РАСПОЗНАВАНИЕ СЦЕН В ЗАДАЧЕ ГЛОБАЛЬНОЙ ЛОКАЛИЗАЦИИ МОБИЛЬНОГО РОБОТА С ИСПОЛЬЗОВАНИЕМ МОДЕЛЕЙ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ ИЗОБРАЖЕНИЙ И ГРАФОВЫХ ПОДХОДОВ[1]

Московский А. Д.[2]
(*НИЦ «Курчатовский институт», Москва*)

Работа посвящена задаче локализации мобильных роботов по визуальным семантическим данным. Центральным элементом такой задачи является распознавание сцен – поиск соответствия между наблюдаемыми объектами и объектами, нанесенными на карту местности (семантическая карта). Предлагаются два метода, использующие определение геометрических особенностей на наблюдаемой сцене и поиск их на карте с помощью различных подходов на графах. Предложенный способ определения отношений между объектами, использующийся в обоих методах, позволяет учитывать погрешности оценки расстояний бортовыми сенсорами. Помимо использования геометрических особенностей в работе также рассматривается применение нейросетевых моделей, которые формируют вектор признаков по изображению, тем самым позволяя определить их визуальное сходство. Визуальное сходство используется для нормирования и оценки результатов, полученных предложенными методами на основе графовых подходов. Кроме того, был модифицирован открытый набор данных KITTI-360 для оценки точности решения задач распознавания сцен. Эксперименты на полученном наборе данных продемонстрировали, что предлагаемый подход, сочетающий геометрические особенности и визуальное сходство, значительно повышает точность рассмотренных методов распознавания сцен. По результатм экспериментов сформированы некоторые рекомендации по использованию данных подходов на практике.

Ключевые слова: локализация мобильных роботов, распознавание сцен, векторное представление изображения, теория графов.

1. Введение

Задача локализации в мобильной робототехнике исконно является одной из важнеших. Навигация мобильных устройств

[1] *Работа проведена в рамках выполнения государственного задания НИЦ «Курчатовский институт».*
[2] *Антон Дмитриевич Московский, начальник группы «Системы управления», лаборатории робототехники* (moscowskyad@yandex.ru).

в большинстве случаев опирается на знание о своём местоположении. Для решения задачи локализации разработано большое количество различных методов, подходящих под те или иные условия. Одним из самых распространённых на сегодня средством локализации мобильного робота без средств GNSS (спутниковой навигации) является построение плотной карты окружающей среды [29] на основе данных дальномеров, таких как LiDAR, а также средств получения глубины из изображений, например, при помощи стереозрения [23], камер со специальной подсветкой и нейросетевых моделей извлечения глубины. Карта, построенная при помощи данных средств, представляет собой трёхмерное облако точек. Из плюсов такой карты можно отметить то, что она достаточно подробна, в том числе и для того, чтобы использовать её для дальнейшей навигации, так как содержит информацию о препятствиях и других элементах среды. Выделяются две задачи: построение такой карты (методы SLAM) и локализация по готовой карте, в в том числе без начальных знаний о положении (global localization).

Локализация на плотной карте происходит путём поиска места, где текущий скан (набор получаемых с сенсоров данных в момент времени) лучшим образом накладывается на данные карты, однако у методов, решающих данную задачу, есть ряд трудностей. Согласно обзору по данной тематике [26], среди открытых вопросов можно выделить проблематику ракурсов, когда лидарные данные одного и того же пространства, полученные с разных точек, имеют мало пересечений. Ещё выделяются проблемы, связанные с масштабированием подходов при увеличении размера карты. Безусловно, размер карты имеет огромное значение и для поиска точки, в которой достигается лучшее совпадение, требуется перебрать много кандидатов и, разумеется, их число растёт с ростом размера карты. Другой проблемой является симметричность и повторяемость окружения. Зачастую здания внутри и снаружи достаточно однообразны с точки зрения геометрии: может существовать несколько мест, на которые полученный с сенсоров робота скан одинаково хорошо накладывается.

В задачах SLAM эти проблемы решаются отслеживанием перемещения робота и использованием информации, получен-

ной с предыдущих шагов. То есть на больших картах требуется проводить поиск нового положения лишь в небольшой области вокруг прошлого положения робота и т.п. Однако всё ещё существует ряд случаев, когда требуется произвести локализацию «с нуля». Первый случай – это задача глобальной локализации, которая происходит при старте робота и когда предыдущее положение ему не известно. Также существует вероятность того, что метод мог ошибиться и это привело к тому, что текущие наблюдения с сенсоров не совпадают с положением робота на карте, значит, придётся пересчитывать положение заново. Также в литературе выделяется ситуация kidnapping – перемещение робота третьими лицами, когда он не в состоянии отследить своё смещение, а следовательно, корректно сузить пространство поиска.

Таким образом, определение положения робота без начальной информации является важной задачей в области локализации по плотным картам, однако затруднена такими факторами как симметричность, повторяемость и большими размерами пространства. Эта проблема может быть решена на **семантическом уровне** карты, так как позволяет «сузить» исследуемое пространство, оперируя объектами, а не облаками точек. Требуется видимый роботом набор объектов сопоставить с семантической картой, т.е. решить задачу **распознавания сцены**, иногда также называемую задачей сопоставления данных\объектов (data association).

Наличие семантической карты помимо помощи в локализации также позволяет решать ряд других задач в робототехнике. Например, задачи в области планирования [27], которые опираются на задачи человеко-машинного взаимодействия, в том числе и на естественном [22] языке.

Структура работы такова: в разделе 2 приведен обзор и анализ методов семантитческой локализации, в том числе с использованием графов для выделения геометрических особенностей, а также нейросетевых моделей для создания векторных представлений изображений и определения визуального сходства объектов. В разделе 3 приведена постановка задачи распознавания сцен и предложены два метода, основанные на комбинации выделения геометрических особенностей и визу-

ального сходства объектов. Раздел 4 посвящен порядку получения данных для тестирования исследуемых методов распознавания сцен и оценки их эффективности. Завершает работу раздел 5, в котором приводятся результаты проведенных экспериментов и их обсуждение.

2. Обзор работ по семантической визуальной локализации

Данная работа основана на использовании семантической карты – размеченных (с указанием класса и положения) в пространстве объектах. Однако стоить сказать, что несмотря на доминирование данного подхода в визуально-семантической локализации, он не является единственным. Активно исследуется подход локализации в поле ключевых кадров [7], в котором отсутствует семантическая карта в виде объектов. В этом подходе карта – набор компактных представлений окружающей среды, полученной по сенсорам робота. Задача локализации сводится в поиске наиболее похожего «кадра» на текущий и дальнейшее определение смещения между ними. К преимуществам такого подхода можно отнести большее «вовлечение» окружающей среды по сравнению с картой объектов, так как затрагивает еще и их окружение. Однако для данной постановки острее встают вопросы со сменой ракурса, процедурой картирования и модификации карты.

Задача построения семантической карты активно исследуется в наши дни [28]. Как упоминалось, задача сопоставления объектов решается по-разному для задач картирования и глобальной локализации. Для задач картирования преимущественно используются оптимизационные алгоритмы, примененные к данным продолжительного движения робота. Существуют и другие подходы, основанные на разных математических методах (например логический вывод, формальные грамматики), однако в этой работе будет сделан акцент на методах, использующих графовые представления, как наиболее популярном.

2.1. ГРАФОВЫЕ ПОДХОДЫ В ЛОКАЛИЗАЦИИ

Подходы на основе представления данных в виде графа привлекательны в силу своей естественности и наработанного математического аппарата. Обычно как наблюдаемая сцена, так и семантическая карта представляются в виде графа, вершины которого соответствуют наблюдаемым объектам, а рёбра – отношениям между объектами. Тут надо отметить, что нет единого подхода к тому, как формировать подобные отношения, однако так или иначе они формируются на основе евклидового расстояния между объектами.

В задачах картирования (в том числе в рамках SLAM) активно применяется семантическая информация для решения подзадачи закольцовывания (loop closing) – определения положения, в котором робот уже был, для исправления накопившихся ошибок. Этой тематике посвящен ряд работ, в которых были использованы графы [13, 20, 30]. В данной постановке требуется найти соответствие между двумя кадрами, на которых содержатся объекты.

Задача же глобальной локализации, рассматриваемая в данной работе, подразумевает поиск соответствия между одним кадром и общей картой, что не позволяет использовать наработки из [13, 20, 30], однако такой постановке также посвящен ряд работ. Например, в работах [8, 15] используется метод случайных дескрипторов для сравнения графов, а в работах [4, 16] используется поиск наибольшей клики на графе специального вида, который отражает соответствия между двумя исходными графами. Этот же подход применяется в работе [24] наряду с другими техниками, в том числе с использованием понятия изоморфизма графов.

Автором данной работы также был предложен метод распознавания сцен, в основе которого лежит поиск изоморфного подграфа [2]. Этот же метод был расширен и для работы в условиях неопределённости [3], включающих в себя ошибки первого и второго рода в системах распознавания объектов, а также высокие погрешности в локализации объектов. Метод сравнивался с методами на основе случайных дескрипторов и поиска максимальной клики и показал своё преимущество в точности работы.

Однако все рассмотренные работы опираются исключительно на геометрические особенности наблюдаемой сцены. Любые два отдельно взятых объекта одного класса номинально идентичны друг другу с точки зрения работы методов. В то же время применение моделей, генерирующих векторное представление изображения, может существенно повысить качество работы методов распознавания сцен.

2.2. МОДЕЛИ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ ИЗОБРАЖЕНИЙ

Нейросетевые модели [25] позволяют из изображения получить вектор чисел, который обладает таким свойством, что векторы для визуально похожих объектов близки по некоторой метрике (часто – косинусному расстоянию), а для непохожих далеки. Такой подход позволяет сравнивать различные изображения, например, он активно используется для распознавания лиц. Одним из ограничений данного подхода являлось то, что получить универсальную модель для произвольных визуальных данных было проблематично и они не обладали достаточной точностью. Однако определённый прорыв в этом направлении был с введением архитектуры CLIP [21]. CLIP является примером языково-визуальной модели и одновременно обучается предсказывать векторы как для изображения, так и для текста, описывающего данное изображение. Это привело к результату, что такая модель очень робастна по отношению к произвольным визуальным данным, что позволяет её применять и в задаче локализации по естественным ориентирам.

Работа [18] использует CLIP для задачи определения типа помещения. На первом этапе при помощи CLIP определяются объекты на входном изображении путем сравнения его вектора с текстовыми запросами, содержащими название объекта. Далее при помощи большой языковой модели по распознанным объектам делается предположение о типе помещения. На последнем этапе полученный текстовый запрос вида «фото рабочего офиса» сравнивается при помощи CLIP с изображением с камеры робота. Авторы отмечают преимущество своего подхода в том, что не требуется подготавливать специально обученные на

определённые типы изображений детекторы, ровно, как и не требуется заранее определять типы помещений.

Идея напрямую использовать кодировку отдельных объектов была опробована в работе [17]. Авторами была вручную размечена трёхмерная карта, где каждый объект задавался эллипсоидом и неким текстом, его описывающим, например «деревянный офисный стол», который потом транслируется в вектор. На этапе локализации на входном изображении определяются объекты и по ним также строятся векторы с помощью CLIP. Далее происходит процедура поиска лучших соответствий между двумя наборами векторов. Авторы никак не используют на этом этапе информацию о геометрическом соотношении объектов. В результате им приходится применять многоитерационный алгоритм определения положения камеры на основе полученных кандидатов. На не очень больших картах, размером с один рабочий кабинет, предложенный метод показывает приемлемые результаты. Авторы также планируют распространить свой подход на решение задачи закольцовывания.

Таким образом, напрашивается соединение методов, работающих по «геометрическому» принципу, и методов, позволяющих определить визуальное сходство объектов путем построения их векторов.

3. Предлагаемые подходы

3.1. ПОСТАНОВКА ЗАДАЧИ

Задача распознавания сцен представляет собой поиск поэлементного соответствия между набором объектов S, что наблюдает робот посредством своих сенсоров (сцена), и полным набором объектов M – семантической картой. Каждый объект o характеризуется своим классом (k), положением (p) и визуальным образом (I):

$$S = \{o = \langle k, p, I \rangle\}. \quad (1)$$

Визуальный образ I может представляться как одним, так и несколькими изображениями объекта. Требуется найти такое соответствие между S и M, которое бы лучшим образом отража-

ло как взаиморасположение объектов, так и их визуальное сходство, при этом сохраняя класс объекта k.

3.2. ОПРЕДЕЛЕНИЕ ОТНОШЕНИЙ

Как уже упоминалось, данная работа будет использовать графовую абстракцию для описания множеств сцены и карты. Для формирования графа и добавления в него рёбер требуется определить так называемые отношения между объектами. В большинстве случаев делается это на основе расчёта евклидового расстояния между объектами и дальнейшей его интерпритацией. Однако на семантической карте объект обычно характеризуется положением своего центра, а на наблюдаемой сцене расстояние можно определить с помощью сенсоров, и зачастую это расстояние до видимой поверхности объекта, что часто не совпадает с физическим центром. Развитие систем трёхмерной детекции [31] по данным лидара отчасти решает такие проблемы, но не для всех конфигураций сенсоров такие методы доступны. Более классический подход – это определение объектов по двумерным изображениям и дальнейшее определение расстояния до него по карте глубины или облаку точек. При таком подходе расстояния между объектами на карте и сцене будут отличаться друг от друга. В связи с этим требуется определить способ сравнения двух расстояний между объектами в разных множествах с учётом описанной специфики.

Чтобы оценить погрешность опредления расстояния, полученного таким способом, предлагается следующий подход. Положение каждого объекта в локальных координатах робота задаётся двумя величинами: углом на объект φ, для определения которого достаточно лишь изображения и параметров камеры, и расстояние до объекта r, получаемого при помощи сенсора глубины. Для определения расстояния с помощью сенсора глубины, формирующего облако точек, требуется совершить перевод этого облака в пространство изображения, а далее выделить те точки, что относятся к объекту. Для определения точек используется обрамляющий прямоугольник объекта или сегментационная маска, позволяющая отсеять лишние точки. Для определения расстояния обычно используется усреднение полученных точек, обычно в виде медианы, так как помимо усредне-

ния требуется еще отбросить точки других объектов, случайно попавших в маску. Но такой подход может плохо работать для не сплошных и частично прозрачных объектов, где попадает слишком много лишних точек. Также в зависимости от формы объекта и ракурса, с которого объект наблюдается, медиана может давать разное расстояние. Из-за этого имеет смысл пытаться определить кратчайшее расстояние до объекта и переводить его в центр объекта, зная примерные размеры данного объекта. Так как шумы за счет неточного выделения маски всё еще могут быть, то рекомендуется брать не минимальное расстояние, а отсекать выбросы, например рассчитывая первый квартиль всех расстояний до объекта.

Получив расстояние до ближайшей точки объекта описанным способом, предлагается перевести его в расстояние до центра объекта, основываясь на знании о физических размерах тех или иных объектов. На заранее размеченных данных возможно рассчитать среднюю ошибку *dr* между расстоянием от робота до центра объекта и расстоянием до ближайшей точки, а также среднеквадратичное отклонение этой величины σ. Данные величины должны быть рассчитаны для каждого класса объектов. Имея данные значения для расстояния и угла на объект, можно оценить погрешность расстояния между двумя объектами, которое в свою очередь рассчитывается как

(2) $$D(o_1, o_2) = \sqrt{(r_1 + dr_1)^2 + (r_2 + dr_2)^2 - 2(r_1 + dr_2)(r_1 + dr_2)\cos(\varphi_1 - \varphi_2)}.$$

Для оценки погрешности данного значения требуется определить параметры эллипса рассеивания для каждого объекта – σr и $\sigma\varphi$, т.е. значения дисперсий данных величин. Согласно [1], дисперсии рассчитываются следующим образом из значений среднеквадратический отклонений:

(3) $$\sigma r = dr^2,$$

(4) $$\sigma\varphi = r^2\left(\left(\frac{1}{\cos d\varphi}\right)^4 d\varphi + \frac{\sin d\varphi}{\cos^2 d\varphi} d\varphi^2\right).$$

Далее, следуя принципу композиции нормальных законов, можно рассчитать результирующий эллипс рассеивания для расстояния между исходными объектами и взять радиус этого эллипса

по направлению рассчитываемого расстояния. Подробные выкладки представлены в [1], параграф 12.8, искомая величина является функцией от указанных, рассчитанных ранее значений:

(5) $\sigma D(o_1, o_2) = F(r_1 + dr_1, r_2 + dr_2, \sigma r_1, \sigma r_2, \sigma\varphi_1, \sigma\varphi_2, D)$.

Полученное значение предлагается использовать как приближение среднеквадратичного отклонения расстояния между объектами, рассчитанного по входным данным (рис. 1).

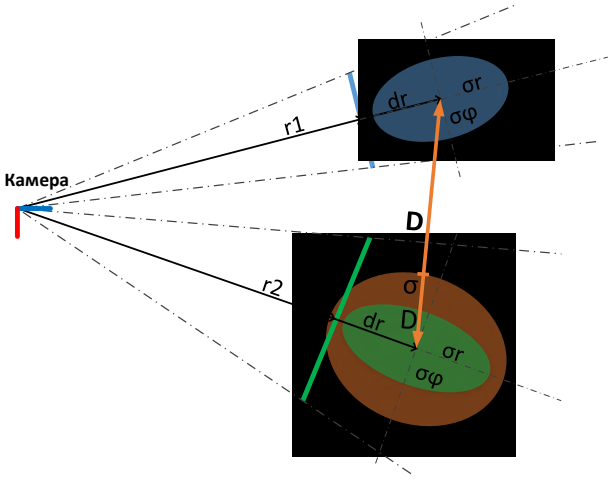

Рис. 1. Оценка погрешности измерения расстояния между двумя объектами. Синий и зелёный эллипсы –распределения положений объектов, оранжевый эллипс – результат композиции двух распределений

Рассчитав расстояние между центрами объектами по сенсорам робота и оценив его погрешность, а также имея расстояние между объектами с карты, предлагается использовать следующую меру сравнения:

(6) $p(r_1, r_2, \sigma) = e^{\frac{1}{2}\left(\frac{r_1 - r_2}{\sigma}\right)^2}$.

Такая мера нормирована на отрезок [0, 1] и учитывает погрешность измерений расстояний.

3.3. МЕТОДЫ НА ОСНОВЕ ПОИСКА ИЗОМОРФНЫХ ПОДГРАФОВ

Первые итерации данного метода излагались в работах [2, 3]. В основе метода используется алгоритм поиска изоморфного подграфа семейства VF [5, 6, 11]. Особенность данного семейства заключается в том, что оно позволяет добавлять дополнительные критерии для сравнения вершин и рёбер, например, позволяет учитывать разметку графов. Учёт разметки естественным образом может быть использован в задаче распознавания сцен. Для семантической карты и наблюдаемой сцены строятся свои графы. Каждая вершина графов соответствует своему объекту в исходных множествах. Это соответствие может выступать в качестве разметки вершин полученных графов:

$$(7)\quad G_S = \langle V_S, E_S \rangle,\ f_{V_S} : V_S \to S, v \in V_S,\ f_{V_S}(v) = o,$$

где G_S – граф сцены, состоящий из множества вершин V_S и множества рёбер E_S, а f_{Vs} – функция разметки, определяемая биекций множества вершин в S. Аналогичным образом происходит для графа карты G_M.

Далее требуется определить рёбра графов и произвести их разметку. Цель разметки – указать близкие друг к другу рёбра, при этом разметка должна быть одинаковой для графа карты и графа сцены. Для того чтобы этого добиться, предлагается сначала рассчитать все попарные расстояния (2) в множестве сцены, разбить их на подмножества по классам объектов (ворота – фонарь, светофор – столб и т.п.), а далее для каждого подмножества произвести кластеризацию. Для кластеризации предлагается воспользоваться методами иерархической кластеризации, использующими порог по сходству и разбивающими на заранее неизвестное число кластеров. Критерием кластеризации является расстояние (6) между двумя элементами, удовлетворяющее порогу τ_1:

$$(8)\quad p(D_i, D_j, \sigma D_i + \sigma D_j) < \tau_1.$$

Таким образом разметка каждого ребра e_{ij}, соединяющего вершины v_i и v_j, есть тип отношения t (идентификатор подмножества, построенный по парам классов объектов, соответствующих вершинам) и идентификатор получившегося кластера c:

(9) $f_{E_S}(e_{ij}=(v_i,v_j))=\left\langle t(k^{o_i},k^{o_j}),c\right\rangle.$

Получив кластеры и рассчитав их центроиды $\bar{D}_C$ (среднее расстояние между объектами в кластере) и среднюю ошибку кластера $\sigma\bar{D}_C$, можно выполнить классификацию рёбер в множестве карты. Для этого также считаются попарные расстояния в множестве карты, а далее каждое из них проверяется на принадлежность получившимся кластерам в соответствующем подмножестве отношений в множестве сцены. Для классификации используется та же метрика (6), однако порог для классификации τ_2 предлагается выбрать отличный от порога для кластеризации τ_1 (8):

(10) $p(D_{map},\bar{D}_C,\sigma\bar{D}_C)<\tau_2.$

В таком случае возможна ситуация, в которой отношение с множества карты будет отнесено сразу к нескольким классам из множества сцены. Чтобы это учитывать, на графе карты вводится свой тип разметки рёбер, который ставит ребру в соответствие тип отношения и набор идентификаторов кластеров, которым оно удовлетворяет. Все отношения, которые удовлетворяют хотя бы одному классу, добавляются на граф карты в виде рёбер с указанной разметкой:

(11) $f_{E_M}(e_{ij}=(v_i,v_j))=\left\langle t(k_{o_i},k_{o_j}),\{c\}\right\rangle.$

Чтобы воспользоваться алгоритмом поиска изоморфного подграфа, требуется определить критерии сравнения вершин и рёбер. В случае с вершинами требуется лишь совпадения разметки, т.е. идентификаторов класса объекта:

(12) $v_i\in V_S, v_j\in V_M : k^{f_{V_S}(v_i)}=k^{f_{V_M}(v_j)}\rightarrow v_i\cong v_j.$

В случае с рёбрами два ребра будут считаться идентичным, если у них совпадает тип отношения и кластер ребра сцены присутствует среди набора кластеров ребра карты:

(13) $e_i\in E_S, e_j\in E_M : t^{f_{E_S}(e_i)}=t^{f_{E_M}(e_j)}, c^{f_{E_S}(e_i)}\in\{c\}^{f_{E_M}(e_i)}\rightarrow e_i\cong e_j.$

Реализовав эти критерии, возможно воспользоваться алгоритмом из семейства VF. На выходе алгоритм даст набор соответствий $\{f\}$ между двумя графами удовлетворяющий поставленным условиям и, разумеется, сохраняющих изоморфизм:

(14) $\{f\}, f : S \to M$.

Для каждого решения из $\{f\}$ требуется ввести численную оценку, выделив лучшие из них. Сделать это можно на основе степени совпадения геометрических ограничений и на основе степени совпадения визуальных образов. Как уже говорилось, задача сопоставления визуальных образов I может быть решена с помощью архитектур H формирующих векторное представление изображения h:

(15) $H(I) = h$.

Меры схожести полученных векторов для каждой пары объектов могут быть рассчитаны с использованием косинусного расстояния:

(16) $u(h_1, h_2) = \dfrac{h_1 \cdot h_2}{\|h_1\| \|h_2\|}$,

эта метрика нормирована на отрезок [0, 1]. Также для каждой пары соответствий определить схожесть их расстояний по уже многократно использованной метрике (6). Обе полученные меры – для объектов и отношений – можно скомпоновать следующим образом, получив общую степень уверенности dc для каждого из полученных ответов:

(17) $dc(f) = \dfrac{1}{|E_S|} \sum_{(v_i, v_j) \in E_S} u(H(I^{o_i^S}), H(I^{o_i^M})) \cdot u(H(I^{o_j^S}), H(I^{o_j^M})) \cdot$

$\cdot p(D(o_i^S, o_j^S), D(o_i^M, o_j^M), \sigma D(o_i^S, o_j^S)), o_i^S = f_{V_S}(v_i), o_i^M = f_{V_S}(f(v_i))$.

Однако у предложенного метода распознавания сцен на основе поиска изоморфного подграфа есть существенное ограничение. Из-за ошибок в системах распознавания на сцене могут быть распознаны «лишние» объекты, которые не представлены на семантической карте. Также такая ситуация может произойти ввиду изменчивости окружающей среды. Если на графе сцены присутствуют такие лишние объекты, то метод поиска изоморфных подграфов не найдёт решение. Для этого было предложено расширение метода, которое изложено в работе [3]. Основная идея состоит в добавлении в граф карты «ложных» вершин. Их цель соответствовать лишним объектам на графе сцены, не нарушая изоморфизм. Добавляется фиксированное число

ложных объектов для каждого класса, представленного в сцене. Число объектов зависит от качества системы распознавания и от того, как много ожидается ложных срабатываний на кадрах. Чтобы сохранять изоморфизм, к ложным вершинам проводятся рёбра для всех типов отношений t (9), что присутствуют на графе сцены. Разметка для ложных вершин – это также класс объекта, а разметка для ложных отношений (11) сводится только к его типу, построенному по соединяемым объектам. Критерий сравнения рёбер (13) для ложных элементов учитывает только сравнение типов. При такой модификации графа G_M в ответы $\{f\}$ (14) будут попадать ложные ребра и отношения, для них в результирующем коэффициенте уверенности все функции u и H, аргументом которых является ложный элемент, выбираются равным 0. Это понижает уверенность ответов с ложными элементами, однако позволяет получать решения в ситуациях с ложными распознаваниями. Конечно, такое, *мягкое* расширение метода приводит к увеличению времени исполнения. Общая схема предложенного метода представлена на рис. 2. На схеме используются обозначения S и M (1), G_S и G_M (7), f (9), u (16), p (8), D (2), dc (17).

Для снижения числа расчетов, была предложена модифицированная версия данного метода (рис. 3, на схеме используются те же обозначения, что и для рис. 2). Её отличие от ранее изложенной заключается в том, что вместо разбиения множеств рёбер на кластеры предлагается использовать критерий (10) сразу на этапе определения изоморфных подграфов, встроив в методы семейства VF. С одной стороны, это позволит быть более чувствительным к сравнению рёбер, поскольку они сравниваются напрямую, а не посредством их классов в виде центроидов. Это также потребует подбора только одного параметра τ вместо двух. С другой стороны, может увеличиться время исполнения, так как процедура сравнения требует больше вычислений, что в том числе зависит от выбора параметров методов. Все остальные моменты, включая расчёт коэффициента уверенности (17) и мягкое расширение метода, аналогичны вышеописанному.

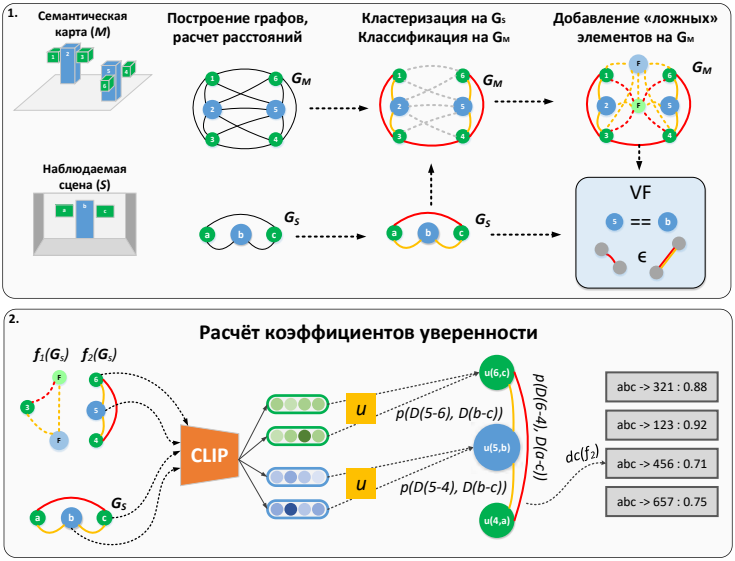

Рис. 2. Схема метода на основе поиска изоморфных подграфов

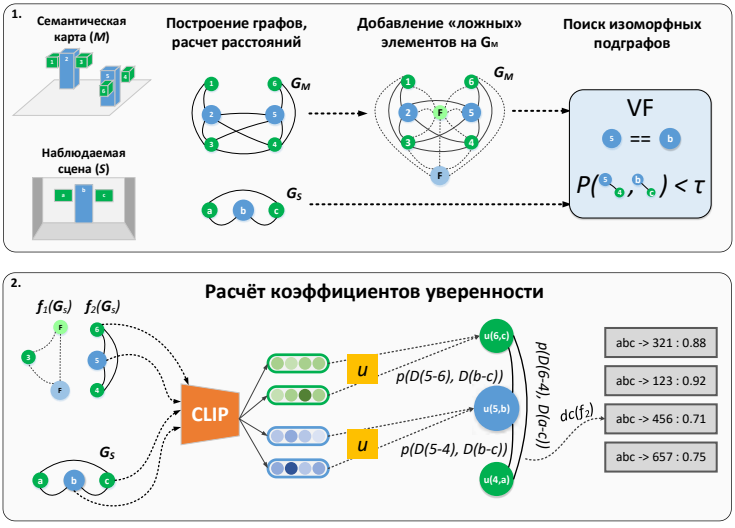

Рис. 3. Схема облегченного метода на основе поиска изоморфного подграфа

3.4. МЕТОД НА ОСНОВЕ ПОИСКА НАИБОЛЬШЕЙ КЛИКИ

Другой метод распознавания сцен основан на процедуре поиска наибольшей клики и изложен в работах [4, 16]. В основе метода лежит построение графа специального вида, где каждая вершина соответствует паре объектов одного типа из множества карты и сцены. Две вершины такого графа соединяются ребром, если разница расстояний между соответствующими исходными объектами в своих множествах не превышает заданного порога. Далее для данного графа осуществляется поиск наибольшей клики, т.е. полносвязного подграфа максимального размера. Такой подграф содержит наибольшее число геометрических парсовпадений между двумя исходными множествами.

В работе [3] была предложена модификация данного метода с использованием предложенной меры оценки сходства расстояний (6) вместо оригинального порога по разнице расстояний. Это добавило дополнительную «чувствительность» и позволило также ввести численную оценку для до этого не различимых вариантов ответов. По проведённым экспериментам метод показал близкие результаты с методом поиска изоморфных подграфов, а также изначально адаптирован к ситуациям «лишних» данных в сцене.

Учитывая плюсы этого метода, в настоящей работе также приводится его расширение с использованием меры схожести объектов, полученной при помощи нейросетевых моделей. На этапе расчета коэффициента уверенности полученных решений добавляется попарная сверка векторизованных представлений h (15) объектов карты и сцены, как это было предложено для метода на основе поиска изоморфного подграфа (17):

$$(18)\ dc(G_C) = \frac{1}{|E_C|} \sum_{\left[\left(o_i^M, o_i^S\right),\left(o_j^M, o_j^S\right)\right] \in E_C} u(H(I^{o_i^S}), H(I^{o_i^M})) \cdot$$

$$u(H(I^{o_j^S}), H(I^{o_j^M})) \cdot p(D(o_i^S, o_j^S), D(o_i^M, o_j^M), \sigma D(o_i^S, o_j^S)),$$

где G_C – клика, а E_C множество ребер данной клики. Схема предложенной модификации представлена на рис. 4. На схеме используются обозначения S и M (1), u (16), p (8), D (2), dc (18).

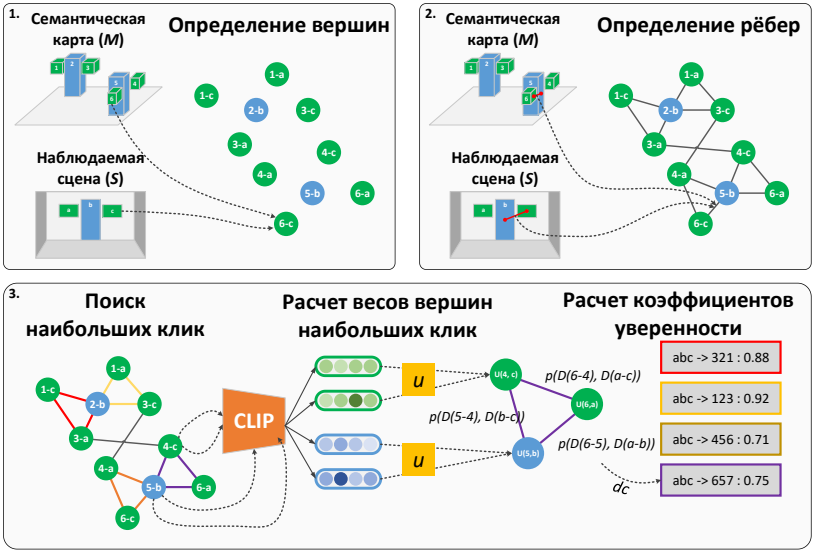

Рис. 4. Схема модифицированного метода на основе поиска наибольших клик

4. Данные для тестирования

Исследования в области распознавания сцен затрудняются отсутствием специализированных открытых наборов данных, позволяющих проводить эксперименты и осуществлять сравнение эффективности методов друг с другом. В приведённых в обзоре работах авторам приходится либо создавать собственные наборы данных, либо адаптировать наборы данных из смежных областей, например, по задаче SLAM. Среди них можно выделить наборы [9, 10, 12]. Эти наборы данных предоставляют траекторию, по которой следовала камера, изображения с неё и ряд дополнительных сенсоров, в том числе определяющих расстояния до объектов, а иногда и некую семантическую разметку. Однако для систем распознавания сцен требуется информация о семантической карте и сопоставленная с ней разметка объектов на изображениях. Поскольку о существовании такого набора данных неизвестно, приходится проводить ручную разметку объектов на карте, что, безусловно, трудоёмко

и не позволяет произвести сравнение между методами, если авторы не публикуют свои дополнительные изменения исходных наборов данных.

В представленной работе также пришлось пойти по пути модификации подходящего набора данных из смежной задачи. Для этих целей интерес представляет набор данных KITTI-360 [14], который, в отличие от своего популярного предшественника KITTY [9], обладает более подробной семантической разметкой и расширенной сенсорикой. Набор данных состоит из нескольких независимых проездов (именуемых далее последовательностями), пронумерованных двузначными числами типа 00, 02 и т.п.

4.1. ПОДГОТОВКА KITTI-360

KITTI-360 обладает трёхмерной глобальной разметкой объектов в виде параллелепипедов, которая фактически является готовой семантической картой, необходимой для задачи распознавания сцен. Также в этом наборе данных существует двумерная разметка объектов для задачи семантической сегментации изображений, что является готовой сценой. При этом для некоторых типов объектов имеется разметка экземплярной сегментации, т.е. разделение на отдельные объекты. Там, где произведена экземплярная сегментация, также выполнена глобальная привязка к идентификатору объекта, т.е. он сохраняется от кадра к кадру для одного и того же объекта. Однако эта глобальная идентификация никак не связана с упомянутой ранее трёхмерной глобальной разметкой объектов, что не позволяет использовать этот набор для задачи распознавания сцен в первозданном виде. Связь в идентификаторах объектов между картой и сценой нужна для возможности оценить точность предлагаемых методов распознавания сцен. Для решения данной проблемы и установки этого соответствия была реализована процедура сопоставления (рис. 5). Стоит отметить, что предлагаемая процедура была выполнена для ограниченного набора объектов. Так как предлагаемые методы работают со статическими и компактными в пространстве объектами, то были исключены все объекты, помеченные как динамические, а также протяженные объекты классов «дорога», «тротуар», «ограждения», «забор» и другие.

Также был удалён класс «растительность», так как помимо подходящих отдельно стоящих деревьев содержит и протяженный кустарник.

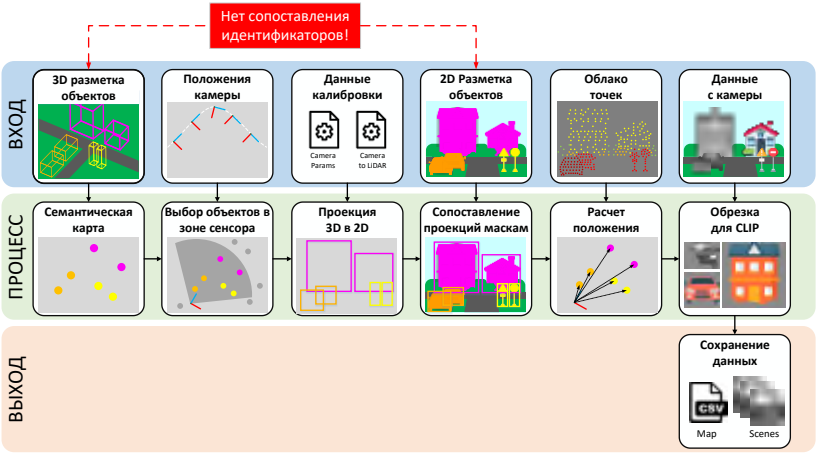

Рис. 5. Схема подготовки набора данных KITTI-360 для задачи распознавания сцен

На первом шаге трёхмерная разметка объектов переводится в семантическую карту посредством расчёта центров этих объектов. Далее для каждого положения камеры определяются видимые объекты, на основании выбранной дальности и угла раствора используемой камеры. Параллелепипеды этих объектов с трехмерной разметки далее проецируются в плоскость изображения и описываются прямоугольниками со сторонами параллельными границам изображения.

Для объектов, для которых есть экземплярная сегментация, происходит сопоставление с полученными прямоугольниками при помощи классического алгоритма решения задачи о назначениях Куна – Манкреса (Венгерский алгоритм) на основе критерия совпадения площадей этого прямоугольника и маски объекта. Для объектов без экземплярной сегментации общая маска класса объектов обрезается по полученному прямоугольнику. Полученные в результате двумерные объекты имеют общие с семантической картой идентификаторы, что позволяет чис-

ленно оценить качество сопоставления в задаче распознавания сцен.

На основе полученной маски определяется также расстояние до объектов по данным лидара. В данном подходе определяются точки лидара, попавшие в маску объекта, фильтруются точки с низкой интенсивностью. Далее несколькими методами рассчитывается расстояние до объекта: определение ближайшей точки, среднее, медианное среднее, первый квартиль, а также взвешенное среднее с учетом интенсивности точек. На основе посчитанного сопоставления был проведен анализ ошибок для каждого из предложенных методов, на рис. 6 приведены результаты для некоторых объектов.

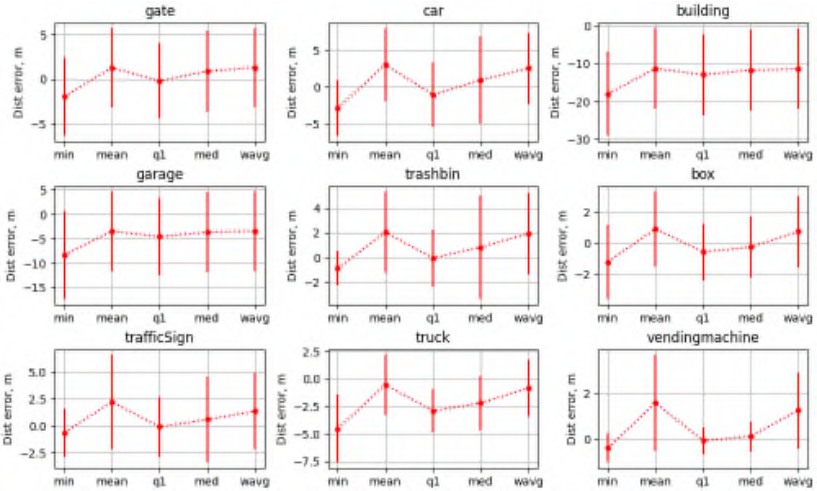

Рис. 6. Значения среднего и среднеквадратичного отклонения для каждого способа расчёта ошибки определения расстояния для некоторых типов объектов из последовательности 00

Медианное расстояние показывает меньшую ошибку, однако у первого квартиля меньше среднеквадратичное отклонение, поэтому далее в работе во всех экспериментах использовался данный метод расчёта расстояния до объектов. Помимо расстояния до объектов определяются углы на него, используя центр обрамляющего прямоугольника, который описывает полученную маску объекта. Для углов также происходит расчёт ошибок,

однако ввиду симметричности средняя ошибка выбирается равной нулю.

Также на основе обрамляющих прямоугольников и маски подготавливается изображение для входа на графический кодировщик CLIP [32] . Все пиксели обрамляющего прямоугольника, не попавшие в маску, закрашиваются нейтральным серым цветом, так как случайно попавшие другие объекты могут дать свой отрицательный вклад в результирующий вектор. Для семантической карты вектор объекта рассчитывается как среднее среди всех векторов этого объекта со всех сцен.

Объекты, попавшие в сцену, но имеющие малую площадь маски или малое число точек лидара, удаляются как ненадёжные. На финальном шаге полученная семантическая карта очищается от объектов, которые не были замечены ни на одной из сцен. Полученные таким образом данные, а также скрипт, их генерирующий, доступны онлайн [33].

Набор KITTI-360 включает себя девять разных последовательностей данных, отличающихся продолжительностью и размером карты, окружением (город\пригород) и характером движения (возвращение в ту же точку, движение в обратную сторону и т.п.). Описанная процедура сопоставления семантических данных была выполнена для каждой из последовательностей.

4.2. РАЗБИЕНИЕ НА ПОДКАРТЫ

Предлагаемые алгоритмы, описанные в работе, так или иначе основаны на методах из теории графов, которые традиционно относятся к классу NP-полных задач; это касается как поиска максимальной клики, так и поиска изоморфных подграфов. Ввиду экспоненциальной зависимости между временем исполнения и количеством объектов на исходном графе целесообразно будет разбить исходную карту на кластеры (подкарты) и производить распознавания сцен на них. Также, имея подкарты, есть возможность обрабатывать их параллельно, что тоже ускорит время исполнения. Проблемой является наличие сцен, объекты которых будут иметь истинное соответствие в разных подкартах, что может привести к падению точности вблизи границ карт. Чтобы этого избежать, предлагается разбивать на подкар-

ты с пересечениями такого размера, чтобы объекты сцены целиком попадали в одну из подкарт.

Для реализации такого разбиения выполняются следующие шаги (рис. 7):

1. Производится начальная иерархическая кластеризация объектов карты. Число кластеров определяется на основе желаемого размера подкарты. Определяются центроиды кластеров.

2. Для каждого кластера рассчитываются диаграммы направленности. С некоторым шагом вокруг центроида определяются секторы и для каждого сектора выбирается наиболее удаленный от центроида объект и рассчитывается расстояние до него.

3. Далее все полученные расстояния увеличиваются на рабочую дальность сенсора дистанции, например, лидара или стереокамеры.

4. Все объекты других начальных кластеров, попавшие в расширенный сектор, добавляются к текущему.

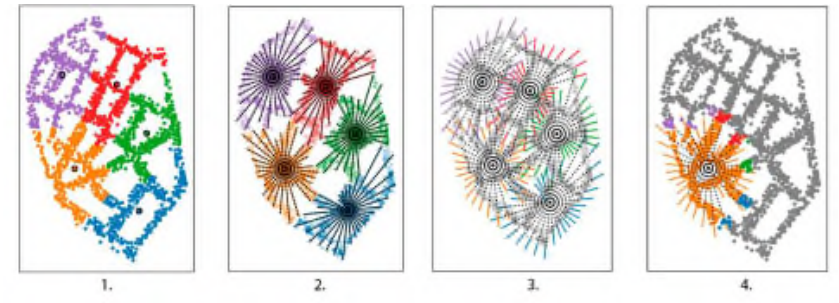

Рис. 7. Пример разделения на подкраты последовательности 09. 1 – начальная кластеризация и определение центроидов; 2 – построение диаграмм направленности; 3 – расширение диаграмм направленности; 4 – добавление новых объектов к одному из кластеров

5. Эксперименты

5.1. МЕТОДЫ И КРИТЕРИИ ТОЧНОСТИ

На полученном наборе данных сравнивались следующие методы:

– оригинальный метод на основе поиска изоморфного подграфа (*sig*), описанный в работе [3], и предлагаемое расширение метода с использованием CLIP-архитектуры (*sig_clip*);

– облегчённый вариант метода *sig_clip* (*sig_lite_clip*) и его мягкое расширение с одним ложным объектом каждого типа (*sig_lite_soft1_clip*);

– методы на основе поиска максимальной клики: реализация, выполненная по оригинальной работе [4] (*mc*), предлагаемая модификация из работы [3] (*mc+*) и предлагаемая модификация из текущей работы с использованием CLIP (*mc+_clip*);

– метод без использования геометрической информации CLIP-Loc из работы [17] (*clip_loc*). Для данного метода был введен критерий оценки решений как среднее сходства полученных соотвествий (16).

Все сторонние исследуемые методы были реализованы автором лично по описаниям из соответствующих работ ввиду отсутствия их открытой реализации. Реализация методов выполнялась на языке Python и не является максимально оптимизованной с точки зрения времени исполнения, однако все действия с графами, являющиеся достаточно ресурсоёмкими процедурами, были взяты или реализованы с помощью библиотеки *igraph* [34].

Основной задачей данной работы было исследовать разницу в подходах к задаче распознавания сцен, т.е. замерить точность сопоставления объектов исходной сцены к карте, хотя обычно в литературе исследуется более общая задача и производится процедура локализации. Но поскольку в этом случае работают сразу два метода – распознавания сцен и локализации по видимым объектам, то вклад каждого метода в отдельности сложно оценить.

Зная из набора данных истинное соотношение f': $S \to M$, в ходе проведения экспериментов замерялись следующие критерии для полученного набора соотвествий $\{f\}$ (14), выдаваемого исследуемыми методами:

– доля лучшего ответа (rate) – отношение правильно найденных соотношений между сценой и картой к размеру сце-

ны в ответе с наивысшем коэффициентом уверенности (17), (18); иллюстрирует число верных соотвествий в лучшем ответе;

– счёт (score) – величина, обратно пропорциональная номеру полностью верного ответа среди всех полученных ответов, упорядоченных по убыванию коэффициента уверенности (17), (18). В случае отсутствия верного ответа равен 0; иллюстрирует позицию верного ответа в общем множестве;

– степень (degree) ответа – четыре величины, показывающие распределение ответов между категориями *полностью верный ответ* (super true) – ответ с единичным счётом; *имеется верный ответ* (has true) – ответ с не нулевым и не единичным счётом; *нет результата* (no result) – нулевой счёт при отсутствии ответов, и *нет верного результата* (wrong) – нулевой счёт при наличии ответов, т.е. полностью верного ответа нет в списке ответов, иллюстрирует общее распределение полученных ответов;

– время работы методов. В поставленных экспериментах время работы было ограничено сверху 120 секундами.

Ограничение по времени было выбрано из тех соображений, что задача глобальной локализации достаточно редко производится: в начале работы и в проблемные моменты, поэтому можно выделить намного больше времени, чем обычно требуется от непрерывных систем локализации, вынужденных работать с субсекундным временем. Также данное органичение было выбрано ввиду большого количества данных, требующих обработки. Если метод превышал данный порог, то его доля и счёт полагались равными нулю, а степень принимала значение *нет результата*.

Данные критерии оценки точности были введены, чтобы показать с разных сторон способность методов находить правильные решения, пусть и не отмеченные как лучшие. Эта способность интересна тем, что лучшие результаты задачи распознавания сцен обычно проходят верификацию дополнительными методами, позволяющими отсеять неверные решения по различным критериям, и введенные метрики позволят определить, какие методы лучше использовать и сколько гипотез стоит выбрать.

5.2. ПОСТАНОВКА ЭКСПЕРИМЕНТА

Поскольку все методы имеют ряд настраиваемых параметров, влияющих как на точность, так и на время исполнения, то проводилась оптимизация параметров на отобранном наборе из 100 сцен средством *optuna* [35]. Отобранный набор из последовательности 02 содержал сцены размером от двух до двенадцати объектов в равной пропорции, т.е. по десять каждого вида. Оптимизация длилась 100 шагов и максимизировался введёный критерий доли (rate) полученного ответа и одновременно минимизировалось время работы алгоритма. Полученные параметры методов использовались во всех дальнейших экспериментах.

В поставленных экспериментах объекты «строение» и «гараж» были исключены, так как используемый метод локализации объектов даёт слишком высокую погрешность для данных классов (рис. 6). Важно отметить, что были оставлены все типы припаркованных транспортных средств («автомобиль», «грузовик», «велосипед» и т.п.) с целью увеличения среднего числа объектов на сценах. Конечно, в реальном использовании невозможно опираться на эти типы объектов при составлении семантических карт на долгосрочный период, однако в данном исследовании в первую очередь ставилась цель проверки работоспособности предложенных методов на реальных визуальных данных. В реальном использовании можно увеличить число объектов в сцене за счет таких, как «окно», «дверь», «канализационный люк», «дерево» и т.п.

Расчёты производились на компьютере с процессором AMD Ryzen7 2700X Eight-Core 3.70GHz для всех исследуемых методов и для каждой последовательности из набора данных KITTI-360 в отдельности.

5.3. РЕЗУЛЬТАТЫ

Так как размер семантической карты (в числе объектов) для каждой последовательности свой, то проведено исследование зависимости выбранных параметров от этого числа. На графиках ниже представлена доля (рис. 8) и счёт (рис. 9) для каждого метода в зависимости от числа объектов на карте.

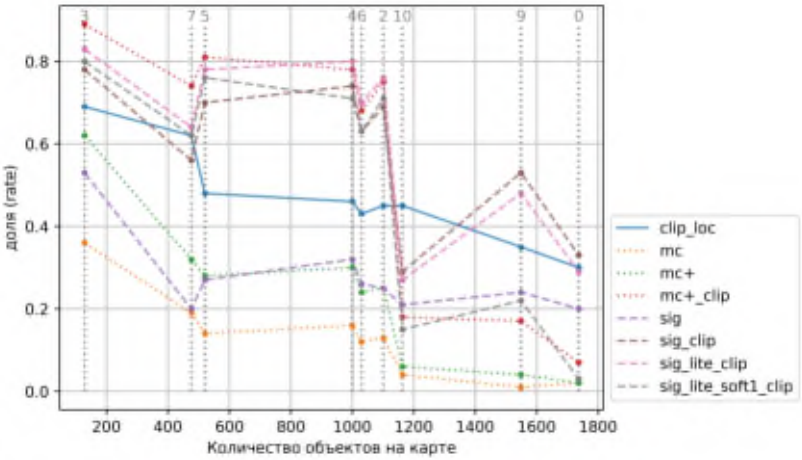

Рис. 8. Средняя для каждой последовательности доля от числа объектов на карте, числами вверху графика отмечен номер последовательности

За исключением некоторых случаев видно убывание обоих критериев точности при увеличении числа объектов. Очевидно, что чем больше карта, тем больше вероятность, что на ней будут похожие сцены, уменьшающие точность. Также размер карты влияет на время исполнения методов, что также важно при введенных ограничениях. Временные показатели приведены на рис. 10.

Общая тенденция на графиках точности (рис. 8, рис. 9) показывает, что методы с использованием визуального сходства всегда превышают свои же реализации, но без использования визуального сходства, причём значительно.

Методы семейства *mc* показывают лучшие результаты по сравнению с аналогичными им (с точки зрения использования\не использования визуального сходства) методами семейства *sig*, за исключением трёх самых больших карт. Это также может быть следствием того, что с выбранным ограничением по времени методы семейства *sig* быстрее (рис. 10), т.е. методы *mc* чаще не успевают обработать данные. Данная скорость, конечно, может быть результатом конкретной реализации используемых методов теории графов.

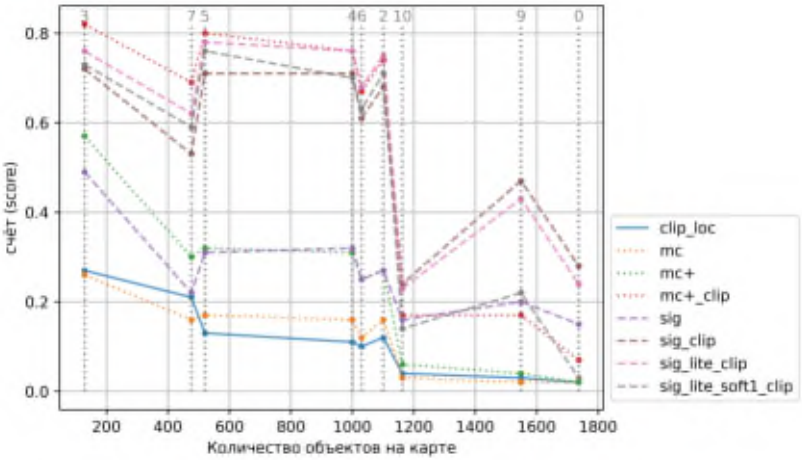

Рис. 9. Средний для каждой последовательности счёт от числа объектов на карте, числами сверху графика отмечен номер последовательности

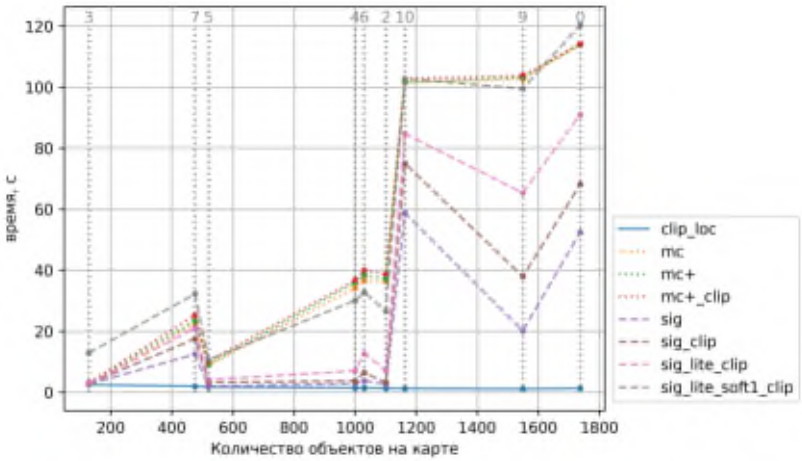

Рис. 10. Среднее время исполнения методов для каждой последовательности в зависимости от числа объектов на карте, числом сверху отмечен номер последовательности

Методы семейства *mc* также примечательны тем, что каждая последовательная их модификация (*mc+* и *mc+_clip*) увеличивает точность по обоим параметрам, при этом практически не увеличивая время.

Последовательные модификации метода *sig* также увеличивают точность за рядом нескольких исключений. Добавление визуального сходства с помощью CLIP также дало существенный скачок в точности. Облегченный вариант метода (*sig_lite_clip*) увеличивает точность и уменьшает скорость. Мягкое расширение метода (*sig_lite_soft1_clip*) уменьшает точность, вероятно также за счёт увеличения времени исполнения.

Отдельного обсуждения заслуживает метод *clip_loc* [17] поскольку он единственный из исследуемых, что использует только визуальное сходство и применяется несколько иначе, чем в оригинальной работе, и к значительно большему числу объектов на карте. По критерию доли (рис. 8) этот метод показывает весьма приемлемые результаты, опережающие все подходы только на геометрическом сравнении и иногда превосходящие даже комбинированные. Однако по критерию счёта метод показывает одни из самых низких показателей. Это говорит о том, что в целом метод в среднем верно распознаёт около половины объектов в сцене, однако верное распознавание сцены целиком для него достаточно редкая ситуация. При этом он показывает наилучшие показатели по времени, очень слабо зависящие от числа объектов на карте.

На графике среднего времени исполнения (рис. 10) примечателен скачок для последовательности 07 при сравнимом с последовательностью 05 числом объектов для всех методов, кроме *clip_loc*. Этот же скачок в виде резкого падения наблюдается и на графиках точности (рис. 8, рис. 9). Несмотря на близкое число объектов и схожую вытянутую топологию, карты этих последовательностей разнятся по наполнению, последовательность 07 содержит много однотипных объектов типа smallPole (рис. 11) – это столбики вдоль дорог, установленные с одинаковым интервалом, а так как они произведены заводским способом, то и выглядят одинаково. Поэтому методы, основывающиеся на геометрических особенностях и визуальном сходстве,

испытывают проблемы с такими окружениями как по точности, так и по времени.

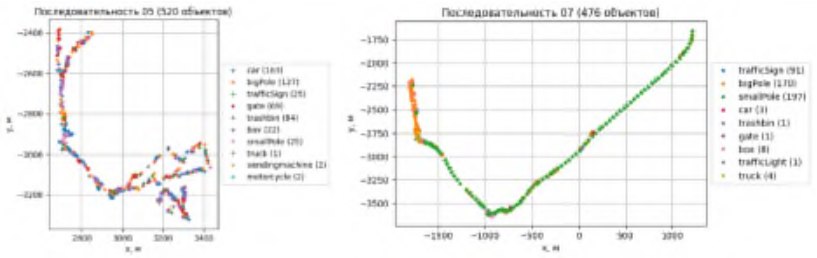

Рис. 11. Карты последовательностей 05 и 07

Были посчитаны эти же критерии для полного набора данных KITTI-360, включающего все последовательности (таблица 1). Указаны средние значения доли (|rate|), счёта (|score|) и времени (|time|), а также их среднеквадратичные ошибки, обозначенные символом σ.

Таблица 1. Численные результаты методов на полном наборе данных KITTI-360. Первый блок содержит метод, основанный на визуальном сходстве, второй – методы на геометрическом сходстве, и третий – методы на обоих критериях

Метод	**|rate|**	**σrate**	**|score|**	**σscore**	**|time|**	**σtime**
clip_loc	0,4	0,28	0,08	0,26	**1,39**	9,63
mc	0,08	0,25	0,09	0,22	66,57	48,58
mc+	0,16	0,35	0,17	0,33	67,39	48,52
sig	0,25	0,41	0,23	0,37	19,72	37,23
sig_clip	0,56	0,47	0,53	0,46	28,4	44,97
mc+_clip	0,46	0,48	0,45	0,48	68,28	48,42
sig_lite_clip	**0,57**	0,46	**0,54**	0,46	41,55	46,5
sig_lite_soft1_clip	0,43	0,48	0,43	0,47	65,3	51,09

Достаточно высокие ошибки объясняются тем, что размер сцены также влияет на скорость и качество распознавания (рис. 12).

Падение точности от размера сцены на рис. 12 также может быть следствием ограничения по времени. Общие показания для степени (degree) представлены на рис. 13.

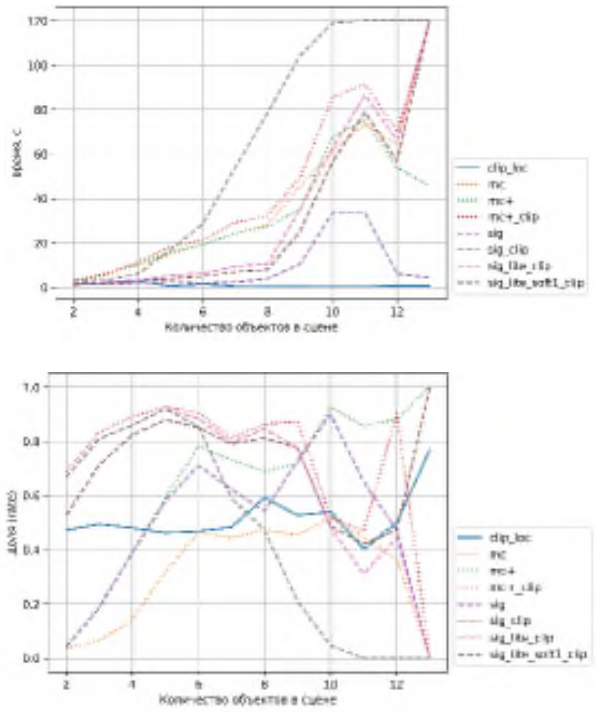

Рис. 12. Графики зависимости времени (сверху) и доли (снизу) для исследуемых методов в зависимости от числа объектов в сцене

Степень методов, изображенная на рис. 13, призвана показать возможность методов находить верный ответ. От этого будет зависеть дальнейшая стратегия по использованию результатов распознавания сцен.

Данный график также показывает, что соединение подходов на основе геометрических особенностей и визуального сходства приводит к наилучшим результатам. Однако можно отметить, что идеальное совпадение лучшего и верного ответа (*super true*) не достигает даже 50% случаев для лучших методов. Немаленький процент наличия верного ответа среди всех полученных (*has true*), а также имеющийся процент неверных решений (*wrong*) говорит о том, что требуется внедрение дополни-

тельных методов, способных оценить качество полученных решений, как, например, упомянутый метод недоопределённой локализации [19]. Такие методы должны обладать возможностью оценивать положение робота на основе нескольких гипотез и определять ошибочные объекты в ответах.

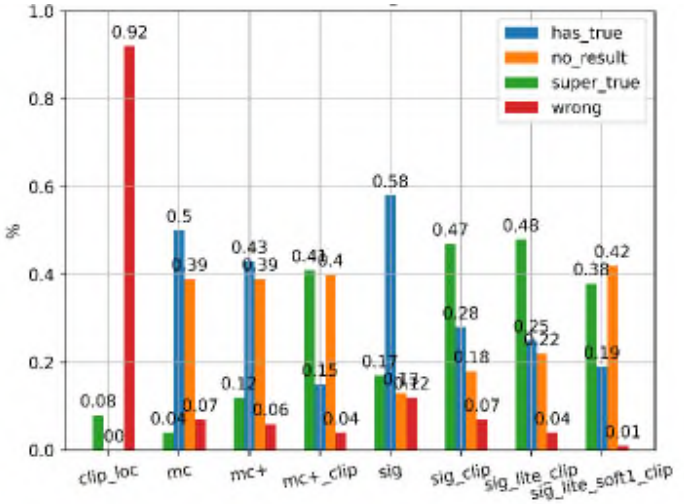

Рис. 13. Степени для исследуемых методов на наборе данных KITTI-360

Также было проведено исследование применимости описанного подхода разбиения на подкарты. Тестировались методы, показавшие лучшие результаты (*mc+_clip* и *sig_lite_clip*) и они же, только примененные к подкартам (*mc+_clip_sm*, *sig_lite_clip_sm*) последовательности 00. Эта последовательность является самой большой по числу объектов и была разделена на 11 подкарт с увеличением общего размера за счёт пересечений на 29%. Несмотря на возможность параллельной обработки подкарт, для чистоты эксперимента расчёты проводились последовательно. Значения точности в виде доли, а также времени исполнения приведены на рис. 14.

У метода *mc+_clip_sm* наблюдается значительный прирост по точности (+285%), выходящий на уровень *sig_lite_clip*, а так-

же ускорение по времени (–20%). У метода *sig_lite_clip_sm* особого изменения в точности не наблюдается, однако также фиксируется ускорение (–14%).

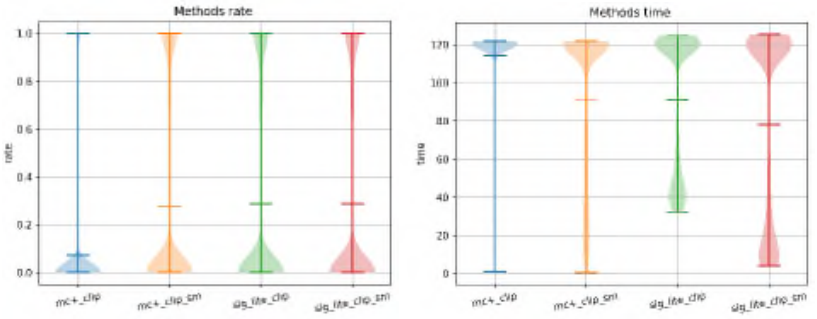

Рис. 14. Сравнение точности (доли) и времени методов работы с полной картой и подкартами

Тут стоит отметить, что в случае, когда любой из методов на подкартах не успевал закончить за отведённое время, фиксировался полученный на этот момент ответ, который мог содержать как реальный ответ, так и ошибочный, но близкий к требуемому. Поэтому на графике степени (рис. 15) также видно увеличение неверных решений (*wrong*) у методов семейства *mc*.

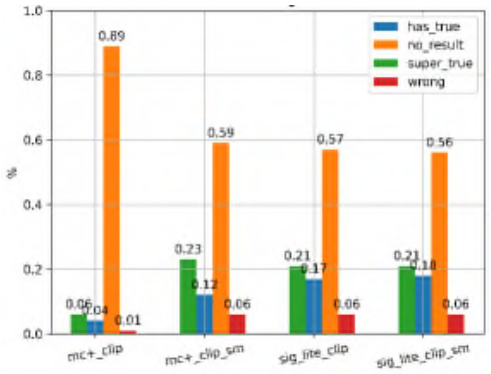

Рис. 15. Сравнение степени методов работы с полной картой и подкартами

У методов семейства *sig* практически отсутствуют изменения в точности. Таким образом, использование разбиения на подкарты потенциально позволит увеличить точность и быстродействие предлагаемых методов даже без учёта параллельного исполнения.

5.4. ОБОБЩЕНИЕ РЕЗУЛЬТАТОВ

По итогам проведённых экспериментов были сделаны следующие заключения:

1. Одновременное использование подходов на основе геометрических особенностей и визуального сходства дают наилучший результат.
2. Увеличение точности достигается путем допустимого увеличения времени исполнения методов, а также требует особого формата семантической карты и предварительного расчёта погрешностей.
3. Метод *sig_lite_clip* показал лучшую точность, однако его «мягкое» расширение *sig_lite_clip1_soft*, адаптированное под сложные ситуации, показывает сравнимые результаты с методом *mc+_clip*, который исходно адаптирован к этим ситуациям. Поэтому выбор «лучшего» метода зависит от условий конкретного применения.
4. Поскольку существует прямая зависимость времени исполнения от размера семантической карты, то была предложена процедура разбиения на подкарты, ускоряющая процессы распознавания сцен.
5. Показано падение производительности и точности в зависимости от размера сцены. Возможно, процедура обрезки сцены до некоторого числа объектов позволит ускорить процесс без потери точности.
6. Помимо зависимости от размеров сцены и карты на исследуемые методы также влияет наполнение их объектами. Множество одинковых и переодично расположенных объектов также является трудностью ввиду отсутсвия геометрических особенностей и визуальных различий.
7. Полученной точности недостаточно для использования результата распознавания сцен без дальнейшей обработки и отсеивания неверных результатов.

6. Заключение

В данной работе приведены два метода, решающие задачу распознавания сцен на основе геометрических особенностей, дополненные технологией визуального сравнения объектов на основе нейросетевой архитектуры CLIP. Результативность полученных методов и их временные характеристики были исследованы на подготовленной модификации реального набора данных KITTI-360. Эта модификация предназначена для тестирования методов распознавания сцен и опубликована в открытом доступе. Проведённые эксперименты показали значительное преимущество предложенных методов в сравнении с методами, которые основываются только на визуальном сходстве или геометрических особенностях. Утверждается, что такой результат получен за счёт использования одновременно обоих подходов. Полученный результат достигается в ущерб времени работы методов, а также требует построение семантической карты, содержащей визуальные образы объектов. Также отмечается, что несмотря на сильное повышение качества задачи распознавания сцен, полученной точности недостаточно, чтобы использовать результат без сторонних методов уточнения результата. Дальнейшими направлениями исследования будут проведение экспериментов совместно с методами оценки результата и использовании систем лидарной локализации.

Литература

1. ВЕНТЦЕЛЬ Е.С. *Теория вероятностей*. – М.: Высшая школа, 1999. – 576 с.
2. МОСКОВСКИЙ А.Д. *Распознавание сцен для задачи глобальной локализации робота* // Труды 34-й Международной научно-технической конференции «Экстремальная робототехника». – 2023. – С. 267–274.
3. МОСКОВСКИЙ А.Д. *Распознавание сцен для задач локализации мобильного робота в условиях неопределённости* // Сборник научных трудов XII Международной научно-практической конференции «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (ИММВ-

2024), Коломна, 14-17 мая 2024 г. В 2-х томах. – 2024. – Т. 2. – С. 255–266.

4. ANKENBAUER J., LUSK P.C., THOMAS A. et al. *Global Localization in Unstructured Environments using Semantic Object Maps Built from Various Viewpoints* // IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS). – IEEE, 2023. – P. 1358–1365.
5. CARLETTI V., FOGGIA P., GRECO A. et al. *VF3-Light: A lightweight subgraph isomorphism algorithm and its experimental evaluation* // Pattern Recognition Letters. – 2019. – Vol. 125. – P. 591–596.
6. CORDELLA L.P., FOGGIA P., SANSONE C. et al. *Performance evaluation of the VF graph matching algorithm* // Proc. of the 10th Int. Conf. on Image Analysis and Processing. – 1999. – P. 1172–1177.
7. GARG S., FISCHER T., MILFORD M. *Where Is Your Place, Visual Place Recognition?* // Proc. of the 13th Int. Joint Conf. on Artificial Intelligence. – 2021. – P. 4416–4425.
8. GAWEL A., DON C. DEL, SIEGWART R. et al. *X-View: Graph-Based Semantic Multi-View Localization* // IEEE Robot. Autom. Lett. – 2018. – Vol. 3, No. 3. – P. 1687–1694.
9. GEIGER A., LENZ P., URTASUN R. *Are we ready for autonomous driving? The KITTI vision benchmark suite* // IEEE Conf. on Computer Vision and Pattern Recognition. – 2012. – P. 3354–3361.
10. HEWITT R.A., BOUKAS E., AZKARATE M. et al. *The Katwijk beach planetary rover dataset* // The Int. Journal of Robotics Research. – 2018. – No. 1(37). – P. 3–12.
11. JÜTTNER A., MADARASI P. VF2++—*An improved subgraph isomorphism algorithm* // *Discrete Applied Mathematics*. – 2018. – Vol. 242. – P. 69–81.
12. KERL C., STURM J., CREMERS D. *Dense visual SLAM for RGB-D cameras* // IEEE Int. Conf. on Intelligent Robots and Systems. – 2013. – P. 2100–2106.
13. KONG X., YANG X., ZHAI G. et al. *Semantic Graph Based Place Recognition for 3D Point Clouds* // IEEE/RSJ Int. Conf. on Intelligent Ro-bots and Systems (IROS). – 2020. – P. 8216–8223.

14. LIAO Y., XIE J., GEIGER A. *KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D* // IEEE Trans. Pattern Anal. Mach. Intell. – 2023. – Vol. 45, No. 3. – P. 3292–3310.
15. LIU Y., PETILLOT Y., LANE D. et al. *Global Localization with Object-Level Semantics and Topology* // Int. Conf. on Robotics and Automation (ICRA). – 2019. – P. 4909–4915.
16. LUSK P.C., FATHIAN K., HOW J.P. *CLIPPER: A Graph-Theoretic Framework for Robust Data Association* // IEEE Int. Conf. on Robotics and Automation (ICRA). – 2021. – P. 13828–13834 .
17. MATSUZAKI S., SUGINO T., TANAKA K. et al. *CLIP-Loc: Multi-modal Landmark Association for Global Localization in Object-based Maps* // IEEE Int. Conf. on Robotics and Automation (ICRA). – 2024. – P. 13673–13679.
18. MIRJALILI R., KRAWEZ M., BURGARD W. *FM-Loc: Using Foundation Models for Improved Vision-based Localization* // IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS). – 2023. – P. 1381–1387.
19. MOSCOWSKY A. *Subdefinite Computations for Reducing the Search Space in Mobile Robot Localization Task* // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2021. – (12948 LNAI) – P. 180–196.
20. QIN C., ZHANG Y., LIU Y. et al. *Semantic loop closure detection based on graph matching in multi-objects scenes* // Journal of Visual Communication and Image Representation. – 2021. – Vol. 76. – P. 103072.
21. RADFORD A., KIM J.W., HALLACY C. et al. *Learning Transferable Visual Models From Natural Language Supervision* // Int. Conf. on Machine Learning. – 2021. – P. 8748–8763.
22. ROVBO M.A., SOROKOUMOV P.S. *Symbolic Control System for a Mobile Robotic Platform Based on Soar Cognitive Architecture* // Smart Electromechanical Systems. – 2022. – P. 259–275.
23. SAHILI A.R., HASSAN S., SAKHRIEH S.M. et al. *A Survey of Visual SLAM Methods* // IEEE Access. – 2023. – Vol. 11. – P. 139643–139677.

24. SHAHEER M., MILLAN-ROMERA J.A., BAVLE H. et al. *Graph-based Global Robot Localization Informing Situational Graphs with Architectural Graphs* // IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS). – 2023. – P. 9155–9162.
25. UEKI K. *Survey of Visual-Semantic Embedding Methods for Zero-Shot Image Retrieval* // 20th IEEE Int. Conf. on Machine Learning and Applications (ICMLA). – 2021. – P. 628–634.
26. YIN H., XU X., LU S. et al. *A Survey on Global LiDAR Localization: Challenges, Advances and Open Problems* // Int. Journal of Computer Vision. – 2024. – No. 8(132). – P. 3139–3171.
27. ZEMSKOVA T., STAROVEROV A., MURAVYEV K. et al. *Interactive Semantic Map Representation for Skill-Based Visual Object Navigation* // IEEE Access. – 2024. – Vol. 12. – P. 44628–44639.
28. ZHANG X.Y., ABD RAHMAN A.H., QAMAR F. *Semantic visual simultaneous localization and mapping (SLAM) using deep learning for dynamic scenes* // PeerJ Computer Science. – 2023. – Vol. 9. – P. e1628.
29. ZHANG Y., SHI P., LI J. *3D LiDAR SLAM: A survey* // The Photogrammetric Record. – 2024. – No. 186(39). – P. 457–517.
30. ZHU Y., MA Y., CHEN L. et al. *GOSMatch: Graph-of-Semantics Matching for Detecting Loop Closures in 3D LiDAR data* // IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS). – 2020. – P. 5151–5157.
31. ZIMMER W., ERCELIK E., ZHOU X. et al. *A Survey of Robust 3D Object Detection Methods in Point Clouds* // arXiv preprint. – arXiv:2204.00106.
32. https://huggingface.co/openai/clip-vit-base-patch32.
33. https://github.com/MoscowskyAnton/scene_recognition_kitti_360.
34. https://python.igraph.org/en/stable/.
35. https://optuna.org/.

SCENE RECOGNITION FOR THE MOBILE ROBOT GLOBAL LOCALIZATION PROBLEM BASED ON IMAGE VECTORIZATION AND GRAPHS APPROACHES

Anton Moscowsky, National Research Ceneter «Kurchatov Institute», Moscow, Head of the group (moscowskyad@yandex.ru)

Abstracat: The paper is devoted to the problem of localization of mobile robots based on visual semantic data. The central element of such a problem is scene recognition task, i.e. searching for a correspondence between observed objects and objects on a semantic map. Paper proposes two methods based on definition of geometric features in the observed scene and searching for them on the map using various graph approaches. The proposed method for determining the relationships between objects, used in both methods, allows taking into account the errors in estimating distances by onboard sensors. In addition to using geometric features, the paper also considers the use of neural network models forming a feature vector based on an image, determining their visual similarity. Visual similarity is used to evaluate and sort the results obtained by the proposed methods based on graph approaches. In addition, the open KITTI-360 dataset was modified to evaluate the accuracy of solving scene recognition problems. Experiments on the resulting dataset demonstrated that the proposed approach, which combines geometric features and visual similarity, significantly increases the accuracy of the considered scene recognition methods. Based on the results of the experiments, some recommendations were formulated for the use of these approaches in practice.

Keywords: mobile robot localaization, scene recognition, image vectorization, graph theory.

УДК 004.896:007.5
ББК 32.813

Статья представлена к публикации
членом редакционной коллегии А.И. Алчиновым.

Поступила в редакцию 12.12.2024.
Опубликована 31.03.2025.