\vdots & \quad\vdots\\ \beta_1+\beta_2(x-b_1)+\beta_3(x-b_2)+\cdots+\\ \beta_{n_b}(x-b_{n_b-1}) & b_{n_b-1}<x\leqslant b_{n_b}\end{cases}\tag{2}$$

式中，$b_1,b_2,\cdots,b_{n_b}$ 依次为 n_b 个断点的位置。式（2）可表示为矩阵形式，具体为

$$\begin{bmatrix}1 & x_1-b_1 & (x_1-b_2)I_{x_1>b_2} & \cdots & (x_1-b_{n_b-1})I_{x_1>b_{n_b-1}}\\ 1 & x_2-b_1 & (x_2-b_2)I_{x_2>b_2} & \cdots & (x_2-b_{n_b-1})I_{x_2>b_{n_b-1}}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_n-b_1 & (x_n-b_2)I_{x_n>b_2} & \cdots & (x_n-b_{n_b-1})I_{x_n>b_{n_b-1}}\end{bmatrix}\times\begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_{n_b}\end{bmatrix}=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}\tag{3}$$

式中，$x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 分别为 n 组系统时间戳和 y 方向对应方位角数值。$I_{x_i>b_j}$ 为指示函数，$x_i\leqslant b_j$ 值为 0，$x_i>b_j$ 值为 1。

式（3）中的矩阵可以简化表示成 $\boldsymbol{A\beta}=\boldsymbol{y}$，$\boldsymbol{A}$ 为 $n\times n_b$ 回归参数矩阵，$\boldsymbol{\beta}$ 为 $n_b\times 1$ 向量，$\boldsymbol{y}$ 为 $n\times 1$ 方位角向量。通过最小二乘最小化残差平方和进行估计求解 $\boldsymbol{\beta}$，表示为 $\boldsymbol{\beta}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$。计算拟合的连续分段线性模型与原始数据的差值，得到 $(n\times 1)$ 残差向量 $\boldsymbol{e}=\boldsymbol{A\beta}-\boldsymbol{y}$，进而残差平方和可以表示为 $SSR=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$。

假设第一个和最后一个断点分别为起始和终止时间戳，需要求解剩下 n_b-2 个未知断点位置。采取双循环迭代优化模型，分为各个区间拟合线性函数的内部优化和整体分段函数的外部优化。

在内部优化中通过最小二乘估计得到各区间最佳拟合折线，得到 $\boldsymbol{A}$ 和 $\boldsymbol{\beta}$ 的值与当前模型的残差平方和；在外部优化中采取差分进化算法（differential evolution，DE）进行全局优化，通过迭代搜索断点位置来更新区间划分，根据不同划分结果的 SSR 决定是否保留当前模型，从而寻找全局最优解，得到最终划分结果的断点位置。

PWLF 算法需要提前设置断点数量，通过对角速度滤波处理，然后设置阈值计算峰值来预先获取断点个数。首先对原始角速度进行加权移动平均滤波（weighted moving average filtering，WMAF），设置高斯权重 σ_w 和滑动窗口大小，每个点的滤波值为窗口内附近数据点的加权平均值，其中第 i 个数据点的权重为

$$w(i)=\exp\left(-\frac{(i-c)^2}{2\sigma_w^2}\right)\tag{4}$$

式中，c 是当前待计算点，距离其越近的数据点权重越高。然后对于滤波后的角速度数据，设置搜索的最大距离和极值阈值，获取局部最大值作为波峰并统计数量，如图 5 所示，蓝线为原始数据，绿线为滤波后数据，红点为波峰计算结果，在该数据下共计 10 个波峰点，由于断点计算考虑线段两端且默认数据起始与终点处为断点，同时从图中可以看出，计算结果中存在一个波峰位于数据末尾处，因此预先设置断点个数为 $10\times 2+2-1=21$ 个。

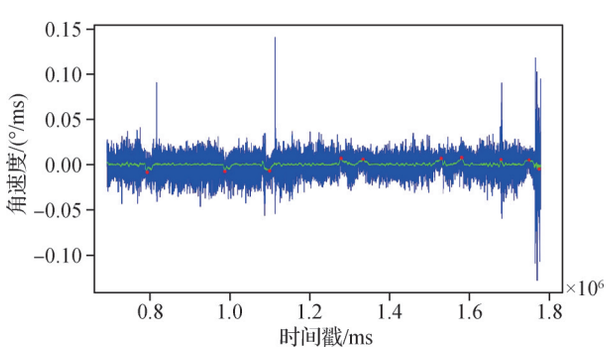


图 5　波峰计算示意图

Fig. 5　Wave crest calculation diagram

将预先设置的断点个数输入 PWLF 算法进行分段线性拟合，拟合结果如图 6 所示，其中红点表示断点位置，即图像场景划分的分割点。最终得到 n_b-1 个区间的分段线性函数，每个区间拟合线性函数的斜率为 a_i，由于相邻帧方位角在转弯道路上变化剧烈，因此 a_i 绝对值较大时则位于转弯位置，在本文

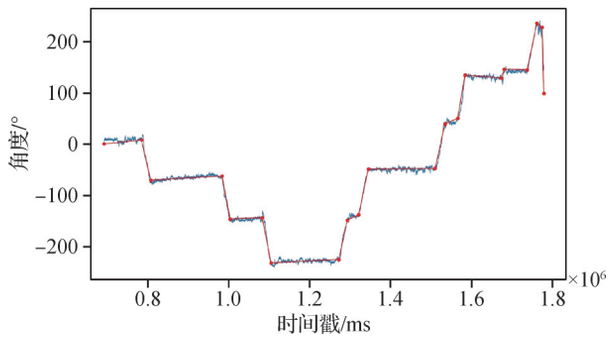


图6　PWLF线性拟合结果

Fig. 6　PWLF linear fitting results

中将斜率阈值设置为3×10^{-4}，对图像场景进行划分：若$|a_i|$小于阈值则认为处于直线道路上，否则处于转弯道路。

值得注意的是，在方位角拟合后的分段线性图中，斜率阈值的选取直接影响了场景预分类的结果，由于视频采集设备的陀螺仪在灵敏度和采样率等方面可能存在差异，因此需要针对性地调整阈值设置。对于不同设备采集的数据拟合结果，通过绘制ROC（receiver operating characteristic curve）评估模型的性能，从而查找最佳阈值，确保本文方法在不同设备上的泛化能力。

2.2　基于互校验加权匹配的帧间光流计算

在本文提出的视频关键帧筛选算法中，提取相邻帧稳定可靠的匹配点对是核心环节。如图7所示，本文使用SIFT检测算法计算图像特征点和描述符，基于FLANN匹配和金字塔LK光流计算获取相邻帧特征点变化，结合两种方法同时捕捉成功的特征点位置差异，筛选得到强匹配点对。

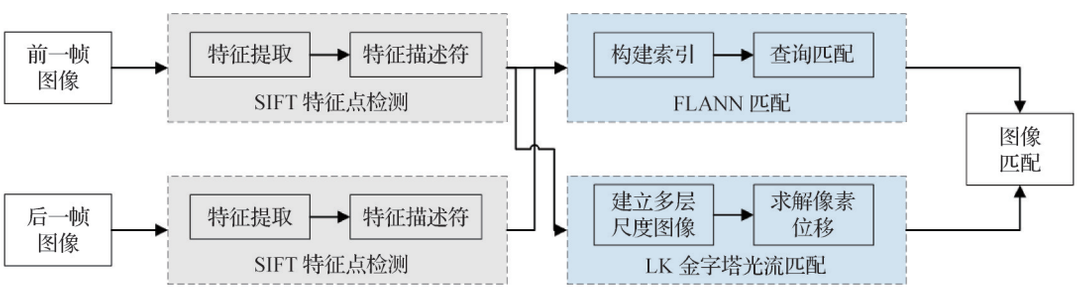


图7　相邻帧互校验匹配流程

Fig. 7　Matching process of adjacent frame mutual check

2.2.1　SIFT特征点检测

尺度不变特征变换（SIFT）对图像局部特征进行检测并生成128维描述符，此方法在尺度不变性和旋转不变性等方面具有较高的鲁棒性（Lowe，2004）。SIFT特征点检测算法主要包括尺度空间极值检测、特征点精确定位、特征点主方向确定和特征点描述符构建4个步骤。

1）尺度空间极值检测。首先通过高斯核函数$G(x, y, \sigma)$对原始图像$\boldsymbol{I}(x, y)$进行卷积运算，构建高斯尺度空间，得到图像的函数表达式，具体为

$$\boldsymbol{L}(x, y, \sigma) = G(x, y, \sigma) \times \boldsymbol{I}(x, y) \tag{5}$$

式中，σ为尺度因子，值越大则图像越模糊，$G(x, y, \sigma)$计算为

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{x^2 + y^2}{2\sigma^2}} \tag{6}$$

SIFT算法将高斯差分金字塔作为图像尺度空间。从最底层图像（原始图像）开始计算高斯金字塔，然后对每组倒数第3幅图像进行降采样处理，得到新的一组底层图像，重复上面步骤直至高斯金字塔构建完成，然后对相邻两幅图像作差构建高斯差分金字塔。最终图像表达式为

$$\begin{aligned}\boldsymbol{D}(x, y, \sigma) &= \big(G(x, y, k\sigma) - G(x, y, \sigma)\big) \times \boldsymbol{I}(x, y) = \\ &\boldsymbol{L}(x, y, k\sigma) - \boldsymbol{L}(x, y, \sigma)\end{aligned} \tag{7}$$

式中，k为常数。

2）特征点精确定位。将高斯差分金字塔中图像的每个像素与其相邻上下两层对应点位附近的9×2个点，以及同层内的8个点进行比较，若存在极值则认为是离散空间中的潜在特征点。接着利用三维二次函数进行插值曲线拟合以找到特征点精确位置，将$\boldsymbol{D}(x, y, \sigma)$在潜在特征点$\boldsymbol{X}_0 = \begin{bmatrix} x_0 & y_0 & \sigma_0 \end{bmatrix}^{\mathrm{T}}$位置处进行泰勒展开，具体为

$$\boldsymbol{D}(\Delta\boldsymbol{X}) = \boldsymbol{D}(\boldsymbol{X}_0) + \frac{\partial \boldsymbol{D}^{\mathrm{T}}}{\partial \boldsymbol{X}_0}\Delta\boldsymbol{X} + \frac{1}{2}\Delta\boldsymbol{X}^{\mathrm{T}}\frac{\partial^2 \boldsymbol{D}}{\partial \boldsymbol{X}_0{}^2}\Delta\boldsymbol{X} \tag{8}$$

当式（8）一阶导数为零时，可以求得位置修正值，具体为

$$\Delta \boldsymbol{X} = -\frac{\partial^2 \boldsymbol{D}^{-1}}{\partial \boldsymbol{X}_0^{\ 2}} \frac{\partial \boldsymbol{D}}{\partial \boldsymbol{X}_0} \tag{9}$$

将$\Delta \boldsymbol{X}$代入泰勒展开式中，设置阈值D_T，若满足$\left|\boldsymbol{D}(\Delta \boldsymbol{X})\right| \geqslant D_T$则保留为最终特征点。

3）特征点主方向确定。若上述高斯图像的尺度为σ，将特征点所在的位置定位圆心，对特征点附近一周内的像素点进行梯度信息统计，根据经验选取半径为1.5σ，梯度信息包含梯度模值$M(x,y)$和梯度方向$\theta(x,y)$，具体为

$$M(x,y) = \sqrt{\left(L(x+1,y)-L(x-1,y)\right)^2 + \left(L(x,y+1)-L(x,y-1)\right)^2}$$

$$\theta(x,y) = \arctan\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)} \tag{10}$$

4）特征点描述符构建。将图像转至主方向θ上，并将特征点邻域划分为4×4个子区域，分别计算邻域点在8个主要方向梯度模值的累加和，将SIFT描述符表示为$4\times4\times8=128$维的特征向量，最后进行归一化处理，可得

$$l_i = \frac{h_i}{\sqrt{\sum_{j=1}^{128} h_j}},\ i = 1,2,3,\cdots,128 \tag{11}$$

式中，h_i为前面计算得到的SIFT描述符128维特征向量，l_i为归一化后的特征向量，从而得到特征点描述符。

2.2.2　FLANN匹配

在SIFT特征点描述符提取完成后，进行图像匹配，对于大规模数据集和高维度特征，暴力匹配计算成本极高，K最近邻（K-nearest neighbors，KNN）匹配会遭遇“维度灾难”问题，本文采用快速近似最近邻搜索库（FLANN）匹配进行计算，此方法能够在高维中提高匹配效率。

首先采用KD树（KD-tree）算法递归划分空间以加速计算最近邻匹配搜索。对于图像中128维特征描述符，分别计算每个维度的方差，具体为

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}{n} \tag{12}$$

式中，n为该维度下数据个数，$\bar{X}$为数据平均数。根据具有最大方差的对应维度进行子空间划分，重复操作直至KD树构建结束。接着利用已构建好的，依次选择KD树从上至下每个节点处的维度类型，加速待匹配点定位到子空间的过程，从所在子空间出发，依次向外计算与其他特征点的欧氏距离，具体为

$$D_{ed}(p,q) = \sum_{i=1}^{128}\left\| p_i - q_i \right\|_2^2 \tag{13}$$

式中，$p(p_1,p_2,\cdots,p_{128})$，$q(q_1,q_2,\cdots,q_{128})$分别为两个特征点描述符。最终得到待匹配点的最近邻距离D_{ed}^{nearest}和次近邻距离D_{ed}^{next}，设置阈值$\alpha=0.6$，若满足下式，则认为最近邻距离对应的特征点为匹配成功的特征点，最终得到相邻帧图像匹配点对，具体为

$$\frac{D_{ed}^{\text{nearest}}}{D_{ed}^{\text{next}}} < \alpha \tag{14}$$

由于存在阈值剔除，对于前一帧图像中的m个特征点：$point_i(i=1,2,\cdots,m)$，并非每个点都能在后一帧中成功进行匹配，假设在后一帧中成功匹配上$n_1(n_1<m)$个点，将FLANN算法获取的匹配点对表示为$MP_i^{\text{flann}}(i=1,2,\cdots,n_1)$。

2.2.3　金字塔LK匹配

传统LK算法假设图像中像素位移较小，当相邻图像中的像素发生较大变化时，LK容易产生错误估计。金字塔LK（pyramid Lucas-Kanade）算法通过计算图像金字塔，根据不同的采样率构建多幅图像，在分辨率上从低到高逐步估计光流，能够较好地处理场景变化过快问题，光流估计更鲁棒。金字塔LK计算步骤如下：

1）建立图像金字塔。图像金字塔将原始图像$\boldsymbol{I}^0$作为金字塔的最底层（第0层），将图像进行降采样得到第1层图像$\boldsymbol{I}^1$，其宽度和高度为$\boldsymbol{I}^0$的一半，依次向上递归计算构建F^m层图像金字塔，本文构建的金字塔取$F^m=4$。

2）计算每层图像光流。对于金字塔内每层图像，使用LK光流法进行求解。假设在t时刻，每层图像$\boldsymbol{I}^F$中位于(x,y)的像素点灰度值为$\boldsymbol{I}^F(x,y,t)$，假设经过Δt时间后灰度值不变，即$\boldsymbol{I}^F(x,y,t)=\boldsymbol{I}^F(x+\Delta x,y+\Delta y,t+\Delta t)$，进行泰勒展开并令高阶项为0，可得

$$\frac{\partial \boldsymbol{I}^z}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial \boldsymbol{I}^F}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial \boldsymbol{I}^F}{\partial t} = 0 \rightarrow \boldsymbol{I}_x^F u^F + \boldsymbol{I}_y^F v^F + \boldsymbol{I}_t^F = 0 \tag{15}$$

假设在大小为(w_x,w_y)的邻域窗口内，所有像素

点光流运动一致，利用最小二乘法迭代求解第F层光流运动$d^F=[u^F \quad v^F]$，其中设像素点(x,y)在后一帧中的灰度值为$\boldsymbol{J}^F(x,y)$，将$\boldsymbol{I}^F$和$\boldsymbol{J}^F$在邻域范围内每个点的匹配误差和作为待最小化的损失函数，具体为

$$\varepsilon^F(d^F)=\sum_{x=u_x^F-w_x}^{u_x^F+w_x}\sum_{y=u_y^F-w_y}^{u_y^F+w_y}\left(\boldsymbol{I}^F(x,y)-\boldsymbol{J}^F(x+u^F,y+v^F)\right)^2 \tag{16}$$

3）迭代更新光流估计。图像金字塔在构建时，图像尺寸每次都缩放了一半，共缩放F^m层，从最上层光流估计开始，依次向下进行反馈，直至金字塔最底层（原始图像），设第F层图像的光流准确值为

$$d^F=d_{\text{ini}}^F+d_{\text{res}}^F \tag{17}$$

式中，d_{ini}^F为光流初始值，d_{res}^F为光流残差。将式（17）代入损失函数，则最小化损失函数求解光流残差表达为

$$\min\sum_{x=u_x^F-w_x}^{u_x^F+w_x}\sum_{y=u_y^F-w_y}^{u_y^F+w_y}(\boldsymbol{I}^F(x,y)-\boldsymbol{J}^F(x+d_{\text{ini}\,x}^F+d_{\text{res}\,x}^F,y+d_{\text{ini}\,y}^F+d_{\text{res}\,y}^F))^2 \tag{18}$$

设最顶层F^m图像光流初始值为$d_{\text{ini}}^{F_m}=[0 \quad 0]^{\text{T}}$，计算损失函数得到残差值$d_{\text{res}}^{F_m}$，将准确值$d^{F_m}$传递到下一层作为初始值：$d_{\text{ini}}^{F_m-1}=2d^{F_m}$，然后重复上述操作直至最底层：$d^0=d_{\text{ini}}^0+d_{\text{res}}^0=2(d_{\text{ini}}^1+d_{\text{res}}^1)+d_{\text{res}}^0=\cdots$，则最底层原始图像的最终光流为

$$d^0=d_{\text{ini}}^0+d_{\text{res}}^0=\sum_{F=0}^{F_m}2^F d_{\text{res}}^F \tag{19}$$

在相邻两帧图像中进行金字塔LK计算，对于前一帧图像中的m个特征点$point_i(i=1,2,\cdots,m)$，都能在后一帧中成功解算对应匹配点，假设在后一帧中匹配上$n_2(n_2=m)$个点，将金字塔LK算法获取的匹配点对表示为$MP_j^{\text{plk}}(j=1,2,\cdots,n_2)$。

2.2.4 互校验加权匹配

为了获取相邻帧更准确的特征匹配，本文采用基于FLANN匹配和金字塔LK光流互校验匹配的方法，通过对两种方法匹配结果进行双向验证，能够有效地过滤掉错误匹配，减小误匹配的影响。

在相邻两帧图像中，假设前一帧图像中有m个待匹配特征点，在后一帧图像中，利用FLANN匹配算法搜索到$n_1(n_1<m)$个匹配点，记为

$$MP^{\text{flann}}=\left(MP_1^{\text{flann}},MP_2^{\text{flann}},\cdots,MP_{n_1}^{\text{flann}}\right) \tag{20}$$

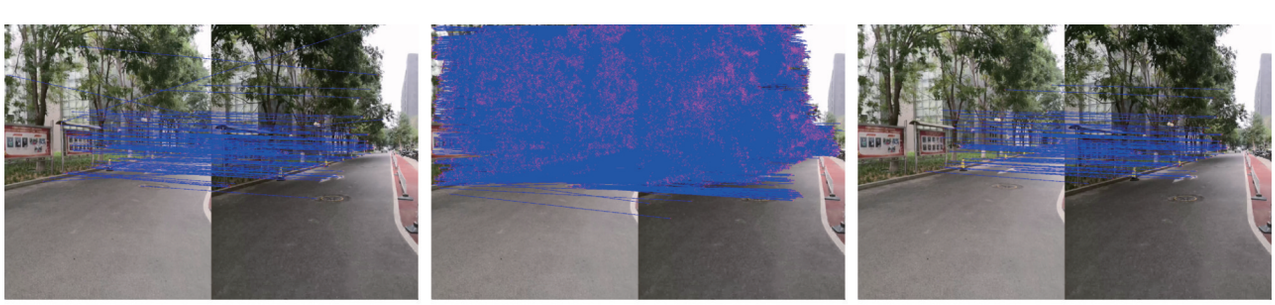
利用金字塔LK光流算法搜索到$n_2(n_2=m)$个特征点，记为

$$MP^{\text{plk}}=\left(MP_1^{\text{plk}},MP_2^{\text{plk}},\cdots,MP_{n_2}^{\text{plk}}\right) \tag{21}$$

遍历第1幅图像点，搜索两种方式都成功匹配上的点MP_i^{flann}和MP_j^{plk}，再计算欧氏距离进行误差量化，设置阈值筛选强匹配点对。从自采集视频中提取出两帧图像，采用SIFT + FLANN算法、金字塔LK光流算法匹配的效果如图8（a）（b）所示，设置阈值为150，互校验后的强匹配点检测算法效果如图8（c）所示。分别对SIFT + FLANN匹配结果、金字塔LK光流匹配结果、互校验匹配结果进行定量对比，如表1所示，可以看出，本文算法能够较好提升匹配正确率。

2.3 相邻帧相似度判别方法

根据前文所述，由于视频帧在直线道路和转弯道路上的变化存在差异，因此帧间匹配特征点对相似度计算的贡献也有所不同。针对这一问题，本文提出一种相邻帧相似度判别方法，如图9所示，完成对场景的预分类后，首先计算图像消失点，在直线道路中，距离图像消失点越远，画面变化越小，因此以



（a）SIFT + FLANN 匹配结果　（b）金字塔 LK 光流匹配结果　（c）强匹配点检测结果

图8　不同算法的匹配结果

Fig. 8　Matching results of different algorithms

（（a）SIFT + FLANN matching result；（b）pyramid LK optical flow matching result；（c）strong match detection result）

表 1　不同算法相邻帧匹配结果对比

Table 1　Comparison of adjacent frame matching results of different algorithms

算法	总匹配点对/对	正确点对/对	正确率/%
SIFT + FLANN 匹配	1 586	1 027	64.75
金字塔 LK 光流法匹配	4 352	2 618	60.16
本文	**986**	**827**	**83.87**

注：加粗字体表示各列最优结果。

消失点为中心，对强匹配特征点对进行高斯分布赋权；在转弯道路中，由于整幅画面都存在变化，因此对所有强匹配特征点对采用均匀分布赋权。

2.3.1　消失点检测

在图像中物体线段通常会在视觉上收敛至消失点，消失点的计算方法如下：

1）提取图像线段。首先，将原始图像转换为灰度图像并采用高斯核函数进行卷积平滑处理。然后，使用 Canny 算子进行边缘检测和 Sobel 算子进一步卷积，计算候选边的强度和方向并得到候选点；通过非极大值抑制剔除非边缘像素值；通过双阈值区分边缘像素强度，在本文中设置高、低阈值分别为 255 和 140；最后，保留强边缘像素以及与其相连的弱边缘像素，从而有效地提取出图像中的边缘线段。图 10 展示了原始图像和 Canny 边缘检测结果。

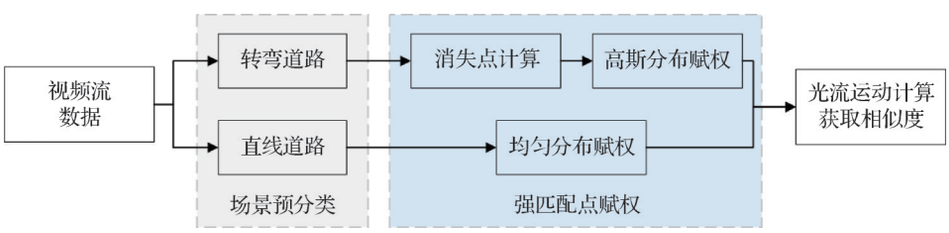
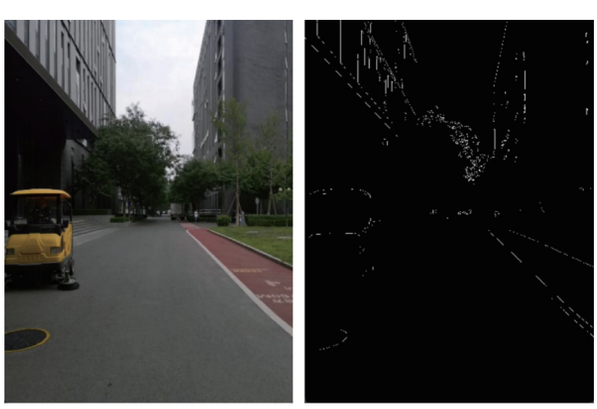


图 9　相邻帧相似度判别

Fig. 9　Similarity discrimination of adjacent frame



(a) 原始图像　(b) Canny 边缘检测结果图

图 10　边缘检测结果

Fig. 10　Edge detection result ((a) original image; (b) Canny edge detection result graph)

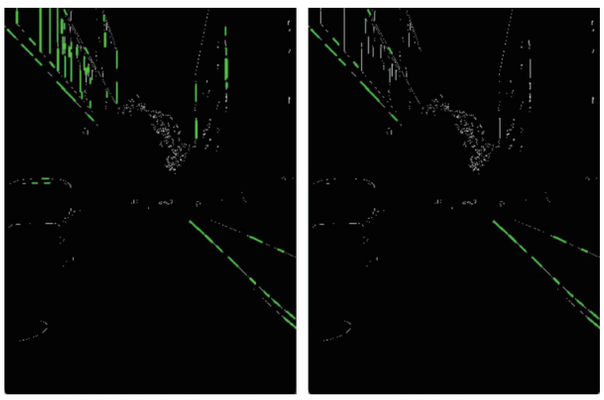
在完成边缘提取后，进一步利用霍夫变换对图像中的线段进行检测。在本文的城市街道场景中，为确保线段的检测精度，设置累加计数的阈值为 50，最小线段长度为 25 像素。

2）线段过滤。从上文中检测到的线段可以看出，位于建筑物或其他地物中的垂直线或水平线在延伸过程中不会经过图像消失点，因此需要进行剔除。为此，设目标线段为 $y_{\text{line}} = mx_{\text{line}} + c$，首先计算图像中每条已检测线段的斜率 m，将该斜率转换为极坐标下的角度表示，记为 $theta = \text{atan}(m)$，单位为弧度。为了剔除接近垂直或水平的线段，设定角度阈值为 4 度，即仅保留角度范围在 $[-86, -4], [4, 86]$ 以内的线段，避免目标线段的角度接近 0 度或 ±90 度，最终实现检测线段过滤。图 11 展示了过滤前后的图像线段检测结果，其中线段用绿色标注。

3）消失点检测。最后利用过滤后的线段进行消失点计算，采用随机抽样一致（random sample consensus，RANSAC）算法迭代估计消失点位置。随机选取两条线段并计算其交点作为候选消失点，依次计算候选点到其余线段的垂直距离 ℓ 并将其累加求和，得到该候选点的误差值，通过迭代搜索，选择误差值最小的候选消失点作为最终消失点。计算结果如图 12 所示，其中红点表示消失点。

2.3.2　相似度计算

计算视频在不同场景下相邻帧的光流变化，利用 2.2 节得到的强匹配点对计算帧间光流移动，如图 13 所示，第 1 幅图中黄线为光流线段，方向和长度分别表示运动矢量的方向和大小；第 2 幅图中对图像的光流运动进行可视化，色彩和亮度的差异分别



(a) 霍夫变换线段检测图　(b) 线段过滤结果图

图11　线段过滤对比图

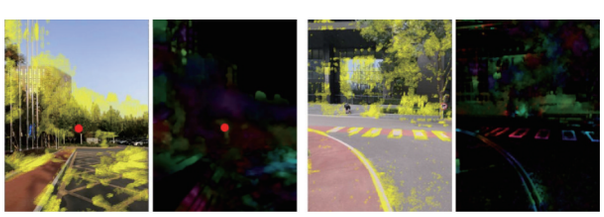
Fig. 11　Line filter comparison diagram

((a) Hough transform line segment detection diagram; (b)line segment filtering result)

图12　图像消失点计算

Fig. 12　Image vanishing point calculation

代表光流矢量方向和大小的不同，颜色越亮说明光流运动变化越大；红点为图像消失点。可以看出：在直线道路处，像素点距离消失点越远，光流运动越大；在转弯道路处，整幅画面持续变化，各个特征点都在向不同的方向移动。



(a) 直线道路光流变化　(b) 转弯道路光流变化

图13　光流场运动可视化

Fig. 13　Visualization of optical flow field motion

((a) change of optical flow on straight road; (b) change of optical flow on turning road)

针对转弯道路图像，采用均匀分布函数赋权，权值均为1；针对直线道路图像，以消失点为中心，采用高斯分布函数对特征点的光流进行赋权。具体为

$$weight = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \sigma = \frac{1}{\sqrt{2\pi}} \tag{22}$$

式中，x为当前像素点与消失点欧氏距离，μ为像素点距离消失点最大欧氏距离，σ为方差。

假设相邻帧图像具有n组强匹配特征点对，在相邻帧中，匹配特征点位置表示分别为$p(p_1, p_2, \cdots, p_n)$和$p'(p'_1, p'_2, \cdots, p'_n)$，其中，$p_i = (x_i, y_i)$，$p'_i = (x'_i, y'_i)$，通过计算匹配点对加权光流变化，得到两幅图像的总光流场聚合运动信息，计算式为

$$optical = \frac{1}{n}\sum_{i=1}^{n} weight_i \cdot \sqrt{(x_i - x'_i)^2 + (y_i - y'_i)^2} \tag{23}$$

式中，$weight_i$为第i组点的光流权重。在完成互校验匹配后，得到强匹配点对分别在FLANN和金字塔LK光流计算下的对应位置，通过式(23)计算两种方法的总光流场运动信息$optical_{\text{flann}}$和$optical_{\text{plk}}$，最后取其均值作为相邻帧最终相似度。具体为

$$sim = \frac{optical_{\text{flann}} + optical_{\text{plk}}}{2} \tag{24}$$

通过上述步骤即可完成一组相邻帧之间的相似度计算，然后根据预先设置的阈值判断是否为关键帧，若小于阈值则保留当前帧为关键帧，否则舍弃当前帧。输入自采集视频，设置视频第1帧为关键帧，并作为前帧图像，依次获取下个视频帧信息作为后帧图像，计算两帧相似度直至出现满足上述条件的关键帧，将其加入到视频关键帧集合中；然后将最新关键帧更新为前帧图像，重复前面操作直至视频帧检测完毕，得到一组最终的视频关键帧集合。通过以上步骤构建了一个高效且可靠的关键帧提取流程，为后续视觉场景重建变得更加精准和高效提供支撑，关键帧提取的整体流程图如图14所示。

3　实　验

3.1　实验环境及数据来源

实验在Ubuntu 20.24系统进行测试，搭载Intel i9-13900HX处理器和NVIDIA GeForce GTX 4070显卡，利用Python 3.8编程实现。

实验使用消费级终端HUAWEI Mate 10中课题组自研开发的视频采集APP(application)，对目标环

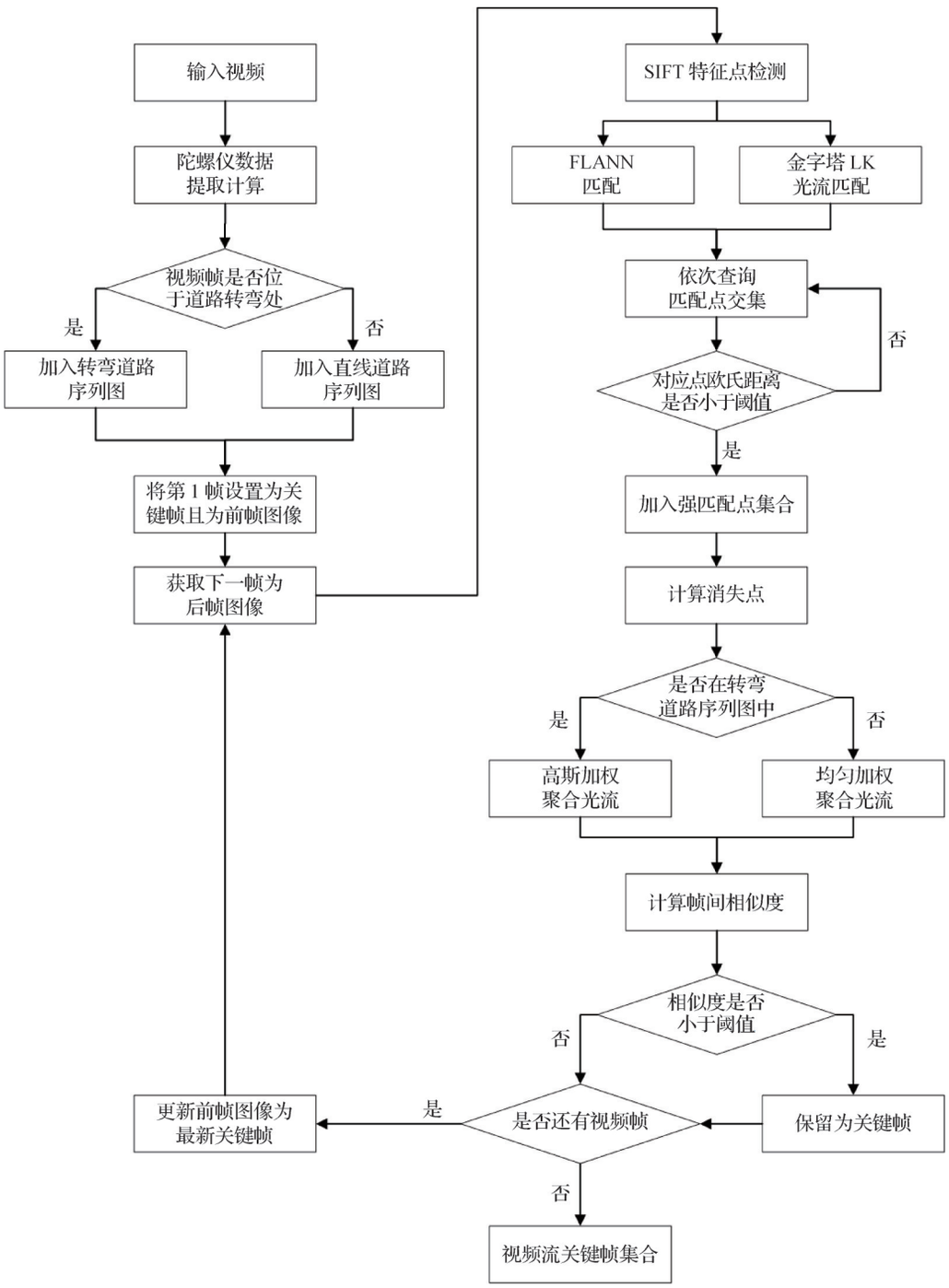


图 14　关键帧提取流程图

Fig. 14　Key frame extraction flow chart

境进行数据自采集，在具有光照和地物分布差异的不同场景中，采用不同的行进路线和速度分别获取了 4 组数据，其中每组都对场景视频信息和陀螺仪数据进行同步记录。图 15 展示了视频数据中的部分视频帧情况。视频帧尺寸均为 1 536 × 2 048 像素，视频帧率为 30 帧/s，各组视频时长、总帧数和采集路线长度具体参数如表 2 所示。

3.2　关键帧提取方法实验验证

实验分别采用本文算法、时间间隔采样提取算法（Yang 等，2015）、光流运动提取算法（Bao 等，2020）和颜色直方图提取算法（Zhao 等，2019）对视频进行关键帧筛选；同时为了验证本文提出的场景预分类以及互校验匹配计算的算法性能和有效性，将光流运动提取算法、场景预分类 + 光流提取算法

图15　部分视频帧

Fig. 15　Partial video frame

和本文算法(场景预分类+光流+特征互校验)进行进一步对比。统计从原视频中筛选出来的关键帧数量,然后使用结构相似性指数(structure similarity index measure,SSIM)计算图像之间的视觉相似度,统计高相似度图像帧数,其中,高相似度帧为提取出来的视频关键帧中两幅图像之间具备高度视觉相像的关键帧,数量越大说明提取出来的关键帧冗余越大,表3展示了以上方法提取关键帧的结果对比。

表2　视频数据参数

Table 2　Video data parameter

名称	采集时间	时长/min	总帧数/帧	路线长度/m
数据1	2024.11.19 10:28:51	6.62	11 017	425
数据2	2024.11.19 10:58:42	6.42	10 687	381
数据3	2024.11.09 16:38:39	6.19	10 325	459
数据4	2024.04.10 16:28:26	17.91	29 797	862

表3　关键帧提取结果对比

Table 3　Comparison of key frame extraction results

算法	提取关键帧数量/帧				高相似度帧数量/帧				提取时间/min			
	数据1	数据2	数据3	数据4	数据1	数据2	数据3	数据4	数据1	数据2	数据3	数据4
时间间隔采样	**736**	**716**	**690**	1 988	92	37	29	31	**1.86**	**1.81**	**1.75**	**5.04**
颜色直方图	1 230	848	1 375	**1 239**	127	102	134	97	21.64	19.86	17.29	53.88
光流运动	1 408	960	1 634	1 852	173	89	91	279	28.56	25.41	22.57	62.96
场景预分类+光流	1 097	869	1 412	1 682	106	62	69	156	36.74	32.67	27.83	76.92
本文	876	725	1 143	1 359	**19**	**15**	**32**	**26**	49.65	47.53	46.48	137.49

注:加粗字体表示各列最优结果。

在不同场景中,直线和转弯处用于三维重建的图像数量有所不同。直线处由于场景变化较小,需要在满足重建需求的情况下尽量减小图像帧数量,从而降低BA(bundle adjustment)优化计算代价和累积误差;而转弯处由于场景变化较大,图像帧数量过少可能影响前后帧特征点匹配,造成图像匹配错误或因特征点对不足导致模型构建中断的情况,因此预期提取出来的关键帧应在直线道路尽可能少,同时在转弯道路满足一定数量。对上面几种算法筛选的关键帧和原视频帧进行分析,分别统计在直线和转弯处的数量,并计算在不同场景中关键帧占原视频帧百分比情况(即直线或转弯道路处提取出来的关键帧数量/视频直线或转弯道路处视频帧总数量),从而对算法的场景自适应性进行评估,统计结果如表4所示。

从实验结果可以看出,相较于其他算法,本文算法为了提高相邻帧匹配准确性,同时融合了特征点检测匹配和光流计算,一定程度上增加了时间成本,在关键帧提取效率方面存在劣势,但是从最终整体的提取效果来看,本文算法提取的关键帧数量相对较少,能够将图像总数量降低到10%左右,最低可以达到4.56%;同时关键帧中高相似度帧数量占比均在3%以下,最低达到1.91%,明显优于其他算法;并且本文算法提取的关键帧在转弯道路处的数量占比远大于直线道路处的占比,相较于其他算法效果较佳。

将光流运动算法与场景预分类+光流算法的提取结果进行对比,后者通过预先进行场景划分,能够有效降低关键帧的数量和高相似度帧数,同时提高

表 4　直线与转弯处关键帧提取结果对比

Table 4　Comparison of key frame extraction results between straight road and turning road

算法	数据 1						数据 2					
	直线道路			转弯道路			直线道路			转弯道路		
	总视频帧	提取	占比/%	总视频帧	提取	占比/%	总视频帧	提取	占比/%	总视频帧	提取	占比/%
时间间隔采样		638	6.68		98	6.68		611	6.67		105	6.68
颜色直方图		1 044	10.93		186	12.68		632	6.90		216	13.74
光流运动	9550	1 201	12.58	1 467	207	14.11	9 155	796	8.69	1 572	164	10.43
场景预分类 + 光流		786	8.23		311	21.20		562	6.14		307	19.53
本文		**552**	**5.78**		**324**	**22.09**		**356**	**3.89**		**369**	**23.47**

算法	数据 3						数据 4					
	直线道路			转弯道路			直线道路			转弯道路		
	总视频帧	提取	占比/%	总视频帧	提取	占比/%	总视频帧	提取	占比/%	总视频帧	提取	占比/%
时间间隔采样		553	6.68%		137	6.69%		1 348	6.67%		640	6.68%
颜色直方图		1 013	12.24%		362	17.68%		645	3.19%		594	6.20%
光流运动	8 277	1 276	15.42%	2 048	358	17.48%	20 209	1 149	5.69%	9 588	703	7.33%
场景预分类 + 光流		873	10.55%		539	26.32%		801	3.96%		881	9.19%
本文		**665**	**8.03**%		**478**	**23.34**%		**571**	**2.83**%		**788**	**8.22**%

注：加粗字体表示各列最优结果。

关键帧在转弯道路处的占比，验证了场景预分类的有效性；将场景预分类 + 光流算法与本文算法（场景预分类 + 光流 + 特征互校验）的提取结果进行对比，后者通过基于光流和SIFT特征点的互校验匹配计算，进一步在关键帧数量、高相似度帧数以及不同场景下关键帧占比情况上得到了优化，说明利用互校验匹配计算帧间光流差异，能够有效降低相邻帧的错误匹配，并且提高帧间相似度计算的准确性。

综上，本文算法能够在直线道路中适当降低图像帧数量，同时在图像画面运动较大的转弯道路中，适当提高图像抽帧数量，进而保证特征点匹配准确性和增量式重建的完整性，符合在三维重建应用下，在不同场景中对图像数量的预期需求。

3.3　基于关键帧的模型三维重建实验验证

实验分别采用本文算法、时间间隔采样提取算法、光流运动提取算法和颜色直方图提取算法进行视频关键帧筛选，在增量式 SfM（structure from motion）三维重建过程中，使用提取的关键帧进行模型三维重建实验验证。由于重建过程中可能出现模型断裂的情况，故仅统计能够成功进行BA优化并加入重建序列的最大关键帧数量，该数量与提取的总关键帧数量之比越大，表示关键帧的利用率越高；但是因为关键帧冗余会导致多幅图像位姿恢复错误的情况，因此较高的利用率并不一定说明有更好的重建效果。为评估重建效果，利用消费级终端记录的粗略GNSS数据绘制路线图，可以对行进路线进行判断，并与重建的三维模型进行对比，根据重建出来的道路长度评估模型的完整度，统计结果如表5所示；由4类自采集数据的GNSS粗略坐标绘制的路线图和重建效果可视化如图16所示。

由表5结果可见，光流提取算法在数据3重建中，虽然重建成功图像数量占比达到71.79%，但是模型完整度只有15.73%，说明此方法提取的关键帧存在较高的冗余或缺失情况，导致最终重建模型错误；颜色直方图提取算法在数据4重建中，模型完整度达到63.56%，但是从模型可以看出，虽然成功恢复出一定的道路长度，但是理论上连贯的道路，实际模型效果在转弯处存在偏差断裂情况，说明在转弯

表5　三维重建关键帧利用率与模型完整度对比

Table 5　Comparison of key frame utilization and model integrity in 3D reconstruction

算法	重建成功图像占比/%				模型完整度/%				重建时间/min			
	数据1	数据2	数据3	数据4	数据1	数据2	数据3	数据4	数据1	数据2	数据3	数据4
时间间隔采样	29.62	52.23	36.81	63.18	28.76	48.72	41.25	42.57	**4.98**	**17.16**	**25.7**	**146.87**
颜色直方图	78.46	**100.00**	44.36	66.75	88.57	99.52	23.36	63.56	497.98	243.05	524.35	231.15
光流运动	53.70	87.81	71.79	26.67	30.57	67.59	15.73	22.69	663.34	271.07	841.85	289.40
场景预分类+光流	63.16	51.75	78.71	45.82	51.23	47.62	56.37	42.16	821.88	379.75	952.85	721.78
本文	**100.00**	**100.00**	**98.08**	**95.51**	**100.00**	**100.00**	**97.46**	**96.54**	523.27	251.35	497.32	983.69

注:加粗字体表示各列最优结果。

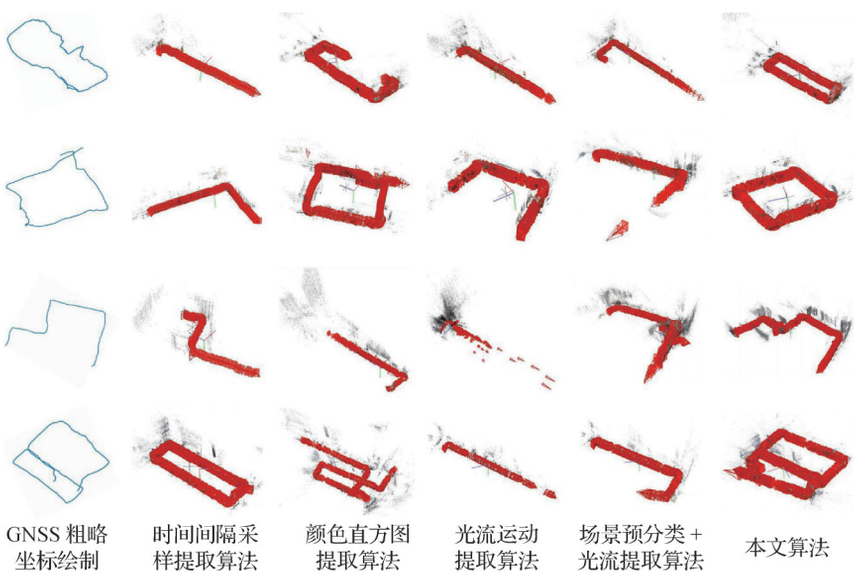


图16　重建效果对比

Fig. 16　Comparision of reconstruction effects

处对图像质量要求较高的场景下,此方法提取的关键帧之间存在错误匹配,并不能很好适用于场景;场景预分类+加权光流提取算法相较于光流提取算法的重建结果,能够在一定程度上提高重建成功图像数量占比和模型完整度,但是相较于本文算法仍然效果较差,验证了前文提出的场景预分类的适应性以及通过互校验匹配计算可以提高相邻帧相似度计算的准确性。本文算法在4组自采集数据上重建出来的模型完整度分别高达100%、100%、97.46%和96.54%。在重建时间方面,由于其他提取算法筛选的关键帧无法成功重建出较完整的场景模型,提前终止计算,从而消耗时间较少;而本文算法在完成目标场景重建的同时,能够有效避免因计算提前结束导致的模型重建不完整问题,并且在与其他算法恢复出相似规模的场景计算相比,能够在一定程度上提高重建效率,同时弥补在关键帧提取时造成的时间成本损失。总体而言,相较于其他方法,本文算法在视频关键帧筛选上,通过互校验匹配计算,提高匹配点对精确性,保证了关键帧间具有较好的视角变化,将直线和转弯场景进行区分,针对性加权计算相邻帧总光流变化情况,构建了一个稳健的相似度判别方法,最终利用提取出来的关键帧恢复得到的场景模型完整度具有一定优势,能够较好地应用于增量式重建,保证三维模型的准确性和鲁棒性。

3.4　算法复杂度分析

本文主要针对基于互校验匹配搜索帧间强匹配点进行复杂度分析。假设待匹配的前后两个视频帧有n_1和n_2个特征点,利用传统暴力搜索匹配计算的复杂度为$O(n_1\times n_2)$。本文使用FLANN匹配计算可以分为KD树构建阶段和查询匹配阶段,在KD树构

建阶段，通常构建一个 KD 树的复杂度为 $O(n/\log(n))$，其中 n 是数据点的数量，对于 n_1+n_2 个特征点，构建 FLANN 的 KD 树复杂度为 $O((n_1+n_2)/\log(n_1+n_2))$；在查询匹配阶段，通过快速最近邻匹配加速查询，复杂度为 $O(n_1/\log(n_2))$。最终 FLANN 计算复杂度为 $O((n_1+n_2)/\log(n_1+n_2))+O(n_1/\log(n_2))$，相较于暴力搜索能够很好地提高效率。在金字塔 LK 匹配阶段，假设构建 F 层图像金字塔，对第 1 帧图像，n_1 个特征点进行光流计算的复杂度为 $O(F\times n_1)$。在互校验匹配阶段，查找两种方法在第 2 帧图像中共同匹配的点位，遍历第 1 帧图像中 n_1 个特征点的复杂度为 $O(n_1)$，假设共有 n_{12} 个同时成功匹配的特征点，对其进行欧氏距离计算筛选强匹配点的计算复杂度为 $O(n_{12})$，最终复杂度为 $O(n_1)+O(n_{12})$。综上，本文基于互校验加权匹配的视频三维重建关键帧提取算法总复杂度为 $O((n_1+n_2)/\log(n_1+n_2))+O(n_1/\log(n_2))+O(F\times n_1)+O(n_1)+O(n_{12})$。

增量式 SfM 重建时，假设有 n 帧图像，每帧提取出 f 个特征点，在特征点提取与匹配阶段，每两个图像之间的匹配复杂度为 $O(f^2)$，总计算复杂度为 $O(n\times f^2)$；在相机位姿估计阶段，每次添加新帧需要估计新帧与已有帧之间的相对位姿，其计算复杂度为 $O(n\times f)$；在 BA 优化阶段，随着帧数和三维点数的增加，每次优化的计算复杂度为 $O((N+M)^2)$，其中，N 和 M 分别为已经恢复的相机和三维点个数。因此，增量式 SfM 的总计算复杂度通常会随着帧数增加而呈二次增长，特别在 BA 阶段存在计算瓶颈，其导致整体复杂度随帧数和三维点数的更新而增加，重建过程中每恢复增加数量的图像就进行一次 BA 优化，假设总共进行 k 次优化，则总复杂度可以表示为 $\sum_{i=1}^{k}O(N_i+M_i)$。

因此，相较于利用所有视频帧进行增量式重建，虽然本文对视频进行了抽帧处理，增加互校验加权光流计算这一步骤，但是此步骤总体计算复杂度较小，对整体流程代价影响较低。然而在重建阶段，假设视频帧总计 10 000 帧，每 10 帧进行一次 BA 优化，将进行 1 000 次计算，复杂度具有 2 次方的指数运算，其中每次 BA 时已重建的相机和三维点数量都在累计增加，总体三维重建计算将具有很大代价，在本文中通过抽帧处理，将图像总数量降低到 1/10，仅需进行 100 次优化，同时大量降低 N 和 M 的数量，处理的数据量相应减小，这直接降低了计算资源的成本，同时抽帧能够去除冗余的图像，减少重复计算，保留更有意义的信息，进而加速后续 BA 和三维模型构建过程。因此，本文算法能够较好地降低总体计算复杂度，提高三维模型构建的效率。

4 结 论

本文总结了当前关键帧提取算法在三维模型重建中存在的问题，提出一种基于互校验加权光流的视频三维重建关键帧提取方法。首先，将自采集的视频流数据作为输入，利用终端陀螺仪数据对直线和转弯场景进行划分，完成视频场景预分类，为后续相似度计算提供基础；然后，对视频帧 SIFT 特征点进行提取描述，将 FLANN 匹配和金字塔 LK 光流法综合纳入考量，采取互校验匹配方法筛选相邻帧强匹配点对；最后，计算图像消失点，考虑不同场景中图像像素对光流计算贡献的差异，以消失点为中心，对其周围像素点采用灵活的加权策略，得到相邻帧总光流场聚合变化获取相似度，最后根据设置的阈值筛选关键帧，从而提取出一组视频流关键帧集合。

在 4 类自采集视频数据上对本文算法开展实验验证，从结果可以看出，相较于其他关键帧提取算法，本文算法在多样化场景中，能够有效降低视频帧数量，针对性提高关键帧筛选的质量，同时提高三维重建的鲁棒性，场景重建模型的效果和完整度最优。但是关键帧提取算法在效率方面有待提升，并且利用关键帧进行场景重构时仍然存在少部分模型不完整的问题，因此在后续需要进一步考虑关键帧筛选与三维重建之间的关联性，同时融合图像多源信息进行辅助提取和重建，提高关键帧筛选效率和质量，从而重建出场景更完整的模型。

参考文献

Bao G B, Li D F and Mei Y L. 2020. Key frames extraction based on optical-flow and mutual information entropy. Journal of Physics: Conference Series, 1646(1): #012112 [DOI: 10.1088/1742-6596/1646/1/012112]

Deshpande A, Bamnote V, Patil B and Tonge A A. 2018. Review of key-frame extraction techniques for video summarization. International Journal of Computer Applications, 180(39): 40-43 [DOI: 10.5120/ijca2018917042]

Dong J S. 2023. Study on video key frame extraction in different scenes based on optical flow. Journal of Physics: Conference Series, 2646(1): #012035 [DOI: 10.1088/1742-6596/2646/1/012035]

Fang Z Y. 2022. Research on Rapid 3D Reconstruction Technology based on Video Keyframe Screening. Chengdu: University of Electronic Science and Technology of China (方子赟. 2022. 基于视频关键帧筛选的快速三维重建技术研究. 成都: 电子科技大学) [DOI: 10.27005/d.cnki.gdzku.2022.003423]

Jekel C F and Venter G. 2019. PWLF: a python library for fitting 1D continuous piecewise linear functions [DOI: 10.13140/RG. 2.2. 28530.56007]

Li Z Q. 2024. A method for recognising wrong actions of martial arts athletes based on keyframe extraction. International Journal of Biometrics, 16(3/4): 256-271 [DOI: 10.1504/IJBM.2024.138228]

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]

Mushan A and Vidap P S. 2020. Video summarization using keyframe extraction methods. International Journal of Recent Technology and Engineering, 9(2): 1030-1032 [DOI: 10.35940/ijrte. B4043. 079220]

Pandian A A and Maheswari S. 2024. A keyframe selection for summarization of informative activities using clustering in surveillance videos. Multimedia Tools and Applications, 83(3): 7021-7034 [DOI: 10.1007/S11042-023-15859-Z]

Paul M K A, Kavitha J and Rani P A J. 2018. Key-frame extraction techniques: a review. Recent Patents on Computer Science, 11(1): 3-16 [DOI: 10.2174/2213275911666180719111118]

Yan G and Woźniak M. 2022. Accurate key frame extraction algorithm of video action for aerobics online teaching. Mobile Networks and Applications, 27(3): 1252-1261 [DOI: 10.1007/s11036-022-01939-1]

Yan J. 2023. Extraction of key frames from dance videos and movement recognition by multi-feature fusion. IEIE Transactions on Smart Processing and Computing, 12(6): 495-501 [DOI: 10.5573/IEIESPC. 2023.12.6.495]

Yang T, Li J, Yu J Y, Wang S B and Zhang Y N. 2015. Diverse scene stitching from a large-scale aerial video dataset. Remote Sensing, 7(6): 6932-6949 [DOI: 10.3390/rs70606932]

Yuan Y, Lu Z, Yang Z, Jian M, Wu L F, Li Z Y and Liu X. 2022. Key frame extraction based on global motion statistics for team-sport videos. Multimedia Systems, 28(2): 387-401 [DOI: 10.1007/s00530-021-00777-7]

Zarco-Tejada P J, Diaz-Varela R, Angileri V and Loudjani P. 2014. Tree height quantification using very high resolution imagery acquired from an unmanned aerial vehicle (UAV) and automatic 3D photo-reconstruction methods. European Journal of Agronomy, 55: 89-99 [DOI: 10.1016/j.eja.2014.01.004]

Zhang H, Lu X P, Zhang X Q and Lu Z Z. 2020. Dynamic extraction method of key frame image of UAV video for mine supervision. Remote Sensing Information, 35(1): 112-116 (张航, 卢小平, 张晓强, 路泽忠. 2020. 面向矿山监管的无人机视频关键帧影像动态提取方法. 遥感信息, 35(1): 112-116) [DOI: 10.3969/j. issn.1000-3177.2020.01.015]

Zhang M L. 2020. Design and Implementation of Visual Positioning System based on 3D Reconstruction. Nanjing: Nanjing University (张铭磊. 2020. 基于三维重建的视觉定位系统设计与实现. 南京: 南京大学) [DOI: 10.27235/d.cnki.gnjiu.2020.000833]

Zhang X. 2022. Research on Image Matching Algorithm in Indoor Visual Localization. Beijing: Beijing University of Posts and Telecommunications (张霄. 2022. 室内视觉定位中图像匹配算法研究. 北京: 北京邮电大学) [DOI: 10.26969/d.cnki.gbydu.2022.001159]

Zhao H, Wang W J, Wang T, Chang Z B and Zeng X Y. 2019. Key-Frame extraction based on HSV histogram and adaptive clustering. Mathematical Problems in Engineering, 2019: #52179611-10 [DOI: 10.1155/2019/5217961]

Zheng Y J, Chen W W, Luo J X, Pan Z S, Zhang Y Y and Sun H X. 2023. Adaptive step size video key frame extraction for 3D reconstruction. Software Guide, 22(9): 159-166 (郑义桀, 陈卫卫, 罗健欣, 潘志松, 张艳艳, 孙海迅. 2023. 面向三维重建的自适应步长视频关键帧提取. 软件导刊, 22(9): 159-166) [DOI: 10. 11907/rjdk.222214]

作者简介

张泷丹,女,硕士研究生,主要研究方向为计算机视觉和视觉三维重建。E-mail:zhangld07@163.com

李雯,通信作者,女,高级工程师,主要研究方向为多源融合导航、众包建图和室内定位。E-mail:wen.li@aircas.ac.cn

魏东岩,男,研究员,主要研究方向为多源融合导航、自主定位和导航通信融合。E-mail:weidy@aircas.ac.cn

马朝伟,男,助理工程师,主要研究方向为视觉三维重建、视觉惯性里程计和图像检索定位。E-mail:macw@aircas.ac.cn

李政祎,男,博士研究生,主要研究方向为计算机视觉和图像检索定位。E-mail:lizhengyi22@mails.ucas.ac.cn

申戈,男,高级工程师,主要研究方向为无线通信和导航通信融合。E-mail:shenge@aircas.ac.cn