

Система многообъектной визуально-инерциальной одометрии беспилотного автомобиля

Ж. Цай, Ц. Цзинь, А.В. Бобков

Московский государственный технический университет и. Н.Э. Баумана

Аннотация: Данная работа посвящена построению системы визуально-инерциальной одометрии беспилотного автомобиля, использующей в качестве источника информации как бинокулярные камеры, так и инерциальные датчики, которая была бы способна одновременно определять собственное положение автомобиля и относительное положение других участников дорожного движения. Для обеспечения точной и непрерывной локализации предлагается использовать инерциальную навигационную систему и два типа особых точек изображения. Модели глубокого обучения используются для точного и надежного отслеживания особых точек. Для достижения эффективного и надежного сопоставления объектов между двумя кадрами предлагается многоуровневый механизм ассоциации данных, учитывающий возможные ошибки различных компонентов системы. Результаты экспериментов демонстрируют осуществимость и потенциал применения предлагаемой системы.

Ключевые слова: многообъектная визуально-инерциальная одометрия, локализация, ассоциация данных, отслеживание 3D - динамических объектов.

1. Введение

Определение собственного положения (локализация) является одной из ключевых задач в мобильных робототехнических системах. Здесь под положением робота будем понимать его трехмерные координаты и углы ориентации в глобальной системе координат. Визуальной одометрией (ВО) будем называть систему, использующую камеру, закрепленную на роботе, для определения его местоположения и ориентации. Особыми точками изображения будем называть точки, имеющие особые характеристики, позволяющие отличать их от обычных пикселей. Визуальная одометрия использует принцип многоракурсной геометрии для оценки движения камеры путем поиска нескольких соответствий особых точек на двух кадрах изображениях. Поскольку инерциальная навигационная система (ИНС) выполняет измерения в реальном масштабе и работает на более высокой частоте, чем камера, измерения от обоих устройств часто объединяются (комплексируются), чтобы получить более надежный результат локализации.

Такая система называется визуально-инерциальной одометрией (ВИО) [1].

Объектом исследования данной работы является беспилотные автомобили. В дорожных сценах обычно присутствуют и другие объекты, такие как различные транспортные средства, пешеходы и велосипедисты. Многие из этих сторонних объектов движутся относительно камеры, что усложняет стабильную работу традиционной визуальной одометрии. В реальной задаче управления движением беспилотного автомобиля при локализации необходимо также оценивать движение других динамических объектов на сцене. Такая задача называется задачей одновременной локализации и многообъектного отслеживания по видеоизображению, или задачей оценки движения нескольких объектов (мультидвижение). Система на основе ИНС и камер, разработанная для решения этой задачи, называется системой многообъектной визуально-инерциальной одометрии (далее МВИО).

В предыдущей работе [2] был приведен обзор методов решения задачи оценки мультидвижения и предлагается комплексная схема решения, объединяющее семантическую информацию и геометрическую информацию. В данной работе на основе этой комплексной схемы строится надежная система МВИО. Основными результатами этой работы являются:

1. Модели глубокого обучения используются для оценки карты оптического потока и карты диспаратности с целью обеспечения более точных прогнозов положения отслеживаемых особых точек.

2. Учет ситуации, когда несколько моделей глубокого обучения как компонентов системы оказываются неточными или дают сбой: предлагается двухуровневый механизм ассоциации данных, обеспечивающий точность и непрерывность отслеживания объектов.

3. На основе комплексирования датчиков и использования графического процессора и многопоточной технологии создана система многообъектной визуально-инерциальной одометрии, которая может работать в режиме

реального времени, а ее эффективность подтверждена экспериментально.

2. Структура системы МВИО

В предыдущей другой работе [3] была описана система робастной визуально-инерциальной одометрии (РВИО), глубоко оптимизированная для дорожных сцен. Эта система РВИО строит надежную оценку собственной траектории движения беспилотного автомобиля путем балансировки количества и качества отслеживания особых точек, строгого контроля распределения глубины особых точек и многоуровневого механизма оптимизации положения камеры. В данной работе построена многообъектная визуальная инерциальная одометрия (МВИО) на основе фреймворка системы РВИО. Метод оценки положения камеры в этой системе точно такой же, как и в РВИО. В системе МВИО введена задача ассоциации данных более высокого уровня, отслеживание объектов, чтобы обеспечить одновременную оценку движения камеры и других динамических объектов. Для эффективного и точного достижения ассоциации нескольких уровней данных объединяются традиционные методы и методы на основе глубокого обучения. Таким образом, система МВИО является расширением системы РВИО.

Структура предлагаемой системы МВИО показана на рисунке 1. Система состоит из трех основных компонентов: блок предварительной обработки изображений, блок отслеживания объектов и блок построения локальной карты. Среди них операции фильтрации и управления распределением обнаружения и сопоставления особых точек, унаследованные от системы РВИО, включены в дочернем модуле А модулей отслеживания камеры и объектов. На рисунке 1 все операции в синем пунктирном поле выполняются на графическом процессоре, тогда как операции в зеленом пунктирном поле выполняются на центральном процессоре. Методы, используемые в основных модулях, описаны ниже.

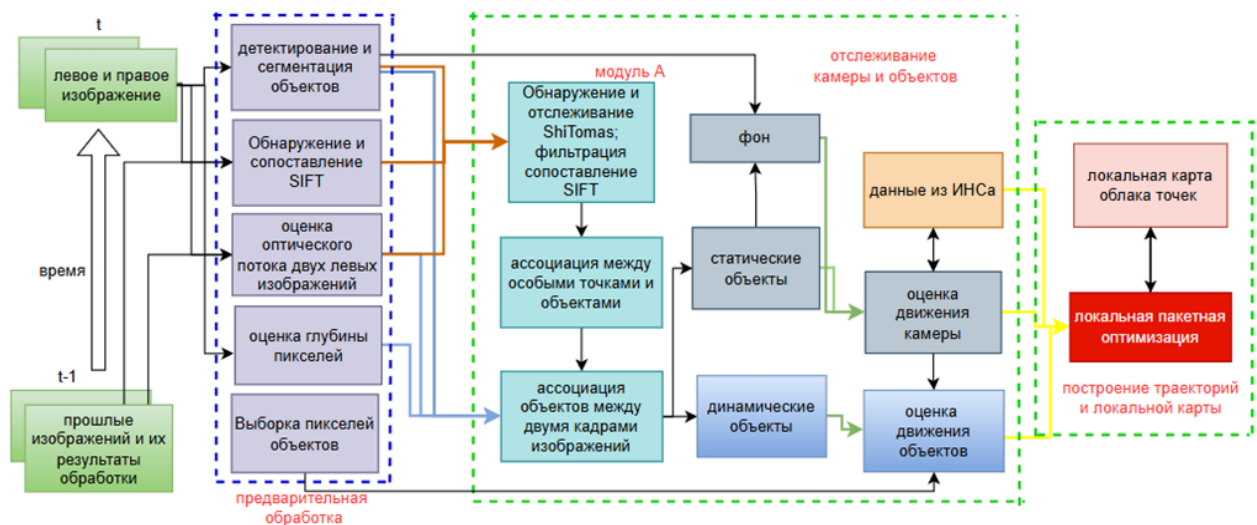


Рис. 1. Архитектура системы надежной многообъектной
визуально-инерциальной одометрии

В данной работе оценка движения выполняется только для объектов, которые можно рассматривать как твердые тела, т. е. обнаруженные пешеходы и велосипедисты исключаются из рассмотрения. На рисунке 2 показано определение системы координат в задаче оценки мультидвижения. На рисунке 2,а показаны независимые траектории движения любых двух объектов А и В и камеры С, где движение между двумя кадрами изображения представлено изменением положения соответствующей системы координат объекта. Камера и ИНС закреплены на автомобиле и предполагается, что их относительное положение не меняется. На рисунке 2,б показаны трёхмерные точки на двух объектах, наблюдаемых камерой в определенный момент.

Здесь и далее обычным шрифтом будут обозначены скалярные величины и объекты, а жирным шрифтом – векторные и матричные величины. Введем следующие обозначения. Обозначим a^N положение некоторой точки a в системе координат N . Индексом w будем обозначать мировую (географическую) систему координат. Она строится таким образом, что направление гравитации совпадает с осью z . Пусть c_k – система координат камеры в момент времени k , то есть во время съемки k -того изображения. Пусть b_k – это система координат тела в момент времени k , который

определяется как такой же, как и в системе координат ИНС. Пусть L_i^k – система координат i -того объекта в момент времени k . В частности, $i=0$ соответствует статическим структурам на сцене, отличные от подвижных объектов. Пусть R_M^N и p_M^N – соответственно матрица вращения и вектор перемещения из системы координат M в систему N . Эти две величины могут образовывать соответствующую матрицу положения T_M^N . Пусть ${}^k_l m_M$ – координата l -той трехмерной точки в некоторой системе координат M в момент времени k . Каждая трехмерная точка и ее проецируемая двухмерная точка на изображении принадлежат определенному объекту.

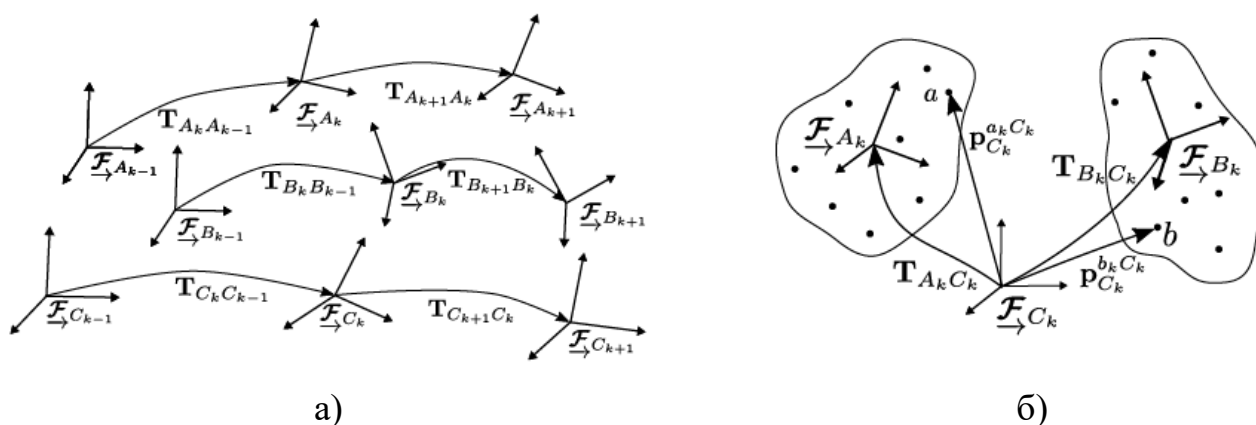


Рис. 2. Системы координат объектов в задаче оценки мультидвижения

2.1. Блок предварительной обработки визуальной информации

Для устранения неопределенности масштаба оценки движения объекта в данной работе предлагается использовать стереоректифицированную бинокулярную камеру, которая позволяет вычислять значение глубины трехмерной точки объекта в системе координат камеры каждого кадра. После получения новой пары изображений три глубокие нейронные сети используются для выполнения сегментации экземпляров в текущем левом изображении, оценки карты диспаратности для текущей пары стереоизображений и оценки оптического потока между двумя левыми изображениями соответственно. Модель сегментации объектов используются для прогнозирования категорий, ограничивающих рамок и масок сегментации

объектов в пределах указанного диапазона категорий, которые появляются на изображении. Оценка оптического потока используется для сопоставления всех пикселей изображений двух кадров. Такое сопоставление недостаточно точно, однако оно позволяет задать достаточные ограничения сопоставления для динамических объектов, занимающих небольшую область изображения, а также помочь выполнить межкадровую ассоциацию данных на уровне объекта. Оценка карты диспаратности стереоскопических изображений на самом деле представляет собой специальную оценку оптического потока, которая ограничивает направление оптического потока пикселей между двумя изображениями только боковым направлением изображения. Все модели нейронных сетей конвертируются в формат TensorRT для достижения максимальной скорости работы на графическом процессоре.

Как и в системе РВИО, для обеспечения качества и количества сопоставления особых точек в системе МВИО также используются два типа угловых точек – SIFT и Shi-Tomasi. Выполняются задачи обнаружения и сопоставления SIFT на графическом процессоре. Наконец, для каждого допустимого объекта, обнаруженного в текущем изображении, равномерно выбирается определенное количество пикселей, и их трехмерные координаты в системе координат камеры вычисляются на основе карты диспаратности. Эти точки выборки будут использоваться для ассоциации объектов между двумя кадрами. Порядок выполнения всех задач на графическом процессоре показан на рисунке 2 (сверху вниз). Это сделано для того, чтобы выполнять их максимально параллельно с другими задачами на центральном процессоре для достижения минимальных временных затрат.

2.2. Блок отслеживания камеры и подвижных объектов

Модуль отслеживания решает три задачи: ассоциация между точками, ассоциация точками и объектами, а также ассоциация между объектами и объектами. Методы обнаружения, отслеживания и фильтрации особых точек,

используемые в МВИО, в основном такие же, как и в РВИО. Из-за ограничений основных допущений метода оптического потока, необходимо задать прогноз положений точек сопоставления на левом и правом изображениях текущего кадра для отслеживаемой точки Shi-Tomasi. В предыдущей работе РВИО использовались либо модель движения с постоянной скоростью, либо интеграция измерений ИНС для вычисления прогнозируемого движения $T_{init}^{L_0}$ камеры. Затем $T_{init}^{L_0}$ используется для расчета прогноза положений точек сопоставления. Аналогично, модель постоянной скорости объекта может использоваться для прогнозирования положений отслеживаемых точек объекта. Однако этот метод нельзя использовать для статических точек без значений глубины в предыдущем кадре или точек на новых объектах, которые появляются в предыдущем или текущем кадре. Более того, ошибка в прогнозировании движения камеры или объектов также будет передана в прогнозирование положения точки сопоставления, что приведет к быстрому накоплению ошибок в оценке траектории камеры или объектов. В этой работе карта оптического потока и карта диспаратности, полученные с помощью нейронной сети, используются для установки значений прогноза положений точек сопоставления на левом и правом изображениях соответственно. Это отделяет сопоставление особых точек от результатов оценки движения соответствующего объекта в предыдущем кадре. Аналогично, карты диспаратности используются для задания прогноза положения точки сопоставления на правом изображении для точек Shi-Tomasi на левом изображении предыдущего кадра или текущего кадра, которым необходимо найти стереосопоставление. Выбросы в 3D-2D сопоставлений статичных особых точек фильтруются на основе ошибки перепроекции с использованием $T_{init}^{L_i}$.

Ассоциация между точками и объектами формируется на основе результатов отслеживания точек, обнаружения и сегментации объектов.

Каждому допустимому объекту, обнаруженному на текущем изображении, присваивается временный идентификатор объекта и категория. Для каждой отслеживаемой точки на текущем изображении записываются глобальный идентификатор и категория объекта этой точки в прошлом кадре, а также временный идентификатор и категория сегментированного объекта, к которому она принадлежит. Все отслеживаемые точки группируются в соответствии с их временными идентификаторами объектов и используются для ассоциации объектов между двумя кадрами. Для каждой особой точки объекта в каждом кадре должно быть доступно значение глубины, которое получается с помощью оценки стереосоответствия или карты диспаратности.

Поскольку нейронная сеть минимизирует общие потери, их результаты часто недостаточно точны в деталях. На качество оценки оптического потока сильно влияют изменения освещенности сцены, а также относительная скорость движения камеры и наблюдаемой структуры. Если объект наблюдается под особым углом или частично перекрыт, то обнаруженная категория объекта подвержена ошибкам, или даже объект не обнаруживается вообще. Край маски объекта, полученный путем сегментации экземпляров, сложно точно совместить с контуром объекта, особенно когда проекции нескольких объектов одной категории на изображении перекрывают друг друга, как показано на рисунке 3. В этом случае, если используются только результаты оптического потока и обнаруженной категории большого количества пикселей в качестве основы для ассоциирования объектов, как в работе [4], легко получить неверные результаты. Напротив, хотя точность сопоставления особых точек намного выше, зачастую отслеживаемых характерных точек, принадлежащих объектам, меньше, особенно объектам, занимающим меньшую площадь на изображении. Поэтому в данной работе предлагается двухуровневый механизм ассоциации объектов, позволяющий максимально повысить точность отслеживания в случае возможного

неправильного или пропуска обнаружения объекта.



Рис. 3. Пример сегментации кадра с использованием модели YOLO-v8

Предыдущий кадр и текущий кадр обозначены как K_1 и K_2 соответственно. Глобальный объект (включая фон) и локальный объект (включая фон) обозначены как ГО и ЛО. Для каждого локального объекта O_N в K_2 наборы $N_{to}^{O_N}$ и $N_{tn}^{O_N}$ используются для записи идентификаторов глобальных объектов, с которыми существуют сопоставления особых точки, и соответствующего количества сопоставления соответственно. Для каждого глобального объекта O_p (включая фон) в K_1 набор $N_{to}^{O_p}$ используется для записи идентификаторов временно связанных локальных объектов. Глобальные объекты и временные объекты, которые должны участвовать в задаче сопоставления двудольного графа, записываются в наборы N_{bip} и N_{bin} соответственно. Набор M_{to}^N используется для записи локальных объектов, завершивших ассоциацию. Наборы M_{lost}^{op} и M_{new}^{op} соответственно используются для записи всех глобальных объектов и локальных объектов, которые в конечном итоге не имеют сопоставленного объекта.

Предполагая, что между глобальным объектом L_i и локальным объектом O_N имеется n сопоставлений особых точек ($\mathbf{p}_k^P \in H^P, \mathbf{p}_k^N \in H^N$). Для проверки сходства двух наборов точек H^P и H^N вычисляются их степени дисперсии d_1

и d_2 соответственно. Для исключения влияния аномальных сопоставлений точек и предположения о распределении наборов точек в качестве меры степени дисперсии используется нормализованное абсолютное медианное отклонение. Если количество точки сопоставления с значением глубины больше порогового значения, вычисляются степени дисперсии d_3 и d_4 двух наборов трехмерных точек. Для набора трехмерных точек значения отклонения не требуют нормализации. Кроме того, если известно прогноза движения $T_{init}^{L_i}$ объекта между двумя кадрами, вычисляется ошибка преобразования движения d_5 этих трехмерных точек сопоставления. Для определения ассоциации между L_i и O_N используется следующая формула:

$$\text{успех}_{acco} = \begin{cases} \frac{\max(d_1, d_2)}{\min(d_1, d_2)} < Th_{var}^{2d}, \text{ если нет 3D точек и оценки } T_{init}^{L_i} \\ \frac{\max(d_1, d_2)}{\min(d_1, d_2)} < Th_{var}^{2d} \wedge \frac{\max(d_1, d_2)}{\min(d_1, d_2)} < Th_{var}^{3d}, \text{ иначе если нет } T_{init}^{L_i} \\ \frac{\max(d_1, d_2)}{\min(d_1, d_2)} < Th_{var}^{2d} \wedge \frac{\max(d_1, d_2)}{\min(d_1, d_2)} < Th_{var}^{3d} \wedge d_s < Th_{dist}, \text{ иначе} \end{cases} \quad (1)$$

Если результат успех_{acco} истинен, то L_i и O_N являются потенциальными объектами ассоциации друг друга. Если n больше порогового значения $Th_{f \in a}$, для проверки ассоциации используется набор сопоставления особых точек. В противном случае, если n больше нуля, то на основе карты оптического потока и карты диспаратности получается множество H^{pixel} сопоставления выбранных пикселей объекта L_i , которые связаны с объектом O_N . Если количество точек в H^{pixel} больше порогового значения, он используется для проверки ассоциации. В частности, если фоном является L_i или O_N , то проверка ассоциации может быть выполнена только путем сопоставления особых точек. Для каждого локального объекта O_N оговаривается, что он имеет потенциальную ассоциацию максимум с двумя глобальными объектами предыдущего кадра, и в этом случае один из них

должен быть фоном прошлого изображения. Это означает, что в предыдущем кадре была обнаружена только часть этого глобального объекта. Эти две части точек можно объединить в один глобальный объект.

Поскольку ассоциация точек не совсем точна, количество потенциальных объектов ассоциации N_{asso} для глобального объекта L_i может быть больше 1. Если N_{asso} равно 2, они могут быть одним глобальным объектом, который на текущем изображении ошибочно сегментирован на два. Если объединение этих двух наборов сопоставления точек удовлетворяет формуле (1), два объекта объединяются в один, который становится потенциальным объектом ассоциации для L_i . В противном случае ассоциации между L_i и этими двумя объектами отменяется. Фактически, этот двухуровневый механизм проверки ассоциации позволяет быстро и точно добиться ассоциации для большинства объектов. Наконец, для всех ГО или ЛО, число потенциальных объектов ассоциации которых равно нулю или больше 2, необходимо найти окончательные ассоциации путем решения задачи сопоставления на основе взвешенного двудольного графа (СВДГ). В этой задаче наилучшая ассоциация объектов достигается путем минимизации общей стоимости сопоставления. Для любой пары глобальных объектов и локальных объектов (L_i, O_N) , участвующих в задаче СВДГ, их стоимость сопоставления $L_i \Upsilon^{O_N}$ выражается с помощью следующей формулы:

$$L_i \Upsilon^{O_N} = \begin{cases} (1.1 - \mu_{pix}) \frac{\max(d_1, d_2)}{\min(d_1, d_2)}, & \text{если нет 3D точек и оценки } T_{init}^{L_i} \\ (1.1 - \mu_{pix}) \frac{\max(d_1, d_2)}{\min(d_1, d_2)} * \frac{\max(d_3, d_4)}{\min(d_3, d_4)}, & \text{иначе если нет } T_{init}^{L_i} \\ (1.1 - \mu_{pix}) \frac{\max(d_1, d_2)}{\min(d_1, d_2)} * \frac{\max(d_1, d_2)}{\min(d_1, d_2)} * \exp\left(\frac{d_s}{Th_{dist}} - 1\right), & \text{иначе} \end{cases} \quad (2)$$

где μ_{pix} – это доля пикселей в наборе точек выборки объекта L_i , которые связаны с объектом O_N . В частности, если μ_{pix} равно нулю, то $L_i \Upsilon^{O_N}$ равно большой постоянной величине Ω . Кроме того, необходимо также учитывать, совпадают ли обнаруженные категории объектов. Если их категории

различны, то $L_i \gamma^{O_N}$ умножается на штрафной коэффициент $\mu = 2/\mu_{pix}$. Для L_i или O_N , если все его стоимости сопоставления равны Ω , он не участвует в задаче СВДГ и напрямую присоединяется к набору M_{lost}^{op} или M_{new}^{op} . Задача СВДГ решается с помощью венгерского алгоритма [5].

Псевдокод предложенного алгоритма ассоциации объектов выражается следующим образом.

Алгоритм 1: алгоритм ассоциации объектов

```
00  Начало
01  успех_ассо = 0;
02  Начало цикла 1: для каждого локального объекта  $O_N$  в  $K_2$ 
03      Начало цикла 2: для каждого элемента  $O_p$  в наборе  $N_{to}^{O_N}$  и  $n_{mn}^{O_p}$  в наборе  $N_{mn}^{O_N}$ 
04          Если  $O_p$  не является фоном и количество элементов в  $N_{to}^{O_N}$  меньше 2, тогда
05              Конец Цикла 2;
06              Конец Цикла 1;
07          Иначе
08              Если  $n_{mn}^{O_p}$  больше порогового значения, тогда
09                  Вычисление по формуле (1) на основе набора сопоставления особых точек:
10                  Если результат истинен, тогда
11                      добавить  $O_N$  в  $M_{to}^{O_N}$ , успех_ассо = 1;
12                  КонецЕсли;
13              КонецЕсли;
14              Если успех_ассо == 0, тогда
15                  Вычисление по формуле (1) на основе набора сопоставления пикселей:
16                  Если результат истинен, тогда
17                      Идентификатор  $O_N$  добавить в  $M_{to}^{O_N}$ , успех_ассо = 1;
18                  КонецЕсли;
19              Конец Если;
20          Конец Если;
21      Конец Цикла 2;
22  Конец Цикла 1;
23  Начало цикла 3: для каждого ГО (кроме фона)
24      Если нет элементов в  $M_{to}^{O_N}$ , тогда
25          добавить  $O_p$  в  $N_{bir}$ ;
26      ИначеЕсли в  $M_{to}^{O_p}$  только один элемент и он является фоном в  $K_2$ , тогда
27          добавить  $O_p$  в  $N_{bir}$ ; удалить все элементы из  $N_{to}^{O_p}$ ;
28      ИначеЕсли в  $M_{to}^{O_p}$  только один элемент и он не является фоном в  $K_2$ , тогда
29          добавить  $O_N$  в  $M_{to}^N$ ;
30      ИначеЕсли в  $M_{to}^{O_p}$  два элемента, тогда
31          Попытка объединить эти два ЛО:
32          Если успех
33              Добавить оба ЛО  $M_{to}^{O_p}$  и в  $M_{to}^N$ ;
```

```
34     Иначе
35         добавить  $O_p$  в  $N_{biP}$ ; оба ЛО добавить в  $N_{biP}$ ; удалить все элементы из  $N_{mo}^{Op}$ ;
36     Иначе
37         добавить  $O_p$  в  $N_{biP}$ ; все ЛО добавить в  $N_{biP}$ ; удалить все элементы из  $N_{mo}^{Op}$ ;
38     Конец Если;
39 Конец цикла 3;
40 Если количество элементов в  $N_{biP}$  и количество элементов в  $N_{biN}$  больше 0, тогда
41     Задача СВДГ выполняется для всех элементов в  $N_{biP}$  и  $N_{biN}$ ;
42     Начало цикла 4: для каждой эффективной пары сопоставление  $(O_p, O_N)$  из СВДГ
43         добавить  $O_N$  в  $M_{mo}^{Op}$  и  $M_{mo}^N$ ;
44     Конец Цикла 4;
45 КонецЕсли;
46 Все глобальные объекты  $O_p$ , не имеющие элементов в их  $M_{mo}^{Op}$ , добавить в  $M_{lost}^{op}$ ;
47 Все локальные объекты, которые не появляются в  $M_{mo}^N$ , добавить в  $M_{new}^{op}$ ;
48 Конец.
```

После выполнения этого алгоритма все локальные объекты преобразуются в ассоциированные с ними глобальные объекты или новые глобальные объекты. Категория отслеживаемого объекта наследуется от его глобальной категории, записанной в предыдущем кадре. В частности, если окончательный объект ассоциации для локального объекта O_N является фоном предыдущего кадра, O_N также преобразуется в новый глобальный объект. Алгоритм надежен и способен правильно отслеживать объекты даже при наличии пропущенных обнаружений и различных типов ложных обнаружений. Кроме того, состояние движения отслеживаемого объекта, который был статичным в предыдущем кадре, можно приблизительно определить по формуле (1). Если объект неподвижен, то сопоставление его точек можно использовать для оценки движения камеры. Это также уменьшает количество потенциальных динамических объектов, требующих оценки их движения.

Как показано на рисунке 1, для оценки движения отслеживаемых объектов требуется известная оценка положения камеры. Для обеспечения точности оценку движения камеры и оценку движения объекта необходимо выполнять последовательно. Когда производительность в реальном времени

имеет решающее значение, две задачи оценки движения можно выполнять параллельно, используя прогноз движения камеры $T_{init}^{L_0}$. Кроме того, если количество потенциальных динамических объектов превышает пороговое значение, многопоточность снова используется при оценке движения объектов. Объекты, оценка движения которых не удалась, становятся новыми глобальными объектами. Если значение движения объекта отличается от движения камеры менее чем на пороговое значение, он помечается как статическое. Для каждой трехмерной точки оптимизируется только ее глубина в кадре, где она была впервые обнаружена. Для объекта с оценкой движения значение глубины его точки в каждом кадре обновляется с использованием преобразования движения. Наконец, траектории и особые точки каждого объекта добавляются и удаляются на локальной карте.

3. Метод оценки движения камеры и объектов в системе МВИО

Для оценки межкадрового движения камеры используется сопоставление точек, принадлежащих статическим структурам фона и статическим объектам. Сопоставление точек, принадлежащих каждому динамическому объекту, используется для оценки движения объекта.

Для оценки траектории движения камеры на основе визуального измерения и инерционного измерения используется тот же многослойный механизм оптимизации, что и в системе РВИО [3]. Разница в том, что в этой работе для дополнения сопоставления статических точек используются только отслеживаемые точки на статических объектах в предыдущем кадре, которые соответствуют требованиям погрешности репроекции.

3.1. Оценка движения динамических объектов

Согласно [4], движения твердого тела от момента времени $k-1$ до момента времени k выражается в мировой системе координат как:

$${}_{L_i^{k-1}}^{L_i^k} T_w = T_{L_i^k}^w \cdot \left(T_{L_i^{k-1}}^w \right)^{-1}. \quad (3)$$

Предположим, что трехмерная точка ${}^k_l \mathbf{m}_w$ принадлежит твердому объекту L_i . Положение этой точки, выраженное в системе координат объекта L_i остается постоянным в каждый момент времени, обозначим его как ${}_l \mathbf{m}_{L_i}$. Тогда преобразование глобального положения ${}^{k-1}_l \mathbf{m}_w$ и ${}^k_l \mathbf{m}_w$ точки ${}_l \mathbf{m}_{L_i}$ в два момента времени выражается, как:

$${}^k_l \mathbf{m}_w = {}_{L_i}^{L_i^k} \mathbf{T}_w \cdot {}^{k-1}_l \mathbf{m}_w. \quad (4)$$

Аналогично преобразование положения ${}^{k-1}_l \mathbf{m}_{c_{k-1}}$ и ${}^{k-1}_l \mathbf{m}_{c_k}$ этой точки в системе координат камеры двух кадров можно выразить, как:

$${}^k_l \mathbf{m}_{c_k} = (\mathbf{T}_{c_k}^w)^{-1} \cdot {}_{L_i}^{L_i^k} \mathbf{T}_w \cdot \mathbf{T}_{c_{k-1}}^w \cdot {}^{k-1}_l \mathbf{m}_{c_k}. \quad (5)$$

Здесь ${}_{L_i}^{L_i^k} \mathbf{T}_{c_k} = (\mathbf{T}_{c_k}^w)^{-1} \cdot {}_{L_i}^{L_i^k} \mathbf{T}_w \cdot \mathbf{T}_{c_{k-1}}^w$ представляет собой движение точки объекта L_i , выраженное в системе координат камеры в момент времени k . Все точки твердого тела L_i удовлетворяют преобразованию движения, показанному в уравнении (5). Это позволяет основывать оценку движения объекта только на значениях положения трехмерных точек объекта в системе координат камеры в двух кадрах и не зависеть от глобального положения объекта и камеры. В работе [4] задача оценки движения объектов тесно связана с задачей оценки положения камеры. Однако в дорожных сценах часто бывает мало надежных статических ориентиров и трехмерных особых точек динамических объектов, что затрудняет формирование эффективных совместных ограничений на оценки положения камеры и движения объектов. На практике, в ситуациях, когда ВО дает сбой между двумя кадрами, можно использовать интеграцию измерений ИНС для получения краткосрочной оценки положения. Оценка положения и размера объекта в системе координат камеры представляет собой более сложную задачу. Одним из важных методов решения является модель глубокого обучения, основанная на многоракурсной геометрии и обзоре сверху (с высоты птичьего полета) [6]. Полученная здесь

оценка движения объекта в системе координат камеры повышает точность расчета глубоких признаков в этих методах.

Если известны трехмерные координаты достаточного количества точек сопоставления $\Upsilon(L_i^k)$ в системе координат камеры двух кадров $k-1$ и k , то движение объекта можно рассчитать по формуле (5). Кроме того, аналогично оценке положения камеры, может быть установлена совместная оптимизация движений объекта L_i и глубины точек объекта в скользящем окне. Однако оптимизация для каждого объекта отдельно также требует много времени.

Эксперименты

Для проверки эффективности предлагаемой многообъектной визуально-инерциальной одометрии был использован открытый набор данных KITTI-Tracking [7]. Набор содержит 21 видеопоследовательность с достоверной информацией о положении камеры и объектов. Для эксперимента выбираются только те последовательности, которые содержат движущиеся объекты. В наборе KITTI аннотировано шесть категорий объектов, включая пешеходов, но в данной работе отслеживаются только три типа транспортных средств.

В экспериментах был использован один графический процессор RTX-3070 для развертывания трех моделей глубоких нейронных сетей, что позволило всесторонне учитывать точность, скорость и потребление памяти. Такие ограничения, с одной стороны, отражают реальные проблемы, с которыми сталкиваются коммерческие продукты, а с другой – позволяют оценить скорость и надежность предлагаемой системы МВИО. Для оценки оптического потока была выбрана модель RapidFlow [8], которая основана на самой совершенной схеме итерационной оптимизации в этой области исследований. Для оценки стереосоответствия и сегментации экземпляров были выбраны модель DP-Stereo [9] и модель YOLO-v8 [10] соответственно. Диапазон оцененной диспаратности в модели DP-Stereo, установлен в диапазоне $(0...192]$. Согласно внешним параметрам стереокамеры в наборе

данных KITTI, в данной работе минимальная эффективная глубина точки, для которой необходимо получить значения глубины через карту диспаратности, установлена равной 2м. С другой стороны, максимальная эффективная глубина точки составляет 30м, исходя из разрешения изображения.

Таблица № 1

Средняя ошибка оценки глобального положения камеры

система	Средняя ошибка оценки перемещения, %	Средняя ошибка оценки поворота, °/м
РВИО (с ИНС)	2.26	0.0812
ORB-SLAM3(без ИНС)	2.57	0.1064
МВИО	1.46	0.0393

Как и в экспериментах на системе РВИО, в данном эксперименте для расчета погрешности результатов глобальной локализации камеры используется критерий, предложенный в наборе данных KITTI. Для сравнения также были проведены эксперименты в системе РВИО и системе ORB-SLAM3, и результаты показаны в таблице 1. Система МВИО достигает значительно более точных результатов глобальной локализации камеры. Первая причина заключается в том, что модели глубокого обучения, используемые в системе МВИО, обеспечивают значения прогноза оптического потока и глубины для особых точек в периоде инициализации системы, которые в целом точнее значений прогноза с использованием модели движения с постоянной скоростью в системах РВИО и ORB-SLAM3. Во-вторых, система МВИО явно делит объекты в сцене на динамические, статические и новые объекты в каждом кадре и в текущем кадре использует только точки отслеживания статических объектов в предыдущем кадре для дополнения сопоставления статических точек. Обе эти причины увеличивают количество сопоставлений статических точек и улучшают их общее качество, что является наиболее важным фактором в повышении точности оценки

движения камеры.

При отсутствии замыкания траектории или использования ССП оценки абсолютного положения чистой визуально-инерциальной одометрии неизбежно будут постепенно расходиться. Поэтому точность результатов оценки межкадрового движения камеры также рассматривается в эксперименте. Точность оценки движения камеры и объектов выражается следующей ошибкой изменения положения:

$$\mathbf{E}_{motion} = \hat{\mathbf{T}}^{-1} \cdot \mathbf{T}, \quad (13)$$

где $\hat{\mathbf{T}}$ - оценка движения камеры или объектов ${}_{L_i^{k-1}}^{L_i^k} \mathbf{T}_{c_k}$. $\mathbf{T}$ - соответствующее истинное значение. Ошибка смещения E_t (м/кадр) вычисляется как квадратичная норма составляющей смещения в $\mathbf{E}_{motion}$. Ошибка вращения E_r (°/кадр) вычисляется как абсолютное значение угла поворота в представлении оси-угла составляющей вращения в $\mathbf{E}_{motion}$. Для разных временных шагов камеры и разных объектов в последовательности вычисляется среднеквадратичная ошибка оценки движения камеры или объектов.

Ошибки оценки движения камеры и объектов показаны в таблице 2. В таблице также приведены для сравнения результаты для аналогичной системы VDO-SLAM [4], в которой представлены результаты только для части последовательностей. Точность системы МВИО немного хуже, чем у системы VDO-SLAM. Для оценки движения камеры, без предоставления точных параметров калибровки ИНС, тесное слияние визуальных измерений и инерциальных измерений трудно улучшить общую точность оценки. Точность оценки движения объектов в МВИО сопоставима с точностью VDO-SLAM. Это связано с тем, что система VDO-SLAM рассчитывает движение объекта на основе оценки оптического потока плотных пикселей на объекте. Оценка в системе МВИО основана на небольшом количестве сопоставления надежных особых точек, а пиксели добавляются в качестве дополнений только в тех случаях, когда особых точек явно недостаточно. Таким образом, метод в VDO-

SLAM является частным случаем метода в МВИО. Кроме того, для снижения требования к памяти графического процессора и повышения скорости работы все выбранные модели нейронных сетей в системе МВИО не являются оптимальными, и все они преобразуются в форме TensorRT с более низкой точностью. Все модели нейронных сетей в VDO-SLAM работают в режиме офлайн, что позволяет выбрать оптимальную модель. Это означает, что точность системы МВИО может быть дополнительно улучшена.

В таблице 2 также указано количество отслеживания объектов, идентифицированных как динамические в каждой последовательности. Для системы VDO-SLAM соответствующие результаты не указаны. В каждом кадре каждый объект, который не отслеживается или чья предполагаемая средняя трехмерная точка не попадает в истинную трехмерную ограничивающую рамку, считается потерей отслеживания. Однако, эта статистика неточна, поскольку файлы аннотаций в KITTI не указывают, какие объекты являются динамическими. В этом эксперименте разница между истинным значением движения объекта и движением камеры используется для определения того, является ли объект динамическим. Кроме того, в МВИО рассматриваются только твердые тела на расстоянии до 25 м, что намного меньше максимального расстояния объектов, аннотированных в наборе данных KITTI. Система МВИО может стабильно отслеживать большинство объектов в указанном диапазоне расстояний, даже если объект неправильно классифицирован или пропущен моделью обнаружения в определенном кадре. Однако, поскольку оценка глубины особых точек или пикселей не всегда точна, динамические объекты иногда могут считаться потерянными из-за большой погрешности среднего координата трехмерных точек. Кроме того, из-за наличия ошибок оценки движения система не может всегда эффективно различать статические объекты и динамические объекты и поэтому рассматривает большинство отслеживаемых объектов как динамические. В

последующей работе семантическую информацию о дороге можно использовать для непосредственной обработки транспортных средств на парковочных местах по обеим сторонам дороги как статических, а порог оценки объектов как динамических может быть увеличен.

Таблица № 2

Ошибки оценки движения камеры и объектов

	Предлагаемый МВИО					VDO-SLAM (RGB-D)			
	Камера		Объекты			Камера		Объекты	
Видео	$E_r, ^\circ$	$E_t, \text{м}$	$E_r, ^\circ$	$E_t, \text{м}$	Кол-во отслеж., прав/пот.	$E_r, ^\circ$	$E_t, \text{м}$	$E_r, ^\circ$	$E_t, \text{м}$
0000	0.0898	0.0927	1.1445	0.3786	65/148	0.0512	0.0541	1.0429	0.1061
0001	0.0719	0.2416	0.9722	0.3412	431/365	0.0368	0.1201	1.2504	0.2129
0002	0.0357	0.0648	1.3527	0.4665	84/197	0.0152	0.0427	1.4311	0.2358
0003	0.0664	0.1374	0.4316	0.0974	115/170	0.0377	0.0866	0.3782	0.1055
0004	0.0552	0.2131	1.7112	0.5235	103/226	0.0476	0.1127	1.0211	0.3376
0005	0.0487	0.1597	0.5871	0.1517	96/248	0.0246	0.0982	0.4482	0.1098
0006	0.0638	0.0313	0.9834	0.1067	73/199	0.0477	0.0187	1.2156	0.1304
0007	0.0785	0.1592	0.5069	0.1103	484/455	-	-	-	-
0008	0.0488	0.0936	0.3687	0.1628	122/357	-	-	-	-
0009	0.0843	0.1837	0.5206	0.0985	453/577	-	-	-	-
0010	0.0556	0.1429	0.2348	0.0992	316/236	-	-	-	-
0011	0.0648	0.0347	0.1837	0.1081	284/337	-	-	-	-
0018	0.0367	0.0782	0.1859	0.0794	237/423	0.0192	0.0715	0.1470	0.1131
0019	0.0645	0.1134	0.4903	0.1584	498/334	-	-	-	-
0020	0.0436	0.3521	0.6315	0.1273	645/3949	0.0293	0.1643	0.5401	0.0944

С точки зрения скорости работы системы VDO-SLAM не является системой реального времени, поскольку все ее модели нейронных сетей работают в режиме офлайн. Остальная часть той системы может работать только со скоростью 5–8 кадров в секунду. Предлагаемая система МВИО – это система, полностью работающая в реальном времени, поскольку в

большинстве случаев вся ее оценка движения основана на разреженных особых точках. Объединение инерциальных и визуальных измерений позволяет выполнять оценку движения камеры и оценку движения объектов параллельно. Среднее время выполнения МВИО в тестовой последовательности составляет 94-105 мс на кадр в зависимости от количества динамических объектов в кадре и количества итераций в задаче совместной оптимизации. Если аппаратные условия позволяют одновременно запускать все модели нейронных сетей, скорость работы системы МВИО может быть увеличена примерно до 14 кадров в секунду.

Заключение

В данной работе предложена система многообъектной визуально-инерционной одометрии для одновременной оценки собственного положения и оценки движения других динамических объектов дорожной сцены. На основе измерений ИНС и отслеживания двух типов особых точек достигается точная и непрерывная оценка движения камеры и большинства объектов. Учитывая возможные ошибки в модуле предварительной обработки визуальной информации, предлагаемый двухслойный механизм ассоциации на основе особых точек и плотных пикселей обеспечивает надежность ассоциации объектов. Применение технологии слияния датчиков и многопоточности позволяет системе работать в режиме реального времени. Эффективность предлагаемой системы проверена с использованием набора данных KITTI.

Литература

1. Huang, G. Visual-inertial navigation: A concise review // International conference on robotics and automation (ICRA). 2019. pp. 9572-9582.
2. Жуйшэн Ц., 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.

3. Цай Ж., Цзинь Ц., Бобков А.В. Система робастной визуально-инерциальной одометрии беспилотного автомобиля. // Инженерный вестник Дона, 2025, №8. URL: ivdon.ru/ru/magazine/archive/n8y2025/10306.

4. Zhang J., Henein M., Mahony R., Ila, V. VDO-SLAM: A visual dynamic object-aware SLAM system // arXiv preprint arXiv:2005.11052, 2020. URL: doi.org/10.48550/arXiv.2005.11052.

5. Kuhn, H.W. The Hungarian Method for the assignment problem / Naval Research Logistics Quarterly. 1955, vol. 2, pp. 83-97.

6. Z. Li et al. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers IEEE Transactions on Pattern Analysis and Machine Intelligence. 2025, vol. 47, no. 3, pp. 2020-2036.

7. Geiger A., Lenz P., Urtasun R. Vision meets robotics: The KITTI dataset. // The International Journal of Robotics Research. 2013, Res. 32. pp. 1231–1237.

8. Sun S., Sun X., He B., and Luo Q. RapidFlow: an efficient approach to continuous subgraph matching. Proc. VLDB Endow. 2022. 15, 11, 2415–2427. URL: doi.org/10.14778/3551793.3551803.

9. Duggal S., Wang S., Ma W.C., Hu R. and Urtasun R. Deeppruner: Learning efficient stereo matching via differentiable patchmatch // Proceedings of the IEEE/CVF international conference on computer vision. 2019. pp. 4384-4393.

10. Yaseen Muhammad. What is yolov8: An in-depth exploration of the internal features of the next-generation object detector. arXiv preprint arXiv:2409.07813. 2024. URL: doi.org/10.48550/arXiv.2408.15857.

References

1 Huang, G. International conference on robotics and automation (ICRA). 2019. pp. 9572-9582.

2. Czaj Zh., Bobkov A.V. Avtomatizaciya. Sovremennye texnologii. 2024. N.



3. pp. 105-118.

3. Czaj Zh., Czzin Cz., Bobkov A.V. Inzhenernyj vestnik Dona. 2025. N. 8.
URL: ivdon.ru/ru/magazine/archive/n8y2025/10306.

4. Zhang J., Henein M., Mahony R., Ila V. arXiv preprint arXiv:2005.11052, 2020. URL: doi.org/10.48550/arXiv.2005.11052.

5. Kuhn, H.W. Naval Research Logistics Quarterly, 1955, vol. 2, pp. 83-97.

6. Z. Li et al. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2025, vol. 47, no. 3, pp. 2020-2036.

7. Geiger A., Lenz P., Urtasun R. The International Journal of Robotics Research. 2013, Res. 32. pp. 1231–1237.

8. Sun S., Sun X., He B., and Luo Q. Proc. VLDB Endow. 2022. 15, 11, 2415–2427. URL: doi.org/10.14778/3551793.3551803.

9. Duggal S., Wang S., Ma W.C., Hu R. and Urtasun R. 2019. pp. 4384-4393.

10. Yaseen Muhammad. arXiv preprint arXiv:2409.07813. 2024. URL: doi.org/10.48550/arXiv.2408.15857.

Дата поступления: 15.07.2025

Дата публикации: 25.08.2025