Система робастной визуально-инерциальной одометрии беспилотного автомобиля

Ж. Цай, Ц. Цзинь, А.В. Бобков

Московский государственный технический университет им. Н.Э. Баумана

Аннотация: Данная работа посвящена построению робастной системы визуально-инерциальной одометрии беспилотного автомобиля, использующей бинокулярные камеры и инерциальные датчики в качестве источников информации. Система основана на модифицированной структуре системы VINS-FUSION. Для лучшего баланса количества и качества точек отслеживания предложено использовать два типа особых точек. Для фильтрации неверных совпадений особых точек предложено использовать несколько разных методов. Семантическая и геометрическая информация объединяются для быстрого удаления динамических объектов. Особые точки статических объектов используются для дополнения точек отслеживания. Предлагается многослойный механизм оптимизации для полного использования всех сопоставлений точек и повышения точности оценки движения. Результаты экспериментов демонстрируют эффективность системы.

Ключевые слова: робастная визуально-инерциальная одометрия, локализация, дорожная сцена, многоуровневый механизм оптимизации.

1. Введение

Визуальной одометрией [1] называется определение собственного положения объекта по изображениям, получаемым с закрепленной на нем камеры. Поскольку алгоритмы визуальной одометрии работают довольно медленно, система навигации обычно дополняется инерциальной навигационной системой (ИНС), менее точной, но позволяющей работать с большей скоростью. Такая комплексированная система называется визуально-инерциальной одометрией [1].

Определение собственного положения по видеоизображению, как правило, производится путем сравнения особых точек на последовательности кадров. *Особой точкой* называют точку изображения, которая существенно отличается от соседних точек и положение которой можно однозначно определить на текущем и последующем кадре. Получение достаточно точного сопоставления статических особых точек является важной основой для всех визуальных систем локализации. В качестве особых точек в

визуальной одометрии широко используются различные угловые точки, такие как признаки SIFT, SURF и ORB [2]. Сопоставление двух особых точек основано на простой идее – на вычислении сходства их особенностей. Если сходство превышает заданный порог, то сопоставление считается успешным.

В качестве особенности особой точки можно просто взять некоторую окрестность этой точки, считая, что окрестность мало меняется между соседними кадрами. Однако это далеко не всегда выполняется в реальных задачах. Более надежная особенность – это некоторые характеристики точек окрестности, которые не зависят от освещенности и небольших геометрических искажений. Такие характеристики называют *дескрипторами* особой точки. Чем больше информации содержит дескриптор, тем выше различимость особой точки, но обычно тем меньше особых точек, которые соответствуют требованиям. Для того чтобы уменьшить количество попыток сопоставления между особыми точками, обычно используется априорная информация о движении или другие ограничения, чтобы сузить диапазон поиска точки сопоставления. Например, базовый метод оптического потока предполагает, что яркость пикселей в двух кадрах изображений остается неизменной, а положение пикселей меняется очень мало. Метод оптического потока Лукаса-Канаде [3] предполагает, что пиксели объекта одной и той же модели движения имеют плавно изменяющееся поле оптического потока, а затем выводит, что пиксели в небольшой области изображения имеют приблизительно постоянное двумерное движение. Это значительно улучшает скорость сопоставления плотных или полуплотных пикселей. Однако это сильное ограничение также приводит к низкой надежности метода. Если известно прогнозирование движения камеры между двумя кадрами, диапазон поиска точки сопоставления может быть сужен ограничением перепроекции или ограничением эпиполярной линии [4]. Однако, если ошибка прогноза

движения велика, это легко приведет к быстрому накоплению ошибок в оценке движения и сопоставлении особых точек.

Рис. 1. Отслеживание точек на нестандартной дороге с большим количеством статических объектов. Синяя линия – оптический поток точек SIFT, красная линия – оптический поток точек Shi-Tomasi

Статические структуры с богатыми текстурами в дорожных сценах могут предоставить надежные особые точки для визуальной одометрии. Существует класс статических объектов, которые являются постоянно неподвижными, например, различные линии разметки на дороге, заборы, дорожные знаки и дома по обеим сторонам дороги. Сопоставление особых точек на этих структурах проще, и вероятность неверного сопоставления относительно мала. Другой класс статических структур не всегда остается неподвижным, в основном включая различные транспортные средства и пешеходов на дороге. В жилом районе, показанном на рисунке 1, сложно получить высококачественное сопоставление особых точек на его нестандартных дорогах. Однако большое количество автомобилей, припаркованных по обеим сторонам дороги, может предоставить больше статических особых точек на близком расстоянии. Если точки таких объектов можно эффективно и точно определить как статические, использование этих точек значительно улучшит точность оценки движения одометра. Для нестандартных дорог на автомагистралях или в горных районах, как показано

на рисунке 3, два вышеуказанных типа статических структур встречаются редко, и большинство особых точек, которые можно обнаружить, исходят от поверхности дороги или травы и деревьев на обочине дороги. Однако, хотя эти структуры также, скорее всего, останутся неподвижными, у них отсутствует текстура или они имеют очень похожие текстуры. Поэтому, даже если особые точки обнаружены, для них трудно найти верное сопоставление. В этом случае метод оптического потока можно использовать для оценки сопоставления полуплотных особых точек в статической области изображения, и с некоторой априорной информацией можно отфильтровать неверные сопоставления.

Для одометрии на основе стереокамеры, если во время визуального отслеживания слишком мало сопоставлений точек 3D-2D, сложно оценить положение камеры напрямую с помощью метода *PnP* [5]. Однако количество сопоставлений точек 2D-2D обычно более достаточно. Поэтому необходимо предложить метод, который полностью использует все сопоставления точек для оценки движения камеры.

В работе разрабатывается робастная визуально-инерциальная одометрия (РВИО) для дорожных сцен. Эта система использует структуру проекта VINS-FUSION [6], но в ее модуль обнаружения и отслеживания признаков и модуль оценки положения внесено множество улучшений соответственно. Основными результатами этой работы являются:

1. Ввиду существенных различий в текстурах элементов дорожных сцен, предложено использовать сразу два типа угловых точек и соответствующие методы их сопоставления, чтобы всегда получать достаточное количество отслеживаемых точек;

2. Потенциальные динамические объекты на сцене отделяются от фона на основе семантической информации, а набор сопоставления статических точек дополняется особыми точками на статических объектах;

3. В случае автомобиля, движущегося по прямой, большинство явно неверных сопоставлений особых точек отфильтровываются ограничением оптического потока;

4. Оценка движения камеры оптимизируется шаг за шагом на основе сопоставлений 2D-2D точек и 3D-2D точек для повышения надежности одометра в разных сценах.

2. Методика построения системы РВИО

Структура предлагаемой системы РВИО показана на рисунке 2. Система состоит из трех основных компонентов:

- модуль обнаружения и отслеживания особых точек,
- модуль оценки положения камеры,
- модуль построения локальной карты точек.

Рис. 2. Архитектура предлагаемой системы РВИО

По сравнению с системой VINS-FUSION, в предлагаемой системе внесены существенные изменения в первые два модуля.

2.1 Модуль обнаружения и отслеживания особых точек

Из-за высокой различимости и надежности, угловые точки SIFT являются предпочтительными в качестве особых точек в этой работе. Точки

SIFT обнаруживаются в паре стереоизображений I_{L2} и I_{R2} текущего кадра. Затем все обнаруженные точки SIFT сопоставляются методом полного перебора между левыми изображениями I_{L1} и I_{L2} двух кадров, а также между I_{L2} и I_{R2} текущего кадра. Метод подавления немаксимумов используется для выбора наилучшего сопоставления для каждой точки, и вычисляется отношение ζ между сходствами дескрипторов наилучшего сопоставления и субоптимального сопоставления каждой точки. Чем больше значение ζ, тем надежнее сопоставление точки. Поэтому все сопоставления точек SIFT сортируются от больших к малым в соответствии со значением ζ. Для повышения производительности системы в реальном времени обнаружение и сопоставление точек SIFT реализовано в графическом процессоре.

Все точки SIFT, которые получили сопоставление, можно разделить на два типа. Первый тип точек сопоставления – это особые точки, которые были отслежены в предыдущем кадре, то есть эти точки отслежены как минимум в трех последовательных кадрах изображений. Эти точки называются старыми точками и считаются более надежными. Второй тип точек сопоставления – это новые обнаруженные точки в предыдущем кадре, которые имеют всего два кадра наблюдения. Старые точки могут участвовать в совместной оптимизации положения камеры для нескольких кадров. Однако, когда камера быстро движется, количество отслеживаемых старых точек, отслеженных в текущем изображении, может быть серьезно недостаточным. Поэтому в каждом кадре отслеживание старых и новых точек должно поддерживаться в определенном соотношении.

2.1.1. Равномерное распределение глубины особой точки

Для бинокулярной визуальной одометрии определенное количество ориентиров с известными глобальными 3D-координатами должно быть включено в набор особых точек, отслеживаемых в каждом кадре. Эти точки

необходимы для определения абсолютного масштаба оценки смещения камеры, в противном случае система выродится в монокулярную визуальную одометрию в этом кадре. Близкие ориентиры, как правило, имеют большие параллаксы на изображениях и меньше подвержены шуму визуального сопоставления, поэтому они более важны для повышения точности оценки смещения камеры. Сопоставление методом полного перебора подходит для таких точек с большим параллаксом. Далекие ориентиры легко находят сопоставления из-за их малого параллакса, но сильнее подвержены шуму. Поскольку количество сопоставления близких точек слишком мало в некоторых сценах, добавление определенного количества сопоставления дальних точек позволяет использовать метод RANSAC для оценки сущностной матрицы, что важно для оценки вращения камеры. Как правило, для точек стереосопоставления, если их значение глубины превышает базовую длину стереокамеры более чем в 40 раз, стереосопоставление отменяется. Глубина этих дальних точек восстанавливается триангуляцией после получения оценки движения камеры.

Очевидно, что средняя глубина точек в области сверху вниз на изображении примерно постепенно уменьшается. Поэтому изображение делится на 6*6 подблоков, и максимальное количество особых точек, сохраняемых в каждом блоке, ограничено. Поскольку глубина точек в нижней половине изображения часто меньше, порог количества точек в этих блоках увеличивается. Таким образом, значения глубины полученных особых точек распределяются более равномерно.

2.1.2. Добавление точек в сценах с низким содержанием текстур

Когда сцена содержит мало эффективных текстур признаков, количество правильно сопоставляемых точек SIFT очень мало, что может привести к сбою отслеживания или деградации системы в монокулярную

систему. Поэтому, когда количество сопоставлений точек SIFT недостаточно, дополнительно обнаруживается определенное количество точек Shi-Tomasi [7], и затем сопоставляется с использованием метода оптического потока Лукаса-Канаде. Именно такой метод используется в системе VINS-FUSION. Аналогично, модель движения с постоянной скоростью используется для прогнозирования положения точек сопоставления, что может в определенной степени уменьшить ошибку сопоставления, вызванную большим смещением точек пикселей при предположении метода оптического потока. Для пары точек сопоставления вычисляется нормализованная кросс-корреляция их локальных окон в двух кадрах. Чем больше значение корреляции, тем выше качество сопоставления. Все сопоставления точек Shi-Tomasi сортируются от больших к малым в соответствии со значением корреляции. Механизм распределения областей для этих точек Shi-Tomasi согласуется с механизмом точек SIFT, так что точки отслеживания во всем изображении распределяются более равномерно, и гарантируется определенное количество точек соответствия со значениями глубины.

2.1.3. Фильтрация сопоставления точек при движении камеры только в прямом направлении

Большое количество сопоставлений особых точек часто можно получить в структурах с богатыми текстурами, однако если похожие текстуры часто появляются в сцене, вероятно возникновение несовпадений. Поэтому, начиная с третьего кадра системы, оценивается движение камеры от предыдущего кадра к текущему кадру на основе предположения о движении с постоянной

скоростью. Это прогноз движения используется для прогнозирования положения отслеживаемой точки на текущем изображении. Если расстояние между точкой сопоставления и предсказанной точкой превышает пороговое

значение, сопоставление считается неверным. Однако в случае большого вращения камеры ошибка модели постоянного движения будет велика, поэтому такая операция фильтрации сопоставления точек выполняется только тогда, когда камера не вращается или вращается незначительно.

Для дорожных сцен большая часть точек с меньшей глубиной обычно являются точками Shi-Tomasi на поверхности дороги. Однако среди них есть много очевидных неверных сопоставлений, то есть направление оптического потока точки серьезно отклоняется от верного направления, как показано на рисунке 3(а). Это значительно снизит точность оценки движения камеры для сцен с меньшим количеством особых точек, таких как автомагистрали. Легко доказать, что если камера движется только вперед или назад между двумя кадрами, и почти нет смещения в других направлениях или вращения, то прямые линии, на которых расположены оптические потоки пикселей, приблизительно пересекаются в центре проекции изображения. Если оговорено, что камера движется только вперед, положение (u_2, v_2) особой точки в текущем кадре должно быть дальше от центральной точки изображения P_c , чем ее положение (u_1, v_1) в предыдущем кадре. Предположим, что точки в соответствующих позициях в изображении предыдущего кадра – это $P_1(u_1, v_1)$ и $P_2(u_2, v_2)$ соответственно. Когда прогнозируемое движение камеры между двумя кадрами близко к прямому перемещению, если пара точек сопоставления не удовлетворяет условию, показанному в уравнении (1), это считается неверным сопоставлением:

$$dist(P_C, L_{12}) < \varepsilon \ \wedge \ V_{P_1P_2} * V_{P_1P_2} < 0, \tag{1}$$

где L_{12} - линия оптического потока, образованную точками P_1 и P_2 ; $dist(P_C, L_{12})$ – расстояние от точки P_C до линии L_{12} ; ε – пороговое расстояние;

$V_{P_1P_2}$ – вектор из точки P_1 в точку P_2; * – скалярное произведение векторов.

Если доля и количество точек, соответствующих уравнению (1) среди всех текущих точек сопоставления больше пороговых значений, считается, что камера движется примерно по прямой линии, а точки сопоставления, не соответствующие уравнению (1), удаляются. Поскольку движение по прямой является наиболее распространенным режимом движения автомобиля, этот метод позволяет отфильтровать большинство явно неверных сопоставлений точек, как показано на рисунке 3(б). В то же время это повысит точность последующей оценки движения и сократит затраты времени.

a)

b)

Рис.3. Отслеживание точек на шоссе: а) оптический поток особых точек до фильтрации; б) оптический поток особых точек после фильтрации

2.1.4. Добавление особых точек статических объектов

Если в сцене есть динамические объекты, такие как транспортные средства или пешеходы, сопоставления динамических точек на них не

должны использоваться для оценки движения камеры. Чтобы быстро отличить область фона на изображении от области динамических объектов, в этой работе используется модель глубокого обучения для сегментации всех объектов указанной категории. Сначала исключается область маски всех объектов, и только в области фона получаются статические особые точки.

После обработки вышеуказанных шагов (2.1.1 – 2.1.3) в области фона текущего изображения может оказаться слишком мало близких статических точек. Такая ситуация легко возникает на нестандартизированных дорогах. Если на изображении обнаружен потенциальный динамический объект, можно рассмотреть возможность получения точек отслеживания из статических объектов на нем. Существует множество методов, которые можно использовать для проверки того, является ли точка объекта динамической [8 - 9]. В этой работе, поскольку используется стереокамера, а особые точки на объекте часто более заметны и сконцентрированы, проще получить стереосоответствие точки объекта. Поскольку сопоставления точек SIFT используют метод полного перебора, нет необходимости полагаться на прогнозируемое положение точки сопоставления. Если на объекте обнаружена точка SIFT со стереосоответствием, среднюю глубину этих точек можно использовать в качестве прогнозируемого значения глубины всех точек Shi-Tomasi объекта. Значение прогнозирования глубины можно использовать для получения прогноза положения точки стереосоответствия. Это может повысить скорость и точность оптического сопоставления потока. Для точек объектов только те точки со стереосоответствием и значениями глубины меньше указанного порогового значения считаются потенциальными статическими точками. Ошибка репроекции каждой точки объекта предыдущего кадра в текущем изображении рассчитывается с использованием прогнозируемого движения камеры. Если значение ошибки точки больше порогового значения, точка считается статической и

добавляется в набор точек отслеживания фона. Аналогично, поскольку модель движения с постоянной скоростью будет иметь большие ошибки, когда камера совершает большие повороты, статические точки объекта проверяются и добавляются только тогда, когда прогнозируемый поворот камеры невелик.

2.2. Модуль оценки положения камеры

Подобно системе VINS-FUSION, эта система обеспечивает две конфигурации системы: бинокулярную камеру и бинокулярную камеру + ИНС. Измерения ИНС может напрямую предоставить начальную оценку движения камеры, тем самым упрощая весь процесс оценки положения камеры.

2.2.1. Система только с бинокулярной камерой

В случае чистого зрения три задачи оптимизации используются для последовательного уточнения оценки положения камеры.

2.2.1.1. Оценка движения на основе сопоставления 2D-2D точек

После завершения обнаружения, сопоставления, сортировки, распределения и предварительной фильтрации особых точек получается набор точек отслеживания F. Иногда количество точек с известной глубиной в наборе F слишком мало для получения точной оценки движения камеры напрямую путем решения задачи PnP. Поэтому в каждом кадре сначала получается начальная оценка движения камеры с использованием сопоставления 2D-2D точек. Для наборов точек сопоставления SIFT и Shi-Tomasi первые 2/3 и первая 1/3 точек выбираются для присоединения к подмножеству W соответственно. Подобно системе ORB-SLAM3 [10], набор W и метод $RANSAC$ используются для одновременной оценки сущностной матрицы $\boldsymbol{E}$ движения камеры или матрицы гомографии $\boldsymbol{H}$, связанной с

плоскостью дороги. Чтобы уменьшить влияние неверных сопоставлений, используется сложный пятиточечный метод для оценки матрицы $\boldsymbol{E}$ [11]. Набор F используется для проверки количества эффективного точки для каждого результата оценки. Наконец, если сумма ошибок эпиполярного ограничения всех эффективных точек меньше суммы ошибок ограничения гомографии, матрица E сохраняется. В противном случае сохраняется матрица $\boldsymbol{H}$.

Предположим, что текущее изображение – это изображение k-того кадра (k>0). Первоначальная оценка вращения камеры $\boldsymbol{R}_{C_{k-1}}^{C_k}$ и направления смещения $\overline{\boldsymbol{p}}_{C_{k-1}}^{C_k}$ получаются путем разложения матрицы $\boldsymbol{E}$ или $\boldsymbol{H}$. Точки выброса, не удовлетворяющие ограничениям эпиполярной линии $\boldsymbol{E}$ или ограничениям гомографии $\boldsymbol{H}$, будут удалены напрямую. Оставшиеся точки могут по-прежнему иметь неверные сопоставления, что не только не повысит точность последующей оптимизации положения камеры, но и увеличит время расчета. Если количество оставшихся точек сопоставления слишком велико, точки с плохим качеством сопоставления удаляются, чтобы окончательное количество точек не превышало пороговое значение.

2.2.1.2. Оптимизация оценки положения камеры на основе сопоставления 3D-2D точек

После получения направления смещения камеры $\overline{\boldsymbol{p}}_{C_{k-1}}^{C_k}$ используется нелинейная оптимизация для оценки масштаба s_p смещения. Все особые точки отслеживания со значениями глубины в предыдущем кадре выбираются в качестве набора точек S_d. Согласно уравнению проекции камеры, задача нелинейной оптимизации относительно s_p описывается следующим образом:

$$\min_{s_p} \sum_{l \in C_{k-1}} \rho\left(\left\|\boldsymbol{r}_{proj}\left(\hat{z}^{C_k}, s_p\right)\right\|_{p_l}^2\right), \tag{2}$$

$$p(t) = \begin{cases} t, & t \geq 1 \\ 2\sqrt{t-1}, & 0 \leq t < 1 \end{cases}, \tag{3}$$

$$\boldsymbol{r}_{proj}\left(\hat{z}_l, s_p\right) = \left(\frac{\boldsymbol{O}_l^{C_k}}{\left|(0{,}01) \cdot \boldsymbol{O}_l^{C_k}\right|} - \widehat{\boldsymbol{O}}_l^{C_k}\right)_{xy} = \left(\frac{\boldsymbol{O}_l^{C_k}}{dep_l^{C_k}} - \widehat{\boldsymbol{O}}_l^{C_k}\right)_{xy}, \tag{4}$$

$$\widehat{\boldsymbol{O}}_l^{C_k} = \pi_C^{-1}\left(\begin{bmatrix} \hat{u}_l^{C_k} \\ \hat{v}_l^{C_k} \end{bmatrix}\right), \tag{5}$$

$$\boldsymbol{O}_l^{C_k} = \boldsymbol{R}_{C_{k-1}}^{C_k} \frac{1}{d_l^{C_{k-1}}} \pi_C^{-1}\left(\begin{bmatrix} u_l^{C_{k-1}} \\ v_l^{C_{k-1}} \end{bmatrix}\right) + s_p \overline{\boldsymbol{p}}_{C_{k-1}}^{C_k}, \tag{6}$$

где $\left[u_l^{C_{k-1}}, v_l^{C_{k-1}}\right]^T$ – вектор наблюдения l-той трехмерной особой точки на k-том изображении; $d_l^{C_{k-1}}$ – обратная величина глубины l-той трехмерной особой точки в системе координат камеры предыдущего кадра. $dep_l^{C_k}$ – значение глубины, полученному путем перепроецирования l-той особой точки из системы координат камеры предыдущего кадра в систему координат камеры текущего кадра. Использование метода Гаусса-Ньютона для решения уравнения (2) позволяет получить оценки масштаба s_p и абсолютного смещения камеры $\boldsymbol{p}_{C_{k-1}}^{C_k} = s_p * \overline{\boldsymbol{p}}_{C_{k-1}}^{C_k}$. В этой точке получается начальная оценка движения камеры на основе точек отслеживания, и выводится глобальное положение камеры в текущем кадре. После этого, наблюдения всех эффективных точек отслеживания добавляются на локальную карту.

Направление смещения $\overline{\boldsymbol{p}}_{C_{k-1}}^{C_k}$ камеры, оцененное с помощью сопоставления 2D-2D точек, не всегда является точным, и уравнение (2) не может оптимизировать направление смещения. Поэтому далее положения камеры во всех прошлых кадрах и глубина всех точек карты в скользящем окне фиксируются, а положение камеры текущего кадра оптимизируется с

помощью сопоставления 3D-2D точек. Поскольку значение глубины, полученное с помощью сопоставления стерео, обычно более точно, следует отдавать предпочтение трехмерным наблюдениям близких особых точек со сопоставлением стерео. Для любой точки карты P_m, отслеживаемой в текущем кадре F_c, ее наблюдения в прошлых кадрах просматриваются во временном порядке от ближнего к дальнему до тех пор, пока не будет найдено стереосоответствие в определенном кадре F_p. Добавляется 3D-2D сопоставление этой точки из кадра F_p и кадра F_c к набору S_0. Если количество сопоставлений в S_0 меньше порогового значения, точки карты просматриваются еще раз. Если точка в определенном прошлом кадре имеет значение глубины меньше порогового значения φ, то добавляется её 3D-2D сопоставление к набору S_0. Наконец, используются методы RANSAC и PnP для оптимизации положения камеры [12]. Поскольку значение оценки движения, полученное путем сопоставления 2D-2D точек, уже близко к правильному значению, задача оптимизации здесь будет выполнена быстро.

2.2.1.3. Сглаживание траектории камеры на основе совместной оптимизации

Вышеуказанный процесс оценки касается только положения камеры текущего кадра. Однако траектория камеры, полученная таким образом, вероятно, не будет гладкой. Поэтому в скользящем окне для всех точек карты, чьи кадры наблюдения не меньше порогового значения, оптимизируются совместно с глобальными положениями камеры всех кадров в скользящем окне. Совместная оптимизация выражается следующими формулами:

$$\min_{X} \sum_{(l,j)\in C} \rho\left(\left\| r_C\left(\hat{\boldsymbol{z}}_l^{C_k}, \boldsymbol{\chi}\right)\right\|^2_{\boldsymbol{P}_l^{c_j}}\right), \tag{7}$$

$$r_C(\hat{\boldsymbol{z}}_l^{C_k}, \boldsymbol{\chi}) = \left(\frac{\boldsymbol{O}_l^{C_j}}{\left|(0{,}01) \cdot \boldsymbol{O}_l^{C_j}\right|} - \widehat{\boldsymbol{O}}_l^{C_j} \right)_{xy} = \left(\frac{\boldsymbol{O}_l^{C_j}}{dep_l^{C_j}} - \widehat{\boldsymbol{O}}_l^{C_j} \right)_{xy}, \quad (8)$$

$$\widehat{\boldsymbol{O}}_l^{C_j} = \pi_C^{-1}\left(\begin{bmatrix} \hat{u}_l^{C_j} \\ \hat{v}_l^{C_j} \end{bmatrix} \right), \quad (9)$$

$$\boldsymbol{O}_l^{C_j} = \boldsymbol{R}_w^{C_j}\left(\boldsymbol{R}_{C_i}^{w} \frac{1}{d_l} \pi_C^{-1}\left(\begin{bmatrix} u_l^{C_i} \\ v_l^{C_i} \end{bmatrix} \right) + \boldsymbol{p}_{C_j}^{w} \right) - \boldsymbol{p}_w^{C_j}, \quad (10)$$

$$\boldsymbol{\chi} = [\boldsymbol{x_0}, \boldsymbol{x_1}, \dots, \boldsymbol{x_{N-1}}, d_0, d_1, \dots, d_{M-1}]^T, \quad (11)$$

$$\boldsymbol{x}_j = \left[\boldsymbol{p}_{C_j}^{w}, \boldsymbol{q}_{C_j}^{w} \right], \ j = 0 \dots N-1, \quad (12)$$

где $\boldsymbol{\chi}$ – полный вектор переменной состояния в скользящем окне; $\boldsymbol{x}_j$ – переменные состояния камеры на момент времени захвата j-го изображения, которые содержат положение и ориентацию левой камеры в мировой системе координат; N – общее количество ключевых кадров изображений; M – общее количество особых точек в скользящем окне; d_l – обратная величина глубины l-ой трехмерной особой точки в системе координат камеры, в которой она впервые наблюдалась; $\left[u_l^{C_i} \quad v_l^{C_i}\right]^T$ и $\left[\hat{u}_l^{C_j} \quad \hat{v}_l^{C_j}\right]^T$ – наблюдение l-той трехмерной особой точки на i-ом и j-ом кадре изображении; i-й кадр - это кадр, в котором точка впервые наблюдается в текущем скользящем окне; $\pi_C^{-1}(\cdot)$ – функция обратной проекции, которая преобразует местоположение пикселя в его 2D-координату в плоскости нормализации, используя внутренние параметры камеры; $dep_l^{C_j}$ – значение глубины, полученному путем перепроецирования l-той особой точки из системы координат камеры i-того кадра в систему координат камеры j-того кадра. Операция $(\cdot)_{xy}$ – взятие первых двух измерений трехмерного вектора; $\boldsymbol{P}_l^{C_j}$ - стандартная ковариационная матрица при фиксированном значении ошибки обнаружения положения особой точки в плоскости нормализации.

Задача оптимизации, показанная в формуле (7), также решается с использованием метода Гаусса-Ньютона. Для обеспечения производительности системы в реальном времени совместная задача оптимизации используется только для ключевых кадров. Однако для визуальной одометрии, используемой в беспилотных автомобилях, камера часто движется с высокой скоростью в течение длительного времени, и в этом случае каждый кадр является ключевым. Поэтому длина скользящего окна не должна быть слишком большой, в противном случае время совместной оптимизации будет слишком большим. Кроме того, система VINS-FUSION маргинализирует соответствующие переменные для формирования априорных остатков каждый раз, когда отбрасывается самый старый кадр скользящего окна. Однако в ходе экспериментов в данной работе было установлено, что эта априорная остатков не приводит к существенному улучшению точности результатов одометра, но при этом требует много вычислительного времени. Поэтому в данной работе данная операция маргинализации не используется, а соответствующие переменные напрямую отбрасываются. После совместной оптимизации точки карты, общая ошибка перепроецирования которых превышает пороговое значение, удаляются, чтобы сохранить легкость локальной карты.

2.2.2. Система с бинокулярной камерой и ИНС

В данной работе метод оценки положения и ориентации, основанный на тесной связи инерциальной навигационной системы (ИНС) и бинокулярной камеры, ссылается на теорию, изложенную в работе [6]. Поскольку автомобиль может уже находиться в движении, когда одометр начинает работать, аналогично VINS-FUSION, в данной системе для инициализации ИНС используются результаты визуального отслеживания всех кадров в пределах первого скользящего окна. В первом скользящем окне

рабочий процесс системы такой же, как и при использовании только бинокулярной камерой. После инициализации системы ВИО интегрированные измерения ИНС можно использовать в качестве значения прогноза движения камеры в каждом новом кадре. Если результат оценки на этапе инициализации достаточно хорошо, точность этого прогнозирования движения намного выше, чем значение прогноза, основанное на движении камеры с постоянной скоростью. При использовании этого прогнозирования движения для расчета значений прогноза положения точек сопоставления нет необходимости исключать случай больших поворотов камеры. Кроме того, после инициализации система ВИО больше не нужно использовать метод PnP для оптимизации положения камеры на основе сопоставления 3D-2D точек.

Для совместной оптимизации задача оптимизации, показанная в уравнении (7) – (12), частично модифицируется следующим образом:

$$\min_{\chi}\left\{\sum_{k\in B}\left(\left\|\boldsymbol{r}_B\left(\hat{\boldsymbol{z}}_{b_{k+1}}^{b_k},\boldsymbol{\chi}\right)\right\|_{\boldsymbol{P}_{b_{k+1}}^{b_k}}^2\right)+\sum_{(l,j)\in C}\left(\rho\left\|\boldsymbol{r}_C\left(\hat{\boldsymbol{z}}_l^{c_j},\boldsymbol{\chi}\right)\right\|_{\boldsymbol{P}_l^{c_j}}^2\right)\right\}, \tag{13}$$

$$\boldsymbol{O}_l^{C_j}=\boldsymbol{R}_B^C\left(\boldsymbol{R}_w^{b_j}\left(\boldsymbol{R}_{b_i}^w\left(\boldsymbol{R}_C^B\frac{1}{d_l}\pi_C^{-1}\left(\begin{bmatrix}u_l^{C_i}\\ v_l^{C_i}\end{bmatrix}\right)+\boldsymbol{p}_c^b\right)+\boldsymbol{p}_{b_i}^w-\boldsymbol{p}_{b_j}^w\right)-\boldsymbol{p}_c^b\right), \tag{14}$$

$$\begin{gathered}\chi=[\boldsymbol{x_0},\boldsymbol{x_1},\ldots,\boldsymbol{x_{N-1}},\boldsymbol{x}_c^b,\ d_0,d_1,\ldots,d_{M-1}]^T,\\ \boldsymbol{x}_k=\left[\boldsymbol{p}_{b_k}^w,\boldsymbol{v}_{b_k}^w,\boldsymbol{q}_{b_k}^w,\boldsymbol{b}_{a_k},\boldsymbol{b}_{\omega_k}\right],\quad k=0\ldots N-1,\\ x_c^b=\left[\boldsymbol{p}_{C_j}^w,\boldsymbol{q}_{C_j}^w\right],\ j=0\ldots N-1\end{gathered} \tag{15}$$

где $\boldsymbol{x}_k$ – переменные состояние ИНС на момент времени захвата k-го изображения, которые содержат положение, скорость и ориентацию ИНС в мировой системе координат, а также смещение ускорения и смещение гироскопа в системе координат тела; $\boldsymbol{r}_B(\cdot)$ – остатки измерений ИНС; их более подробный расчет можно найти в работе [6]. В совместной оптимизации будут участвовать только точки карты, последовательные наблюдаемые кадры которых больше указанного значения (3 кадра в этой

работе). Если количество точек отслеживания, соответствующих требованиям в текущем кадре, меньше порогового значения, этих точек недостаточно для предоставления ограничений для всех переменных текущего кадра, и отменяется совместная оптимизация в текущем кадре.

3. Эксперименты и анализ результатов

Для проверки эффективности предлагаемой системы РВИО используется открытый набор данных KITTI [13]. В наборе аннотировано шесть категорий объектов, но в этой работе отслеживаются только твердые тела. Разрешение стереоизображения после стереокоррекции составляет примерно 376*1240. В наборе данных KITTI-Odometry содержит в общей сложности 11 последовательностей с достоверной информацией о положениях камеры. Получаются все измерения ИНС из исходных данных соответствующей последовательности. Бинокулярная камера работает на частоте 10 Гц, а ИНС – на частоте 100 Гц. Длина базовой линии стереокамеры составляет 0,54 м, поэтому порог глубины близкой точки установлен на уровне 21 м. С другой стороны, максимальная эффективная глубина точки составляет 40 м, исходя из разрешения изображения. Для сегментации экземпляров на изображении была выбрана модель YOLO-v8 [14]. Эта модель может одновременно получать ограничивающую рамку и маску сегментации объекта.

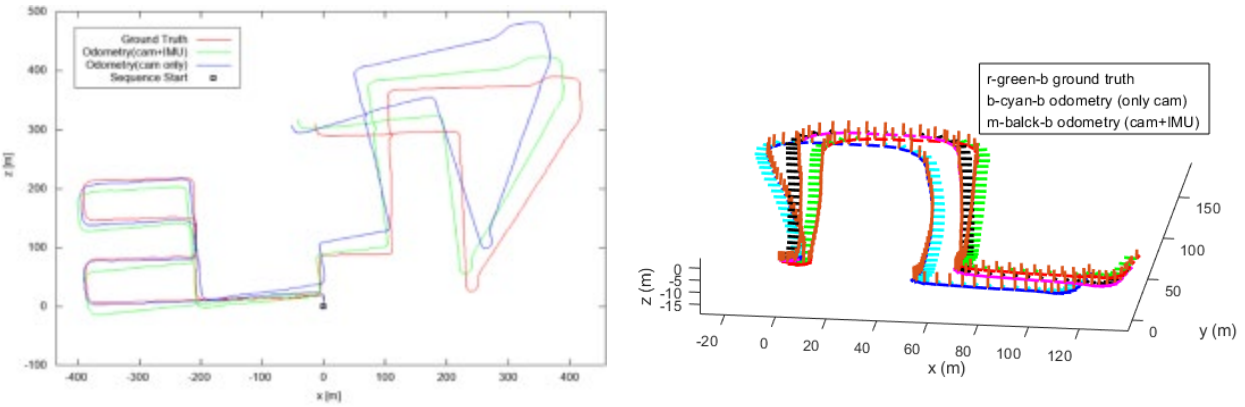
Таблица № 1

Ошибки оценки движения в наборе данных KITTI-Odometry

	Предлагаемый РВИО				VINS-FUSION			
	РВИО (без ИНС)		РВИО + ИНС		VINS-F (без ИНС)		VINS-F + с ИНС	
видео	E_r, м	E_t, %	E_r, м	E_t, %	E_r, м	E_t, %	E_r, м	E_t, %
0000	0.367	0.792	**0.255**	0.471	0.555	1.144	0.376	0.703
0001	**0.516**	**2.849**	0.703	4.493	0.658	5.482	0.781	7.19
0002	0.406	0.775	**0.273**	0.671	0.464	1.001	0.343	0.712
0003	0.227	1.049	**0.169**	0.785	0.273	1.727	0.24	1.146

0004	0.356	0.762	**0.295**	0.542	0.431	1.045	0.349	0.915
0005	0.418	0.535	**0.317**	0.466	0.604	0.877	0.454	0.577
0006	0.331	0.74	**0.242**	0.583	0.486	1.151	0.316	0.859
0007	0.489	3.436	**0.275**	1.618	0.516	4.706	0.382	3.261
0008	0.529	3.541	**0.367**	1.995	0.738	5.075	0.475	3.943
0009	0.562	1.319	**0.296**	0.917	0.759	1.993	0.413	1.454
0010	0.608	2.224	**0.381**	1.532	0.645	3.145	0.449	2,398

Для каждой последовательности вычисляются ошибки смещения и вращения всех возможных подпоследовательностей длиной 100...800 метров. Ошибка смещения E_t измеряется в процентах, и ошибка вращения E_r измеряется в градусах на метр. Мы сравниваем наши результаты с результатами системы VINS-FUSION [6]. Поскольку РВИО является системой одометрии, то для корректности экспериментов функция замыкания петли системы VINS-FUSION в эксперименте отключена. Результаты показаны в таблице 1. По сравнению с системой VINS-FUSION данная система повысила точность локализации в обеих конфигурациях системы. Среди них особенно очевидно улучшение точности оценки смещения камеры. Это связано с тем, что различные методы фильтрации сопоставления особых точек в этой работе в основном используются в случае прямого движения при чистой визуальной конфигурации. Что еще более важно, метод системы VINS-FUSION для оценки глубины особых точек очень ненадежен. РВИО тщательно и строго управлял механизмом распределения и экранирования глубины особой точки для получения более надежных близких точек. Используя сопоставления два типа угловых точек и механизмов многослойной оптимизации, РВИО также продемонстрировал свои преимущества в оценке вращения.



а) б)

Рис. 4. - Пример результатов оценки траектории вместе с истинными данными для двух последовательностей в наборе KITTI: а) вид с верху (видео 0008 KITTI-Odometry); б) трехмерный вид (видео 0007 в KITTI-Tracking)

В большинстве случаев использование ИНС улучшит точность оценки движения. Противоположная ситуация произошла в последовательности 0001. Это связано с тем, что последовательность представляет собой сцену шоссе, камера начинает с большого поворота, и в сцене наблюдается крайняя нехватка близких особых точек, что делает результат инициализации ИНСа очень плохим. В реальных условиях одометр обычно калибруется в начале и начинает работать из стационарного состояния. Результаты оценки траекторий двух типичных последовательностей с использованием РВИО показаны на рисунке 4. Видно, что использование ИНС значительно повышает точность локализации. Фактически, результаты калибровки бинокулярной камеры, предоставленные набором данных KITTI, не очень точны. Если внутренние и внешние параметры камеры можно будет откалибровать точнее [15] или оптимизировать онлайн, производительность РВИО будет еще больше улучшена.

Среднее время выполнения на кадр двух систем показано в таблице 2. Обе системы запускают 6 потоков во время совместной оптимизации. Время

выполнения YOLOv-8 в графическом процессоре составляет около 1,5 мс, а время постобработки – около 2 мс. Обнаружение, сопоставление и предварительная фильтрация занимают около 4,5 мс. Поскольку при использовании ИНС для лучшего объединения измерений с ИНС и камеры требуется большая длина окна, количество точек, участвующих в совместной оптимизации, больше, что требует больше времени. Поскольку используется больше стратегий улучшения, рабочая частота системы РВИО ниже, чем у VINS-FUSION, но она все еще близка к работе в реальном времени.

Таблица № 2

Среднее время работы систем РВИО

	Предлагаемый РВИО		VINS-FUSION	
Время (мс/кадр)	62	74	46	55

Также были провели эксперименты с использованием системы ORB-SLAM3. Поскольку ORB-SLAM3 сложно завершить инициализацию ИНС в дорожных сценах, она использует только конфигурацию стереокамеры. Функция замыкания петли была также отключена. Средняя ошибка результатов двух систем во всех последовательностях вычисляется, как показано в таблице 3. При использовании только стереокамеры точность оценки ORB-SLAM3 в целом выше, чем у РВИО. Это связано с тем, что система ORB-SLAM3 использует только особые точки с дескрипторами и методы сопоставления на основе поиска, поэтому ее точность сопоставления точек в целом, выше. Кроме того, особые точки ORB более универсальны, чем точки SIFT, поэтому число обнаружений точки ORB больше, чем у SIFT.

Таблица № 3

Средняя ошибки оценки движения в наборе данных KITTI-Odometry

Предлагаемый РВИО		ORB-SLAM3
без ИНС	с ИНС	без ИНС

$\bar{E}_r$, м	$\bar{E}_t$,%	$\bar{E}_r$, м	$\bar{E}_t$,%	$\bar{E}_r$, м	$\bar{E}_t$,%
0.437	1.638	0.325	1.269	0.307	1.384

Таблица № 4

Ошибки оценки движения в наборе данных KITTI-Tracking

	Предлагаемый РВИО				ORB-SLAM3	
	без ИНС		с ИНС		без ИНС	
видео	E_r, м	E_t, %	E_r, м	E_t, %	E_r, м	E_t, %
0004	0.302	0.815	**0.186**	0.476	0.293	0.749
0008	**0.431**	**2.317**	0.605	3.139	0.682	4.537
0018	0.447	1.934	**0.316**	1.215	0.423	2.198
0019	0.341	0.787	**0.197**	0.395	0.229	**0.370**
0020	0.425	1.181	**0.329**	**0.873**	0.634	2.294

Для проверки работоспособности системы в сценах с большим количеством динамических объектов были также проведены эксперименты с использованием набора данных для обучения в KITTI-Tracking. Для эксперимента выбираются 5 последовательности из этого набора данных, содержащие более динамические объекты. Поскольку в KITTI-Tracking не указаны истинные данные о положениях камеры, для расчета истинных данных используются данные GPS в исходном наборе данных соответствующей последовательности. Результаты экспериментов показаны в таблице 4. В этих сценариях система РВИО работает значительно лучше, чем ORB-SLAM3, поскольку она напрямую исключает большинство динамических объектов из источника сопоставления точек. ORB-SLAM3 полагается на метод RANSAC для исключения выбросов сопоставлений точек. Однако большое количество точек отслеживания динамических объектов как выбросов серьезно повлияет на оценку движения камеры. Этот

эффект более очевиден в сценах шоссе, таких как последовательности 0008 и 0020, поскольку высокая доля выбросов сделает метод RANSAC неэффективным.

Заключение

В данной работе предлагается робастная визуально-инерциальная одометрия (РВИО) в режиме реального времени для беспилотного автомобиля. Ввиду ситуаций на дорогах с редкими особыми точками в фоне и множеством динамических объектов предлагается несколько методов, чтобы одометрия достигала хорошего баланса между количеством и качеством отслеживания особых точек и полностью использовала сопоставления особых точек для лучшей оценки положения автомобиля. Система РВИО основана на структуре системы VINS-FUSION, но значительно лучше, чем VINS-FUSION в экспериментах. В высокодинамичных и крайне бедных особенностями сценах система РВИО также работает лучше, чем система ORB-SLAM3. Поэтому предлагаемые методы улучшения могут быть использован для улучшения любой существующей визуально-инерциальной одометрии. Дальнейшая работа будет сосредоточена на одновременной локализации автомобиля и 3D-отслеживания динамических объектов на сцене на основе этой одометрии.

Литература

1. Lim K.L., Braunl T. A review of visual odometry methods and its applications for autonomous driving. arXiv preprint arXiv:2009.09193, 2020. URL: doi.org/10.48550/arXiv.2009.09193.

2. Bansal M., Kumar M. & Kumar, M. 2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors. Multimed Tools Appl, 2021. pp. 18839–18857.

3. Lucas, B. D., Kanade, T. An iterative image registration technique with an application to stereo vision // IJCAI'81: 7th international joint conference on Artificial intelligence. 1981. Vol. 2. pp. 674-679.

4. Zhang Z. Determining the Epipolar Geometry and its Uncertainty: A Review // International Journal of Computer Vision 27, 1998, P. 161–195. URL: doi.org/10.1023/A:1007941100561.

5. Xiao Xin Lu. A Review of Solutions for Perspective-n-Point Problem in Camera Pose Estimation // Journal of Physics: Conference Series, 2018. volume 1087. 052009 p.

6. Qin T., Cao S., Pan J., and Shen S. A general optimization-based framework for global pose estimation with multiple sensors // arXiv preprint arXiv:1901.03642. 2019. URL: doi.org/10.48550/arXiv.1901.03642.

7. Shi J., Tomasi. Good features to track // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, 1994. pp. 593-600.

8. Цай Ж., Бобков А.В. Обзор методов решения задачи одновременной локализации и многообъектного 3D-отслеживания с использованием технического зрения // Автоматизация. Современные технологии. 2024. № 3. С. 105-118.

9. Суханов А.В., Артемьев И.С., Долгий А.И., Хатламаджиян, А.Е. Метод оптической идентификации железнодорожных подвижных единиц на основе интегральных устойчивых признаков // Инженерный вестник Дона, 2013. №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2217.

10. Campos C., Elvira R., Rodriguez J. J. G., Montiel, J. M., Tardos, J. D. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam // IEEE transactions on robotics. 2021. 37(6). pp. 1874-1890.

11. Nister D. An efficient solution to the five-point relative pose problem // EEE-T-PAMI. 2004. vol. 26, No. 6. pp. 756-770.

12. Martinez-Otzeta JM, Rodriguez-Moreno I, Mendialdua I, Sierra B. RANSAC for Robotic Applications: A Survey // Sensors. 2023. No. 1. 327 p.

13. Geiger A., Lenz P., Stiller C., and Urtasun R. Vision meets robotics: The KITTI dataset. // The International Journal of Robotics Research. 2013, Res. 32. P. 1231–1237. DOI:10.1177/0278364913491297.

14. Terven J., Cordova-Esparza D.M., Romero-Gonzalez, J.A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS // Machine Learning and Knowledge Extraction 5. 2023. No. 4. pp. 1680-1716.

15. Толкачев Д.С. Повышение точности калибровки внешних параметров видеокамеры // Инженерный вестник Дона, 2013, №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1840.

References

1. Lim K.L., Braunl T. arXiv preprint arXiv:2009.09193, 2020. URL: doi.org/10.48550/arXiv.2009.09193.

2. Bansal M., Kumar M. & Kumar, M. Multimed Tools Appl, 2021. pp. 18839–18857.

3. Lucas, B. D., Kanade, T. IJCAI’81: 7th international joint conference on Artificial intelligence. 1981. Vol. 2. pp. 674-679.

4. Zhang Z. International Journal of Computer Vision 27, 1998, pp. 161–195. URL: doi.org/10.1023/A:1007941100561.

5. Xiao Xin Lu. Journal of Physics: Conference Series, 2018. volume 1087. 052009 p.

6. Qin T., Cao S., Pan J., and Shen S. arXiv preprint arXiv:1901.03642. 2019. URL: doi.org/10.48550/arXiv.1901.03642.

7. Shi J., Tomasi. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, 1994. pp. 593-600.

8. Czaj Zh., Bobkov A.V. Avtomatizaciya. Sovremennye texnologii. 2024. N. 3. pp. 105-118.

9. Suxanov A.V., Artemev I.S., Dolgij A.I., Xatlamadzhiyan, A.E. Inzhenernyj vestnik Dona. 2013. N. 4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2217.

10. Campos C., Elvira R., Rodriguez J. J. G., Montiel, J. M., Tardos, J. D. IEEE transactions on robotics. 2021. 37(6). pp. 1874-1890.

11. Nister D. EEE-T-PAMI. 2004. vol. 26, No. 6. pp. 756-770.

12. Martinez-Otzeta JM, Rodriguez-Moreno I, Mendialdua I, Sierra B. Sensors. 2023. No. 1. 327 p.

13. Geiger A., Lenz P., Stiller C., and Urtasun R. The International Journal of Robotics Research. 2013, Res. 32. pp. 1231–1237. DOI:10.1177/0278364913491297.

14. Terven J., Cordova-Esparza D.M., Romero-Gonzalez, J.A. Machine Learning and Knowledge Extraction 5. 2023. No. 4. pp. 1680-1716.

15. Tolkachev D.S. Inzhenernyj vestnik Dona, 2013, N. 3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1840.

Дата поступления: 11.06.2025

Дата публикации: 25.07.2025