# A Head-Wearable Short-Baseline Stereo System for the Simultaneous Estimation of Structure and Motion

Hern´an Badino and Takeo Kanade Carnegie Mellon University Pittsburgh, PA, USA

# Abstract

This paper presents a short-baseline real-time stereo vision system that is capable of the simultaneous and robust estimation of the ego-motion and of the 3D structure and the independent motion of thousands of points of the environment. Kalman filters estimate the position and velocity of world points in 3D Euclidean space. The six degrees of freedom of the ego-motion are obtained by minimizing the projection error of the current and previous clouds of static points. Experimental results with real data in indoor and outdoor environments demonstrate the robustness, accuracy and efficiency of our approach. Since the baseline is as short as 13cm, the device is head-mountable, and can be used by a visually impaired person. Our proposed system can be used to augment the perception of the user in complex dynamic environments

本文提出了一种短基线实时立体视觉系统，该系统能够同时和稳健地估计自我运动和 3D 结构以及环境中数千个点的独立运动。卡尔曼滤波器估计 3D 欧几里得空间中世界点的位置和速度。自我运动的六个自由度是通过最小化当前和先前静态点云的投影误差来获得的。室内和室外环境中真实数据的实验结果证明了我们方法的稳健性、准确性和效率。由于基线短至13cm，该设备是头戴式的，可供视障人士使用。我们提出的系统可用于增强用户在复杂动态环境中的感知。

# Introduce

In this paper we present a wearable mobile system that provides information of 3D structure, independent motion and ego-motion to augment the perception of the user in complex environments. We have built a prototype consisting of a helmet with stereo cameras connected to a portable computer system (Figure 2). As the user navigates in its environment, the system detects and tracks image features and computes their corresponding stereo disparities. The features and disparities of consecutive frames are used to compute the ego-motion of the camera using a robust least squares algorithm. A Kalman filter then fuses the feature tracking, the stereo disparity and the extracted egomotion to iteratively estimate the 3D position and 3D velocity of each tracked feature in a user oriented coordinate system. Figure 1 shows an example of the output obtained in a crowded scene containing multiple moving objects. Our proposed methods can be used to augment the perception of the visually impaired in complex dynamic environments. The system runs in real time at 17 Hz using a standard laptop, and it was tested online in countless occasions.

在本文中，我们提出了一种可穿戴移动系统，该系统提供 3D 结构、独立运动和自我运动信息，以增强用户在复杂环境中的感知。我们已经构建了一个原型，该原型由一个头盔和连接到便携式计算机系统的立体摄像头组成（图 2）。当用户在其环境中导航时，系统会检测和跟踪图像特征并计算它们相应的立体差异。连续帧的特征和视差用于使用稳健的最小二乘算法计算相机的自我运动。然后卡尔曼滤波器融合特征跟踪，立体视差和提取的自我运动以迭代估计面向用户的坐标系统中每个跟踪特征的 3D 位置和 3D 速度。图 1 显示了在包含多个移动对象的拥挤场景中获得的输出示例。我们提出的方法可用于增强复杂动态环境中视障者的感知。该系统使用标准笔记本电脑以 17 Hz 的频率实时运行，并在无数场合进行了在线测试。

# Related Literature

The simultaneous estimation of structure and motion from stereo images has been heavily covered by the literature. We give here a brief review of the most related methods. Jung and Lacroix [4] uses Kalman filters to refine estimates of ego-motion and 3D landmark position of static world points. Agrawal et al. [1] and Talukder and Matthies [13] estimate independently moving objects by detecting and tracking blobs in the image. The blobs are obtained from image regions that are not in accordance with the computed ego-motion. Similarly, Ess et al. [2] present a framework for the detection of independent motion with a freely moving camera in crowded scenes. Franke et al. [3] use Kalman filters to track independent motion using stereo cameras. The ego-motion of the cameras is obtained from the inertial sensors of a vehicle. Rabe et al. [9] extended this approach to dense motion fields using FPGA and GPU implementations. Klein and Murray [5] also track features using a monocular system for the real time estimation of camera motion and structure for augmented reality applications. Independent motion is not modeled but treated as outlier. Vision has also been used to provide navigation support to the visually impaired. A survey of navigation systems of the visually impaired can be found in [15]. Lu and Manduchi [7] present a stereo system to detect curbs and stairways. S´aez and Escolano [12] detect aerial obstacles in near real time, but only static scenes are considered. Treuillet et al. [14] propose a similar application to localize and guide the walker along a predefined path by using a monocular camera

从立体图像中同时估计结构和运动已被文献大量报道。我们在这里简要回顾最相关的方法。 Jung 和 Lacroix [4] 使用卡尔曼滤波器来改进对静态世界点的自我运动和 3D 地标位置的估计。阿格拉瓦尔等人。[1] 以及 Talukder 和 Matthies [13] 通过检测和跟踪图像中的斑点来估计独立移动的物体。斑点是从不符合计算的自我运动的图像区域获得的。同样，Ess 等人。

[2] 提出了一个框架，用于在拥挤的场景中使用自由移动的相机检测独立运动。弗兰克等人。 [3] 使用卡尔曼滤波器来跟踪使用立体相机的独立运动。摄像机的自我运动是从车辆的惯性传感器获得的。拉贝等人。[9] 使用 FPGA 和 GPU 实现将这种方法扩展到密集运动场。 Klein 和 Murray [5] 还使用单目系统跟踪特征，以实时估计相机运动和增强现实应用的结构。独立凹痕运动未建模，但被视为异常值。视觉也被用于为视障者提供导航支持。可以在 [15] 中找到对视障者导航系统的调查。 Lu 和 Manduchi [7] 提出了一个立体系统来检测路缘石和楼梯。 S´aez 和 Escolano [12] 近乎实时地检测空中障碍物，但只考虑静态场景。

Treuillet 等人。 [14] 提出了一个类似的应用程序，通过使用单目相机来定位和引导步行者沿着预定义的路径

In contrast to the above methods, we track stereo features and use Kalman filters to estimate their position and velocity in Euclidean space while simultaneously estimating the ego-motion of the camera using a robust method. Furthermore, our proposed system runs in real time at 17 Hz using a single laptop.

与上述方法相比，我们跟踪立体特征并使用卡尔曼滤波器来估计它们在欧几里得空间中的位置和速度，同时使用稳健的方法估计相机的自我运动。此外，我们提出的系统使用一台笔记本电脑以 17 Hz 的频率实时运行。

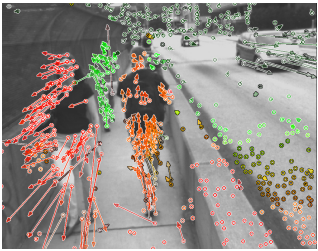


Figure 1. Output of our system while moving in a com plex environment. Tracked features are shown as a cir cle. The vectors show the direction and speed of the features on moving objects.

图 1. 在复杂环境中移动时我们系统的输出。跟踪的特征显示为一个圆圈。矢量显示移动物体上特征的方向和速度。

# Tracking 3D Points

This section presents a Kalman filter method that iteratively estimate the position and velocity of individually tracked feature points. The Kalman filter we describe in this section is derived from Franke et al. [3], where it was used to detect moving object in the automotive domain. We have expanded the system model equations to allow a freely moving camera instead of a motion on the plane, as originally proposed.

本节介绍了一种卡尔曼滤波器方法，该方法迭代地估计单独跟踪的特征点的位置和速度。我们在本节中描述的卡尔曼滤波器源自 Franke 等人。 [3]，它被用来检测汽车领域的运动物体。我们已经扩展了系统模型方程，以允许相机自由移动，而不是像最初提出的那样在平面上运动。

**System Model.** Let pk−1 = (X, Y, Z) T represent the coordinate vector of a world point observed by the system at time k − 1 and vk−1 = (X, ˙ Y , ˙ Z˙) T represent its associated velocity vector. The camera platform moves in its environment with a given translational and angular velocity, changing its relative position to the point. After a time ∆tk the new position of the point from the camera point of view is given by

系统模型。令 表示系统在时间 k - 1 观察到的世界点的坐标向量，而 表示其相关速度向量。相机平台以给定的平移和角速度在其环境中移动，从而改变其与该点的相对位置。经过一段时间 后，从相机视点到该点的新位置由下式给出



where Rk and tk are the rotation matrix and translation vector of the static scene with respect to the camera. The velocity vector vk changes its direction according to:

其中 和 是静态场景相对于相机的旋转矩阵和平移向量。速度矢量 根据以下公式改变其方向：



Combining position and velocity in the state vector

在状态向量中结合位置和速度



leads to the discrete linear system model equation:

导致离散线性系统模型方程：



with the state transition matrix

与状态转移矩阵



and input vector



The term ρk is assumed to be Gaussian white noise. Measurement Model. A measurement is defined by the vector m = (u, v, d) T , where (u, v) corresponds to the image position of the feature point and d is its disparity. The (u, v) components are obtained from the feature tracking algorithm, while the disparity d is obtained from the stereo algorithm. The non-linear measurement equation h for the state vector of Equation 1 is

假设 是高斯白噪声。测量模型。测量由向量 m = (u, v, d) T 定义，其中 (u, v) 对应于特征点的图像位置，d 是其视差。 (u, v) 分量是从特征跟踪算法中获得的，

而视差d是从立体算法中获得的。方程 1 的状态向量的非线性测量方程 h 是



where f is the focal length of the camera and B is the baseline of the stereo system. The term ν is assumed to be Gaussian white noise. Since the measurement equation is non-linear, the extended Kalman Filter is used.

其中 f 是相机的焦距，B 是立体系统的基线。假设 ν 是高斯白噪声。由于测量方程是非线性的，因此使用了扩展卡尔曼滤波器。

# Visual Odometry Estimation

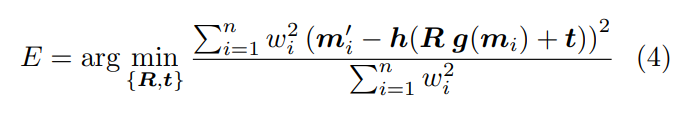
The computation of ego-motion is one of the most fundamental tasks for most mobile vision problems. The accurate knowledge of the motion of the camera allows to integrate estimates in a global coordinate system. It also provides a solid constraint to detect independent motion. Equations 2 and 3 require Rk and tk (the rotation and translation of the static scene, i.e. the inverse motion of the camera). This section presents a method for their robust estimation.

自我运动的计算是大多数移动视觉问题的最基本任务之一。相机运动的准确知识允许在全局坐标系中整合估计。它还提供了可靠的约束来检测独立运动。等式 2 和 3 需要 和 （静态场景的旋转和平移，即相机的反向运动）。本节介绍了一种稳健估计的方法。

**Least Squares Formulation.** Given a set of tracked feature points mi = (ui , vi , di) T for i = 1, 2, · · · , n in the current frame, and the set of corresponding feature points m0 i = (u 0 i , v0 i , d0 i ) T in the previous frame, we seek to estimate the rotation matrix R and translation vector t, such that for all points in the sets, g(m0 i ) = R g(mi) + t, with g() = h −1 (), i.e., the triangulation equation. One way of obtaining the translation and rotation is to calculate the absolute orientation between both set of points. Many solutions to the absolute orientation problem exist when the error in the 3D points is isotropic [6]. However, stereo triangulation error can be highly anisotropic and correlated [11]. Instead of minimizing the residuals in Euclidean space, we minimize them in the image space, where the noise level is similar for all components of the measurement vector:

**最小二乘公式:** 给定当前帧中 i = 1, 2,···, n 的一组跟踪特征点 ，以及对应的特征点集 在前一帧中，我们寻求估计旋转矩阵 R 和平移向量 t，

这样对于集合中的所有点，，其中 ，即三角方程。获得平移和旋转的一种方法是计算两组点之间的绝对方向。当 3D 点中的误差是各向同性的时，存在许多绝对方向问题的解决方案 [6]。然而，立体三角测量误差可能是高度各向异性和相关的[11]。我们不是在欧几里得空间中最小化残差，而是在图像空间中最小化它们，其中，测量向量的所有分量的噪声水平都相似：

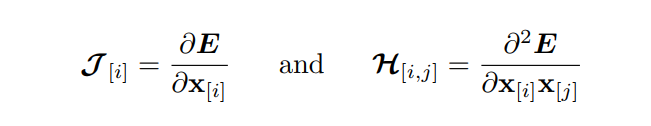


where wi is a weighting factor determining the contribution of the measurement to the least square solution. In order to minimize Equation 4, the rotation matrix R is parameterized by the pseudo-vector r = (wx, wy, wz) T . The matrix R is obtained by rotating the identity matrix |r| radians around the axis r/|r|. Assuming t = (tx, ty, tz) T , the parameter for minimization is then the six-dimensional vector x = (wx, wy, wz, tx, ty, tz) T .

其中是确定测量对最小二乘解的贡献的加权因子。为了最小化等式 4，旋转矩阵 R 由伪向量 参数化。通过旋转单位矩阵 |r| 得到矩阵 R绕轴 的弧度。假设 ，那么最小化的参数就是六维向量 。

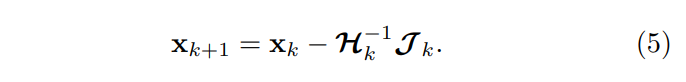
**Newton Minimization.** Because of the non-linearity imposed by the rotation and the projection equation h(), we use an iterative Newton optimization method to solve Equation 4, for which we require the computation of first and second order derivatives of the loss function, i.e.:

牛顿最小化。由于旋转和投影方程 h() 施加的非线性，我们使用迭代牛顿优化方法来求解方程 4，为此我们需要计算损失函数的一阶和二阶导数，即：



Given an initial estimate x0, the Newton method iteratively converge to a local minimum

给定初始估计 ，牛顿法迭代收敛到局部最小值

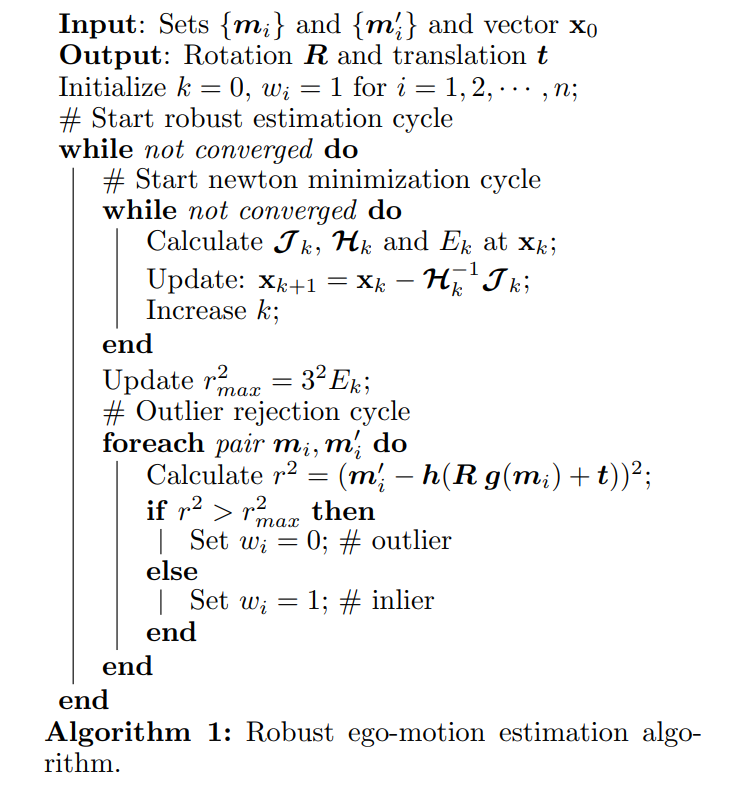


Equation 5 is iteratively applied until the residual Ek of Equation 4 is small enough, or no significant change in the estimate is observed. The closer the initial value is to the real solution, the less iterations are required to find a minimum. For small camera motion, it is usually enough to set x0 to the zero vector. For a large motion between frames, an inertial sensor unit can provide the initial estimate of the motion of the camera. For the experiments shown in Section 5, the method converges in less than six iterations.

迭代地应用等式 5，直到等式 4 的残差 足够小，或者观察到估计值没有显着变化。初始值越接近真实解，找到最小值所需的迭代次数就越少。对于小的摄像机运动，通常将 设置为零向量就足够了。对于帧之间的大运动，惯性传感器单元可以提供相机运动的初始估计。对于第 5 节中所示的实验，该方法在不到六次迭代中收敛。

**Iterative Robust Estimation.** The sets of feature correspondences often contain outliers due to false correspondences or moving objects. In order to cope with the outliers, an optimization approach is applied that iteratively rejects them. Assuming that the outliers are bounded (constraint imposed by the stereo and tracking algorithm), the motion estimated by the Newton method will be approximately correct, and therefore, the outliers will have a larger residual than the inliers. This provides a simple method for outliers rejection: if the residual for a given feature is larger than a threshold, the feature is eliminated from the set, and the whole estimation process is repeated until convergence (Algorithm 1). A nice property of the above procedure is that the Newton method converges faster after each robust estimation cycle, since the resulting estimate is closer to the solution. After the initial estimate, the Newton method usually requires only two cycles to converge.

**迭代稳健估计。**由于错误的对应或移动的对象，特征对应集通常包含异常值。为了处理异常值，应用了一种迭代拒绝它们的优化方法。假设异常值是有界的（由立体和跟踪算法施加的约束），牛顿法估计的运动近似正确，因此离群点的残差比离群点大。**这为异常值拒绝提供了一种简单的方法：如果给定特征的残差大于阈值，则从集合中消除该特征**，并且重复整个估计过程直到收敛（算法1）。上述过程的一个很好的特性是牛顿法在每个稳健估计周期后收敛得更快，因为得到的估计更接近解。在初始估计之后，牛顿法通常只需要两个周期即可收敛。



The final value Ek after each Newton cycle in Algorithm 1 is a measure of the average variability of the residuals. We define inliers as those measurements for which their squared residual is smaller than . Assuming that the residuals are normally distributed, that threshold ensures that 99.7% of the features belong to the same distribution.

算法 1 中每个牛顿循环后的最终值 是残差平均可变性的量度。我们将内点定义为它们的平方残差小于 的那些测量值。假设残差是正态分布的，则该阈值可确保 99。

7% 的特征长于相同的分布。

# Experimental Results

We have implemented the proposed method in C++ using OpenMP technology to benefit from multi-core processing. Our algorithm runs in real time on a standard laptop PC, and we have extensively tested the algorithms on-line in innumerable scenes and situations. We use the KLT algorithm [10] for tracking features. In our configuration, KLT provides up to 1024 tracks with a relatively low computational cost. The stereo disparity of a feature is computed by correlating a window of size 15×15 px centered on the feature position. We use a pyramidal implementation for both, tracking and stereo computation. The baseline of the stereo system is 12.8 cm with a focal length of 654 px and image size of 640 × 480 px.

我们使用 OpenMP 技术在 C++ 中实现了所提出的方法，以从多核处理中受益。我们的算法在标准笔记本电脑上实时运行，我们已经在无数场景和情况下对算法进行了广泛的在线测试。我们使用 KLT 算法 [10] 来跟踪特征。在我们的配置中，KLT 以相对较低的计算成本提供多达 1024 个轨道。通过关联以特征位置为中心的大小为 15×15 px 的窗口来计算特征的立体视差。我们对跟踪和立体计算都使用金字塔实现。**立体声系统的基线为 12.8 厘米，焦距为 654 像素，图像尺寸为 640 × 480 像素**。

”Bridge“ Data Set. More than 2000 images were acquired as the user was walking through the sidewalk of a bridge with a length of approximately 60 meters. Figure 3 shows some excerpts of the sequence. The sequence is challenging because it contains not only repetitive structures, lack of texture, and semitransparencies produced by the railing at the left, but also multiple pedestrians and vehicles moving in both directions. In particular, the middle of the sequence presents a difficult situation for the estimation of the camera’s ego-motion. The images are occupied with up to a 30% of moving objects (see Figure 1 and the second and third row of Figure 3). The robust least squares algorithm presented in Section 4 was still able to provide a correct ego-motion estimate in those situations.

“桥梁”数据集。当用户走过长约 60 米的桥的人行道时，采集了 2000 多张图像。图 3 显示了该序列的一些摘录。序列具有挑战性，因为它不仅包含重复的结构，缺乏纹理和左侧栏杆产生的半透明物体，还有多个行人和车辆双向移动。特别是，序列的中间对于相机的自我运动的估计来说是一个困难的情况。图像被多达 30% 的移动物体占据（参见图 1 和图 3 的第二和第三行）。第 4 节中提出的稳健最小二乘算法仍然能够在这些情况下提供正确的自我运动估计。

Observe that, since we build a local map of the environment, the visual odometry error will grow superlinearly over time [8]. Nevertheless, our robust algorithm is accurate enough to allow the generation of accurate 3D reconstructions of large environments with small drift. In order to demonstrate this, we have performed a reconstruction of the scene by accumulating all observed static 3D points − excluding moving points such as those on pedestrians − into the same reference frame. Figure 4 shows the reconstruction result. It can be seen that the ego-motion algorithm was not only robust throughout the sequence, but also precise enough to produce a coherent spatial perception of the real overall structure.

请注意，由于我们构建了环境的局部地图，因此视觉里程计误差将随着时间的推移超线性增长 [8]。尽管如此，我们强大的算法足够准确，可以生成具有小漂移的大型环境的准确 3D 重建。为了证明这一点，我们通过将所有观察到的静态 3D 点（不包括行人身上的移动点）累积到相同的参考框架中来执行场景的重建。图 4 显示了重建结果。可以看出，自我运动算法不仅在整个序列中是稳健的，而且还足够精确，可以对真实的整体结构产生连贯的空间感知。

**”Footbridge“ Data Set.** A sequence containing 750 images was acquired inside a building. Figure 5a shows a picture of the tested environment. The user started approximately at the camera position of Figure 5a and then turned left to walk on the footbridge. Figures 5b and 5c show the structure obtained by the accumulation of all observed static points of the sequence. As it can be seen from the bird’s eye view of Figure 5b, the estimation was accurate enough to provide an almost perfect planar reconstruction of the lateral footbridge wall. A careful inspection of the ego-path shown in Figure 5c reveals the typical sinusoidal undulation performed when walking.

“人行桥”数据集。在建筑物内采集了包含 750 张图像的序列。图 5a 显示了测试环境的图片。用户大约从图 5a 的相机位置开始，然后左转走在人行天桥上。图 5b 和 5c 显示了由序列的所有观察到的静态点的累积获得的结构。从图 5b 的鸟瞰图可以看出，估计足够准确，可以提供几乎完美的横向人行桥墙的平面重建。仔细检查自我

图 5c 所示的路径显示了行走时典型的正弦波动。