МЕТОДЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СИСТЕМАХ НАВИГАЦИИ МОБИЛЬНЫХ РОБОТОВ

ARTIFICIAL INTELLIGENCE METHODS IN MOBILE ROBOT NAVIGATION SYSTEMS

УДК 004.896 + 007.52 + 004.93'1
DOI: 10.17586/0021-3454-2022-65-3-204-217

МЕТРИКО-СЕМАНТИЧЕСКОЕ КАРТИРОВАНИЕ НА ОСНОВЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ СИСТЕМ АВТОНОМНОЙ НАВИГАЦИИ В ПОМЕЩЕНИЯХ

А. Р. Беркаев[1*], М. Мохрат[1], А. М. Бурков[2], С. А. Колюбин[1]

[1] *Университет ИТМО, Санкт-Петербург, Россия*
[*] *berkaevamiran@mail.ru*

[2] *Сбербанк, Москва, Россия*

Аннотация. Представлены результаты исследования, направленного на разработку интеллектуальной автономной навигационной системы для складской и офисной логистики с использованием глубоких нейронных сетей. Проанализированы современные и наиболее универсальные средства для получения карт глубин и семантической сегментации данных на изображениях в различных средах. Проведен сравнительный анализ карт глубин, формируемых RGB-D-камерой, а также с помощью нейросетевых алгоритмов и модифицированного алгоритма Хиршмюллера. Результаты тестирования, проведенного на специально подготовленном наборе данных, снятых в офисном пространстве, демонстрируют, что предложенное решение превосходит альтернативные по точности и позволяет сократить затраты вычислительных ресурсов.

Ключевые слова: *семантическая сегментация, карты глубин, одновременная локализация и картирование, метрико-семантическая карта, мобильный робот, логистика, глубокие нейронные сети, оценка глубины, интеллектуальные системы*

Ссылка для цитирования: *Беркаев А. Р., Мохрат М., Бурков А. М., Колюбин С. А.* Метрико-семантическое картирование на основе глубоких нейронных сетей для систем автономной навигации в помещениях // Изв. вузов. Приборостроение. 2022. Т. 65, № 3. С. 204—217. DOI: 10.17586/0021-3454-2022-65-3-204-217.

METRIC-SEMANTIC MAPPING BASED ON DEEP NEURAL NETWORKS FOR SYSTEMS OF INDOOR AUTONOMOUS NAVIGATION

A. R. Berkaev [1*], M. Mohrat[1], A. M. Burkov[2], S. A. Kolyubin[1]

[1] *ITMO University, St. Petersburg, Russia*
[*] *berkaevamiran@mail.ru*

[2] *Sberbank, Moscow, Russia*

Abstract. Results of a study aimed at developing an intelligent autonomous navigation system for warehouse and office logistics using deep neural networks, are presented. The modern and most versatile tools for depth maps retrieval and semantic data segmentation on images in different environments are analyzed. A comparison of depth maps retrieved hardware from RGB-D camera, neural network algorithms, and a modified Hirschmuller algorithm is carried out. Results of testing performed with a specially prepared dataset shot in an office space, including many complex objects such as glass, mirrors, and multiple light sources demonstrate that the proposed solution outperforms the alternatives in accuracy and uses fewer computational resources in the process.

Keywords: segmentation, depth maps, simultaneous localization and mapping, metric-semantic map, mobile robot, logistics, deep neural network, depth estimation, intelligent systems

For citation: Berkaev A. R., Mohrat M., Burkov A. M., Kolyubin S. A. Metric-semantic mapping based on deep neural networks for systems of indoor autonomous navigation. *Journal of Instrument Engineering.* 2022. Vol. 65, N 3. P. 204—217 (in Russian). DOI: 10.17586/0021-3454-2022-65-3-204-217.

Введение. Совершенствование методов одновременной локализации и картирования имеет важное значение для функционирования мобильных роботов, способных автономно перемещаться в помещениях различного назначения. Многие известные методы разработаны применительно к условиям конкретного пространства, однако все чаще возникает потребность в универсальных подходах. Это связано с появлением новых сценариев для мобильных роботов, например в сфере логистики: роботу-курьеру необходимо забрать посылку со склада, преодолеть определенное расстояние во внешнем пространстве, а затем снова заехать в помещение для доставки адресату. Поэтому настоящая работа сосредоточена на выявлении универсальных и эффективных методов метрико-семантической одновременной локализации и картирования (Simultaneous Localization and Mapping — SLAM) на основе глубоких нейронных сетей (DNN).

Для создания максимально эффективной метрико-семантической SLAM-системы были протестированы различные методы генерации облаков точек. Одним из наиболее частых решений является использование специальных датчиков с активным зрением, таких как лидар, или датчиков со структурированным светом, таких как RGB-D-камеры.

Использование камер структурированного света — популярное решение для работы в помещении или на открытом воздухе. Однако, как и все датчики, они имеют не только преимущества, но и недостатки. Одно из преимуществ — возможность использования вне помещений, даже при сильном влиянии солнечного света. С другой стороны, наличие высокого уровня шумов обусловливает невысокое качество информации о пространственной глубине: на изображениях с объектами, находящимися на значительном расстоянии от камеры, теряется большая часть деталей или наблюдается сильное искажение. Такие камеры используются для расчета глубины проецирования инфракрасных лучей на поверхность объектов. В связи с этим облака точек, полученные для прозрачных объектов, недостаточно информативны. Кроме того, как правило, получаемые таким образом облака точек имеют невысокий коэффициент заполнения, т.е. процент вокселов, имеющих ненулевые значения глубины.

Следовательно, представляется актуальной разработка альтернативных решений для создания высококачественных семантически аннотированных 3D-карт окружающего пространства с использованием возможностей искусственного интеллекта.

Основное значение и особенность настоящей статьи заключаются в системном анализе существующих подходов и разработке интегрированной метрико-семантической SLAM-системы, использующей современные методы на основе DNN для программного восстановления карт глубин и семантической сегментации и аннотирования получаемых трехмерных карт. Предлагаемое решение базируется на алгоритмах, обеспечивающих наилучшие показатели по качеству (точности и плотности заполнения облаков точек) и скорости работы.

Обзор известных подходов. Построение семантически аннотированной карты — это современная задача в области мобильной робототехники, решение которой позволяет получить 2D- или 3D-карты окружающего пространства, представляющие информацию не только о пустых и заполненных пространствах, но также и о характере окружающих предметов. В метрико-семантических SLAM-системах необходимо добиться уверенного детектирования элементов пространства (полов, потолков, стен, базовой мебели, источников света в помещении) и, конечно, динамических объектов (людей, автомобилей, других роботов и т. д.). Также

важно знать точное местоположение всех объектов и их геометрию. Возникающие в данном случае сложности обусловлены не только шумом датчиков, искажающим объекты на изображении, но и проблемами, связанными с динамичностью самой системы, которая к тому же перемещается в пространстве, содержащем и статические, и динамические объекты. Кроме того, добавляются погрешности измерений, которые накапливаются со временем, что приводит к отклонениям оцениваемой траектории робота от реальной, а также искажает точность формируемых 2D- или 3D-карт окружающего пространства.

Восстановление карт глубин по изображению на основе DNN. Извлечение информации о глубине сцены важно для восприятия окружающей среды и оценки ее состояния. Качественная оценка расстояния до окружающих объектов особенно необходима для работы автономных мобильных систем. С ускоренным развитием глубоких нейронных сетей их использование для данной задачи показало обнадеживающие результаты с точки зрения точности при обучении end-to-end (e2e) способом, требующим в большинстве случаев только одного изображения в качестве входных данных [1].

Расчет расстояния до объектов по данным монокулярной камеры затруднен из-за внутренней неоднозначности, вызванной шумами и вибрациями на изображении. Поэтому вначале для решения этой проблемы были исследованы методы, основанные на статистических признаках.

В ходе исследования было протестировано несколько алгоритмов в целях выбора лучших с точки зрения точности и работы в реальном времени с частотой выше 10 кадров/с. Такие алгоритмы, как DiverseDepth [2] и MiDaS [3], показали впечатляющие результаты в плане точности, поскольку обучение моделей производилось на большом и разнообразном количестве наборов данных. DELTAS [4] и DeepVideoMVS [5] представляют алгоритмы оценки глубины сцены, основанные на многоракурсной технике, где используются два или три последовательных кадра, а также геометрическая информация о положении и ориентации камеры между этими кадрами для создания плотного облака точек, связанного с оценкой метрических значений глубины. Алгоритмы FastDepth [6], PydNet [7] и FCNN_Node [8] показали наилучшую производительность с точки зрения выполнения в реальном времени. С использованием FastDepth и FCNN_Node были реализованы эффективные сетевые архитектуры с программными решениями во время выполнения для снижения сложности моделей, тогда как в PydNet вычисления производятся только на центральном процессоре.

Другим способом косвенного получения карт глубин является реализация бинокулярных алгоритмов оценки глубины на основе DNN, принимающих на вход пары стереоизображений RGB для создания карт смещений, процесс преобразования которых в карту глубин принципиально прост. В работе [9] представлено множество решений с различными архитектурами — от моделей, обеспечивающих производительность в режиме реального времени, до моделей, дающих высокую точность оценки смещений. Более того, существуют алгоритм MADNet [10], производительность которого может быть адаптирована во время работы, и AANet [11], имеющий архитектуру без слоя 3D-свертки для быстрого оценивания глубины.

Семантическое аннотирование. Семантическая сегментация изображений — это разделение изображений на сегменты в зависимости от того, какому объекту принадлежит каждый пиксел.

В последние годы исследования по семантической сегментации изображений значительно активизировались благодаря появлению различных архитектур сверточных нейросетей (CNN). Точность прогнозирования постоянно повышается, и в целом можно заметить, что проблема сегментации изображения на два-три класса уже решена [12]. Кроме того, отмечается достаточно высокая скорость работы DNN-алгоритмов, обученных на конкретных наборах данных или сценах, например DeepLabV3+ [13] и U-Net [14]. В ряде работ описываются классы универсальных архитектур, таких как HRNet [15] и SegNet [16], которые имеют множество

модификаций для обеспечения лучшей сегментации данных и более быстрого прогнозирования. Другие решения считаются приблизительно универсальными, например ESANet [17], построенная на архитектуре ResNet [18], которая первоначально предназначалась для внутренних сцен, но хорошо работает и для сцен вне помещений.

Семантическое картирование. Благодаря заметному увеличению доступных вычислительных ресурсов современных встраиваемых систем появилась возможность построения семантически аннотированных 3D-карт в режиме реального времени. Такие карты меняют уровень понимания роботом окружающей обстановки. Классические карты дают информацию об окружающем пространстве в виде препятствий и свободных пространств. В свою очередь, метрико-семантические карты содержат множество дополнительной информации, которая позволяет роботу понять, какие именно объекты находятся вокруг него и как далеко они расположены. Эта информация используется для перехода на более интеллектуальный уровень локализации.

В настоящей работе рассматриваются фреймворки, использующие воксельное представление карт или усеченное поле знаковых расстояний (Truncated Signed Distance Field — TSDF). Преимуществами таких карт являются визуальная составляющая, а также возможность преобразования в евклидовы поля знаковых расстояний (Euclidean Signed Distance Field — ESDF), используемые для планирования траекторий. Данные определения и их преимущества наиболее полно объясняются в работе по Voxblox [19].

В ходе исследования были изучены такие современные программные пакеты, как Kimera [20], DA-RNN [21] и Voxblox++ [22], являющиеся фреймворками семантического картирования на основе DNN. Voxblox++ базируется на экземплярно-семантической сегментации с использованием архитектуры Mask R-CNN, тогда как DA-RNN построен на собственной архитектуре, основанной на полностью сверточной сети (Fully Convolutional Network — FCN). Также интересной представляется современная работа Voxgraph [23], фреймворк которой позволяет строить глобально согласованную карту и корректировать ее во время построения в соответствии с замыканиями цикла, но при этом здесь отсутствуют семантические данные.

Для исследования был выбран программный пакет Kimera, отличающийся от других следующими характеристиками:

— метрико-семантическая SLAM-система основана на визуально-инерциальной одометрии (VIO);

— использование вычислительных ресурсов исключительно центрального процессора, что важно с точки зрения энергопотребления для мобильной робототехники;

— наличие обширного и гибкого функционала: например, имеется возможность исключения из построения классов объектов, в частности динамических объектов, что позволяет избежать появления артефактов на карте в виде треков движущихся объектов.

С учетом способности эффективно работать со сценами различной сложности, а также уровня производительности, достаточного для работы в режиме реального времени, были выбраны алгоритмы MiDaS и MADNet в качестве лучших для получения карт глубин, формируемых монокулярной и бинокулярной камерами соответственно. Кроме того, на основе сопоставительного анализа WildDash2 [24] выбран пакет MSeg [25] в качестве решения, имеющего универсальную таксономию для большого количества наборов данных и удовлетворительно работающего в различных сценариях. MSeg-semantic построен на базе HRNet-W48 и обучен сегментировать до 194 классов.

Система метрико-семантического картирования. *Структура системы***.** Схема предлагаемой метрико-семантической SLAM-системы представлена на рис. 1 (здесь GPU — графический процессор). На вход системы поступают изображения, формируемые левой и правой видеокамерами, а также синхронизированные со снимками по времени данные IMU-сенсора. Входными данными служат расстояния до некоторых характерных точек, разреженно расположенных в поле зрения камер. На выходе системы формируется плотная,

точная, сглаженная, семантически аннотированная 3D-карта окружающего пространства, на которой отмечены заранее задаваемые ключевые классы объектов (пол, стены, дверные и оконные проемы, мебель и т.п.).

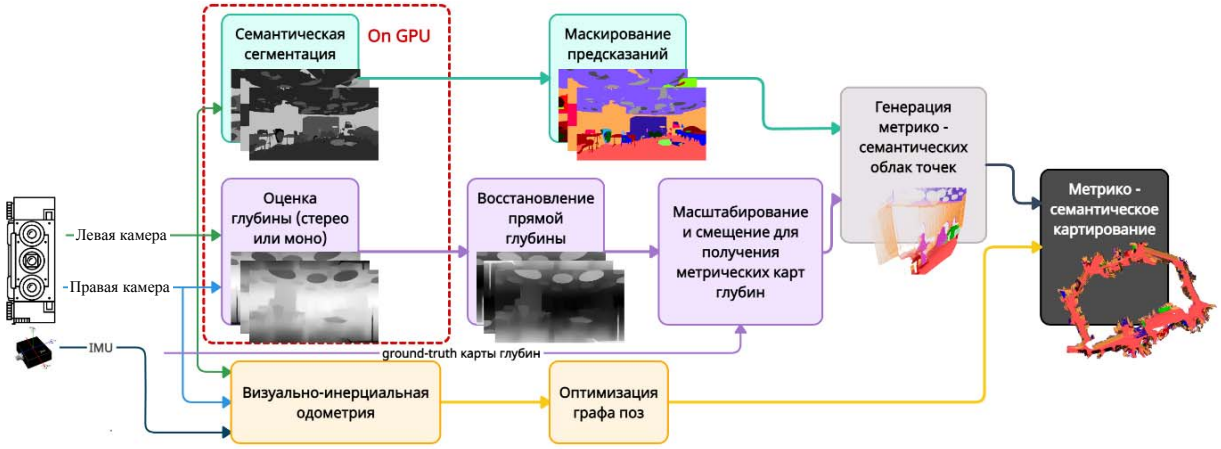

Рис. 1

Выделяются три основных параллельных потока, каждый из которых отвечает за генерацию данных, необходимых для построения метрико-семантической 3D-карты окружающего пространства. *Первый поток* отвечает за генерацию семантически аннотированных изображений и последующее маскирование прогнозируемых сцен в оттенках серого до более наглядных и понятных человеку цветных изображений. *Второй поток* — это расчет плотных RGB-D облаков точек по данным монокулярных или бинокулярных изображений, этот поток отвечает за генерацию 3D-карт: здесь сначала для каждого кадра рассчитывается массив обратных (относительных) глубин, а затем с использованием разреженных прямым измерений расстояния или внутренних параметров камер (в случае бинокулярных изображений) происходит масштабирование полученных обратных глубин и строится карта абсолютных метрических глубин. Далее данные первого и второго потоков объединяются для построения семантического облака точек. *Третий поток* отвечает за работу модуля визуально-инерциальной одометрии (VIO) и модуля оптимизации графа поз (PGO), которые необходимы для получения надежных данных о текущей позиции и перемещениях камеры.

Генерация плотных облаков точек. Для получения облаков точек абсолютной метрической глубины используются обратная глубина, генерируемая монокулярными „оценщиками“ глубины на основе DNN, и глубина, формируемая RGB-D-камерой. Процесс реализации был проведен на наборе данных, который никогда раньше не подавался на вход обученных DNN-алгоритмов MiDaS и MADNet, а схема представлена на рис. 1 вторым потоком.

Монокулярная оценка глубины изображения на основе DNN. Для получения высокого уровня оценки глубины монокулярного изображения были использованы нейросетевая модель MiDaS и 10 наборов данных, имеющих различные характеристики масштаба и сдвига. Исходя из такого большого количества источников, необходимо было предложить функцию потерь, объединяющую стратегии обучения данных.

Процесс генерации обратных карт глубин, являющихся выводом алгоритма MiDaS, происходит параллельно с сохранением полученных облаков точек в виде файлов в формате pfm, которые содержат исходные значения, предсказанные моделью MiDaS.

В зависимости от набора данных, на которых была обучена модель MiDaS, обратные карты глубин получаются с неоднозначными масштабом и сдвигом, которые изменяются от кадра к кадру и не соответствуют реальным метрическим значениям. Возможный способ использования модели MiDaS — генерация карт смещений, которые пропорциональны обратным картам глубин.

В соответствии с этим выполняется покадровая постобработка данных. Для получения абсолютных метрических глубин для некоторых пикселов сцены используются значения глубины d^*, формируемой стереокамерой RealSense D435i. Таким образом, необходимо определить наилучшее соответствие между глубиной d^* и алгоритмическим прогнозом на основе критерия наименьших квадратов. Для восстановления предсказанной глубины в реальном метрическом пространстве используется следующая процедура:

— инвертирование глубины d^* для того, чтобы ее значения соответствовали одной и той же области смещения:

$$\overline{d}^* = 1/d^* ;$$

— расчет масштаба и сдвига карты глубин на основе критерия наименьших квадратов:

$$(s,t) = \underset{s,t}{\operatorname{argmin}} \sum_{i=1}^{P} \left(s\overline{d}_i + t - \overline{d}_i^* \right)^2 ,$$

где $\overline{d}_i$ — глубина, оцениваемая алгоритмом MiDaS; s и t — коэффициенты масштаба и сдвига, P — количество достоверных пикселов, получаемых из карты глубин, формируемой RGB-D-камерой RealSense;

— выравнивание глубины $\overline{d}_i$ с помощью найденных коэффициентов масштаба и сдвига:

$$\hat{\overline{d}}_i = s\overline{d}_i + t,$$

где $\hat{\overline{d}}_i$ — выровненная обратная глубина $\overline{d}_i$;

— инвертирование значения $\hat{\overline{d}}_i$ для получения реальных значений метрической глубины:

$$\hat{D}_i = 1/\hat{\overline{d}}_i ,$$

где $\hat{D}_i$ — метрическое предсказание глубины, согласованное с данными стереокамеры RealSense.

Все сгенерированные облака точек глубины будут упакованы в один файл для дальнейшей работы с наборами данных.

Бинокулярная оценка глубины изображения на основе DNN. Адаптация нейронных сетей к окружающей среде в режиме реального времени при выводе изображений введена в [10]. Адаптация здесь означает применение метода обратного распространения ошибки и его модификаций для редактирования весов модели к данным, которые подаются на вход сети впервые. Это позволяет получать точные карты смещений даже для сцен, на которых модель не была обучена. В MADNet существуют три различных варианта вывода, а именно:

— отсутствие режима адаптации — не применяется метод обратного распространения ошибки и сеть работает так, как она обучалась;

— режим полной адаптации — онлайн-адаптация осуществляется полностью;

— режим MAD — модификации осуществляются быстрее, чем в полном режиме.

Режим полной адаптации работает наилучшим образом, но с более высокой задержкой, тогда как режим MAD представляет собой компромисс по скорости и качеству.

Генерация семантических прогнозов. Как определено выше, одно из наиболее точных универсальных решений — MSeg-semantic. Это решение используется в качестве ядра семантической сегментации.

Предварительно обученная модель MSeg (3 млн параметров, разрешение входящих изображений 480p) была выбрана для выполнения семантической сегментации данных на изображениях по нескольким причинам. Во-первых, эта модель требует минимального объема видеопамяти, что является немаловажным фактором, и, во-вторых, модель характеризуется высоким быстродействием.

Тестирование модели производилось в режиме одномасштабного вывода, что означает использование одного конкретного масштаба изображения (например, оригинального x1). Однако одномасштабный вывод приводит к некачественным прогнозам. Так как при более высоком разрешении входного изображения алгоритм MSeg „разбивает“ изображение на более мелкие фрагменты, то на границах этих фрагментов возникают ошибки в предсказании классов объектов. Во избежание этого был использован многомасштабный вывод, который заключается в выполнении многократного прогнозирования для одних и тех же изображений при различном масштабировании и в последующем усреднении прогнозов для каждого пиксела.

Многомасштабный вывод позволяет исключить отрицательное влияние некачественных семантических прогнозов на SLAM-систему, провести более наглядные эксперименты и получить более точные и плавные предсказания. Однако многомасштабный вывод намного дороже по расходу ресурсов и не работает в режиме реального времени. Поэтому в будущем, при реализации системы в режиме реального времени, планируется использовать одномасштабный вывод.

Изначально на выходе модели MSeg-semantic формируются изображения в оттенках серого, представляющие семантические метки, где, согласно универсальной таксономии [25], 194 метки соответствуют каждому известному классу, а 195-я метка выделяется для неизвестных объектов. Для более корректного представления 2D-семантических изображений, семантического облака точек и результирующей семантической карты была выполнена собственная цветовая маскировка изображений, базирующаяся на Detectron2 [26], и, таким образом, получены RGB семантически-аннотированные изображения.

Экспериментальные результаты. В качестве аппаратного обеспечения для выполнения семантической сегментации 2D-изображений с помощью модели MSeg и для пакета Kimera использовался ноутбук с процессором Intel I5-9300H 2,4 ГГц×8 и графическим процессором NVIDIA GeForce GTX 1050; для работы DNN-алгоритмов по оценке глубины использован одноплатный компьютер NVIDIA Jetson AGX Xavier.

Основной набор данных, предоставленный лабораторией робототехники ПАО „Сбербанк“, представляет собой данные, записанные с мобильного наземного робота в офисном помещении, имеющем большие открытые пространства, узкие коридоры и большое количество окон, что создает дополнительные требования к тестируемым решениям и позволяет более точно определить лучшие из них. Кроме того, этот набор содержит все необходимые данные, включая RGB-изображения, инфракрасные стереоизображения, необработанные измерения IMU-сенсора, карты глубин, формируемые RealSense D435i, характеристики стереокамеры. Для оценки качества фреймворка Kimera-VIO использовались данные одометрии, полученные SLAM-системой робота-курьера ПАО „Сбербанк“.

Были обработаны 11 475 RGB-моноизображений с разрешением 1280×720 с помощью моделей MiDaS и MSeg-semantic, а также 11 282 стереоизображения с разрешением 1280×720, сформированные правой и левой камерами RealSense, с использованием алгоритма MADNet в полном режиме. Результат обработки представлен на рис. 2. Для сравнения дополнительно приведены данные, полученные с помощью модифицированного алгоритма Хиршмюллера (Stereo Semi-Global Matching — SGM), используемого по умолчанию в Kimera в случае отсутствия других данных о глубине.

Для фреймворка Kimera объединены плотные данные различных алгоритмов на основе DNN и семантические 2D-изображения с исходным набором данных в целях получения расширенного набора данных. Поскольку генерируемые данные записываются в режиме реального времени, их необходимо адаптировать ко времени, связанному с исходным набором данных. Поэтому временные метки сгенерированных данных переназначаются, чтобы они соответствовали тем, которые существовали в исходном наборе данных.

Для того чтобы все модули Kimera работали в режиме реального времени и во избежание больших затрат ресурсов, используется масштаб 0,5 для левого и правого изображений. Кроме того, произведена модификация Kimera-Semantics для работы с 195 классами и создана карта цветов для 2D-семантических RGB-изображений, которая подается на вход Kimera-Semantics.

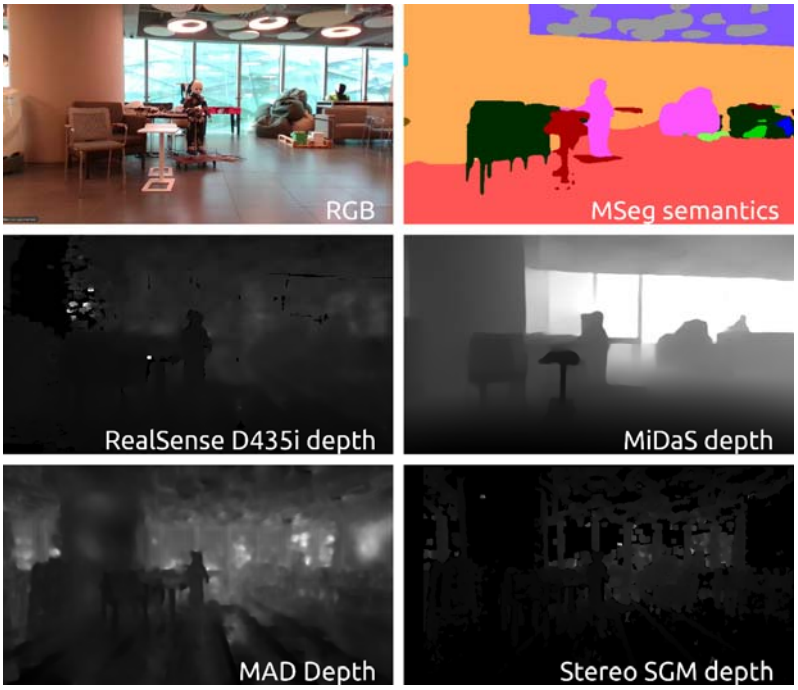

Рис. 2

Объединяя облака точек глубины и семантические метки, получаем семантические облака точек. На рис. 3 представлены семантические облака точек, полученные для каждого решения. Визуальная оценка их качества наглядно демонстрирует превосходство MiDaS над остальными алгоритмами по показателю отсутствия шума. Однако алгоритмом MiDaS не решается проблема детектирования зеркал. Кроме того, использование RealSense обеспечивает наиболее точную глубину для каждого пиксела, тогда как MiDaS не позволяет сделать это напрямую, если отсутствуют дополнительные данные о метрике пространства. Тем не менее плотные облака точек, сформированные RGB-D-камерой RealSense, не обеспечивают получение достоверной информации о прозрачных объектах и отражающих поверхностях, что приводит к большим погрешностям и, следовательно, к повреждению 3D-карт. Более того, облака точек, сгенерированные с использованием MiDaS, имеют более высокое качество с геометрической точки зрения. С другой стороны, облака точек, полученные с помощью MADNet, содержат больше всего артефактов, что объясняется неудовлетворительным предсказанием глубины инфракрасных монохромных изображений, поскольку эта сеть была обучена на стерео RGB-изображениях.

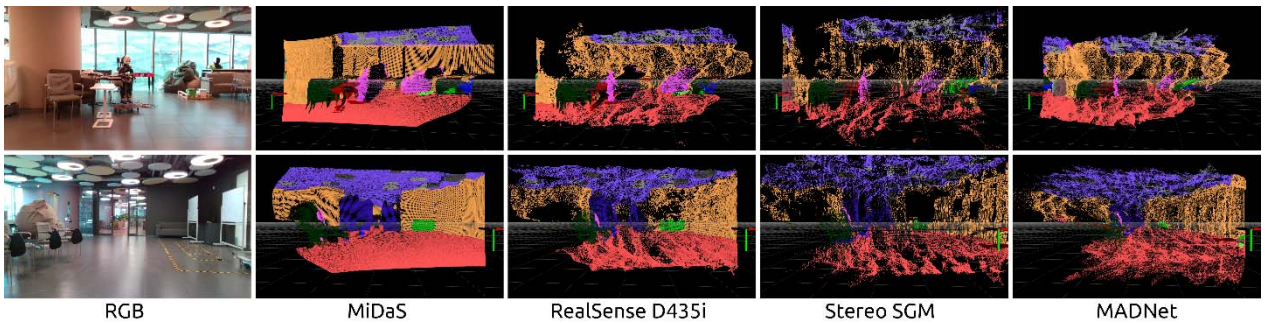

Рис. 3

Для каждого решения были созданы метрико-семантические 3D-карты (рис. 4, вид сверху), а сравнение карт в нескольких сценах представлено на рис. 5 (в скобках указан размер воксела). Визуальный анализ показывает, что качество построенной с помощью MiDaS семантической карты с использованием карт глубин является наилучшим, тогда как решение MADNet наихудшее. Все решения, кроме MiDaS, показали значительные выбросы, особенно в областях сцен, содержащих прозрачные объекты, такие как окна. Визуализация построения метрико-семантической 3D-карты с использованием карт глубин, полученных с помощью алгоритма MiDaS, представлена в электронном ресурсе https://youtu.be/VtPTjaiF8Jw.

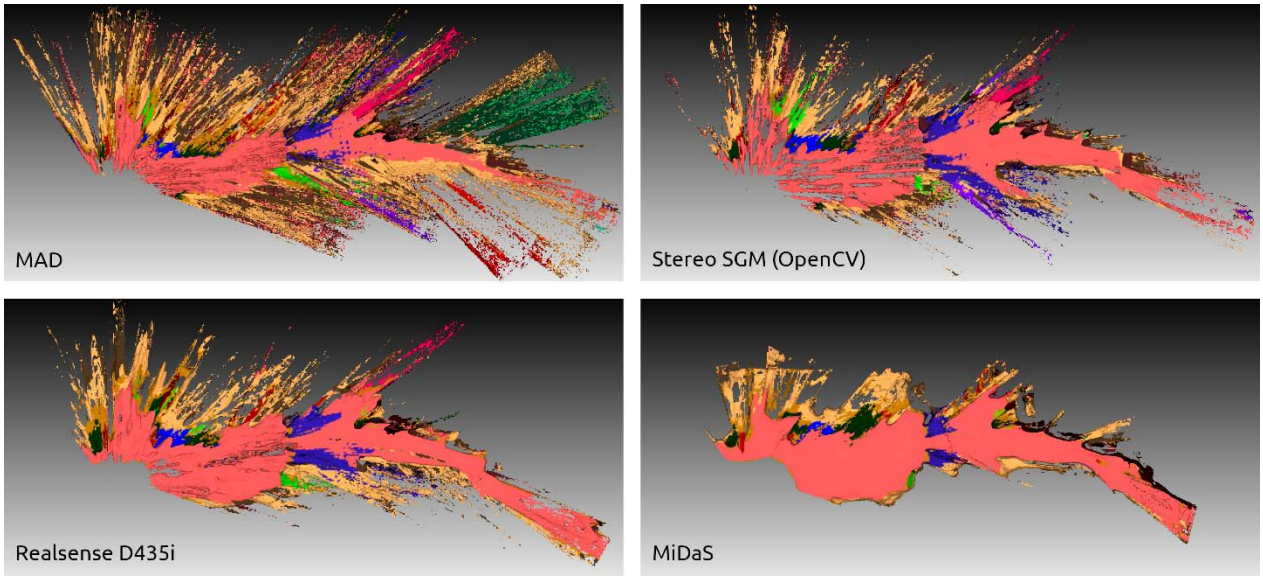

Рис. 4

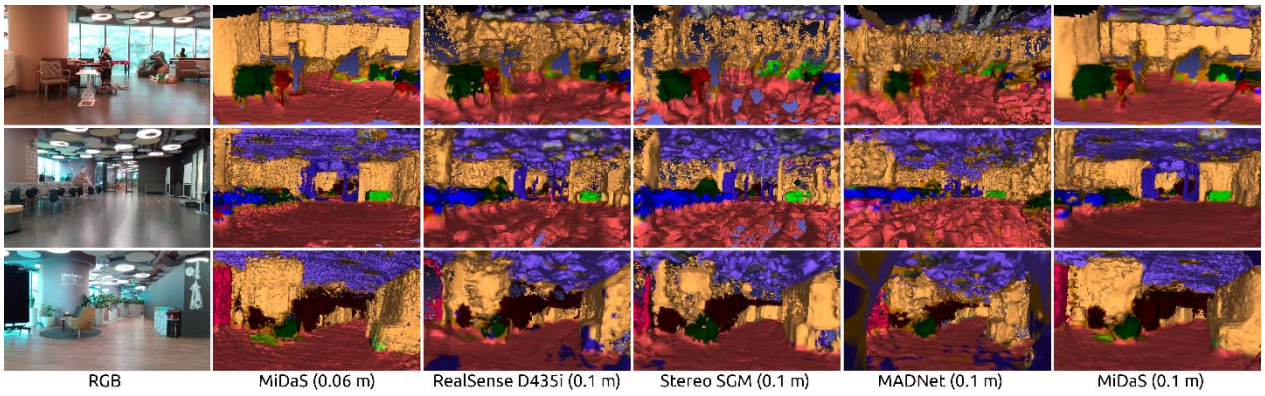

Рис. 5

Для трехмерного случая в качестве основных метрик для оценки качества карт использовались метрики Хаусдорфа. При этом карта, полученная на основе информации о глубине, сформированной камерой RealSense, использовалась как „наземная истина“ (ground-truth). В общем случае одностороннее расстояние Хаусдорфа определяется как максимальное из всех возможных расстояний от каждой точки одного множества до ближайшей к ней точки второго множества. Двустороннее расстояние Хаусдорфа обобщается до максимального из односторонних расстояний Хаусдорфа относительно каждого из множеств, т.е.:

$$D_H\left(X,Y\right)=\max\left\{\sup_{x\in X}\inf_{y\in Y}\left|xy\right|,\sup_{y\in Y}\inf_{x\in X}\left|xy\right|\right\}.$$

Дополнительной метрикой качества является среднее расстояние между полученными 3D-картами, а точнее, между каждым вокселом одной карты и ближайшим к нему вокселом другой карты:

$$\bar{D}(X,Y)=\frac{1}{X}\sum_{x\in X}\inf_{y\in Y}|xy|;\ \ \bar{D}(Y,X)=\frac{1}{Y}\sum_{y\in Y}\inf_{x\in X}|xy|,$$

которое целесообразно представить в виде среднего хаусдорфового расстояния:

$$\bar{D}_H(X,Y)=\frac{\bar{D}(X,Y)+\bar{D}(Y,X)}{2}.$$

Результаты оценивания точности восстановления карт глубин с помощью метрик Хаусдорфа представлены в табл. 1. Следует отметить, что показатели метрик имеют достаточно большие значения, особенно двустороннее расстояние Хаусдорфа. Это вызвано множеством причин, в частности несовершенством выбранных в качестве ground-truth данных. Отдаленные выбросы на 3D-картах в наибольшей степени влияют на двустороннее расстояние Хаусдорфа, поэтому в таблице представлены также другие метрики для более справедливого сравнения. Кроме того, в качестве визуальной метрики для выявления явных выбросов и шумов использована тепловая карта ошибок между метрико-семантическими 3D-картами, построенными с применением карт глубин от MiDaS и RealSense (рис. 6).

Таблица 1

Решение (размер воксела)	**Количество вершин**	**Расстояние Хаусдорфа**		
		Двустороннее	**Среднее**	**Среднеквадратическое**
Stereo SGM (0,1 м)	939667	6,284989	0,315699	0,520958
MADNet (0,1 м)	2342929	8,420206	0,663223	1,148033
MiDaS (0,1 м)	347426	1,084647	0,158118	0,214312
MiDaS (0,06 м)	1105870	1,139520	0,162516	0,223605

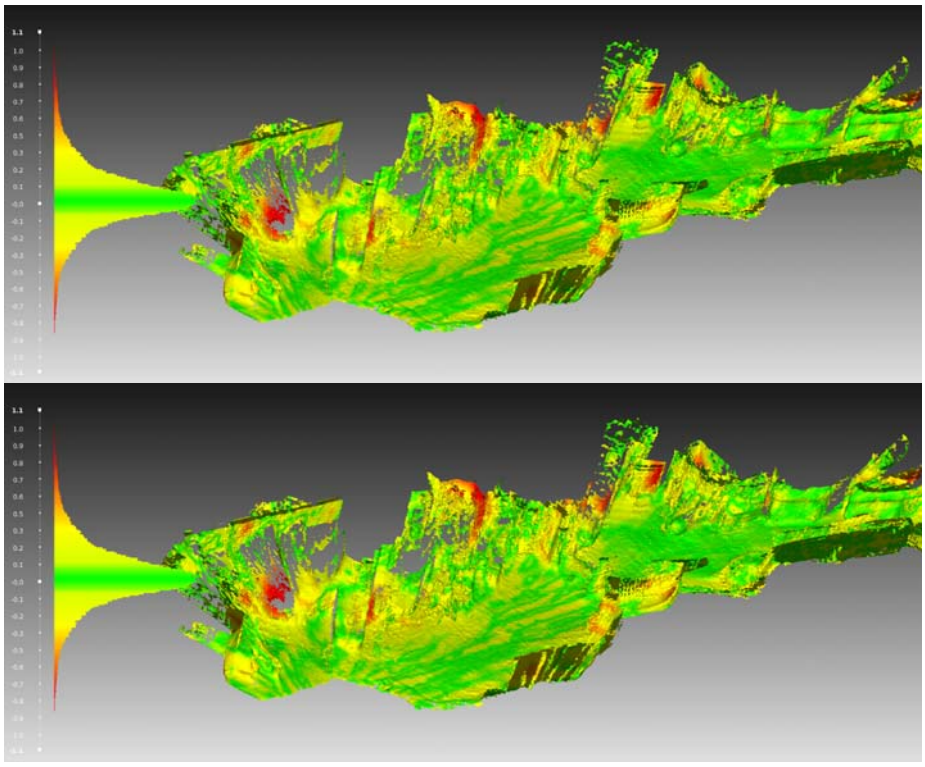

Рис. 6

Еще один важный фактор, который следует отметить при построении карт с использованием глубины, формируемой MiDaS, — меньшее использование оперативной памяти. В табл. 2 приведены данные по использованию оперативной памяти (ОЗУ) в процессе

построения одного и того же фрагмента карты для каждого из решений, а также данные по частоте генерации облаков точек глубины с помощью этих решений. Согласно рис. 5 и табл. 2 заметна зависимость между качеством карты и объемом используемой памяти — чем лучше карта, тем меньше памяти она использует. Это объясняется геометрическим совпадением точек у соседних кадров, в результате чего образуется меньшее количество вокселов.

Таблица 2

Решение (размер воксела)	ОЗУ, Гбайт	Частота, Гц
RealSense D435i (0,1 м)	2,5	30
Stereo SGM (0,1 м)	2,9	30
MADNet (0,1 м)	4,1	1,17
MiDaS (0,1 м)	1,5	16
MiDaS (0,06 м)	4,6	16

Заключение. Для обеспечения качественной работы метрико-семантической VIO-SLAM-системы были протестированы различные решения по генерации облаков точек глубины изображения. Выбор лучших решений основан на их общей производительности в реальном времени и универсальности для различных сцен с естественными условиями. Аналогичный критерий был задан при выборе пакета MSeg в качестве семантического предиктора.

Для получения правильно масштабированных облаков точек глубины, генерируемых алгоритмом MiDaS, необходимо иметь как минимум несколько точных результатов измерений расстояния. Таким образом, было выполнено слияние алгоритма MiDaS и решения, полученного с помощью RGB-D-камеры RealSense, для генерации более реалистичного метрического облака точек глубины.

Представлена разработанная собственная SLAM-система, основанная на нескольких DNN-алгоритмах, которая может работать благодаря всего одному устройству получения данных — RGB-D-стереокамере. Апробация работы системы проведена на реальных данных, предоставленных лабораторией робототехники ПАО „Сбербанк“, и дополнительно проведены эксперименты, результаты которых демонстрируют высокие требования к алгоритмам. Работа системы показала, что карты, созданные с помощью DNN-алгоритмов, более стабильны и детальны по сравнению с картами, сформированными RGB-D-камерой RealSense, особенно в сценах, содержащих яркие объекты.

В дальнейшем планируется развитие системы с использованием только данных визуально-инерциальной одометрии для обеспечения корректного масштабирования получаемых облаков точек с исключением необходимости в подаче на вход карт глубин из RGB-D-камер и расширением, таким образом, разрабатываемого решения для использования с любыми RGB-камерами.

СПИСОК ЛИТЕРАТУРЫ

1. *Zhao C., Sun Q., Zhang C., Tang Y., Qian F.* Monocular depth estimation based on deep learning: An overview // Science China Technological Sciences. 2020. Vol. 63, N 9. P. 1612—1627. DOI: 10.1007/s11431-020-1582-8.

2. *Yin W., Liu Y., Shen C.* Virtual Normal: Enforcing Geometric Constraints for Accurate and Robust Depth Prediction // IEEE Trans. on Pattern Analysis and Machine Intelligence. 2021. DOI: 10.1109/TPAMI.2021.3097396.

3. *Ranftl R., Lasinger K., Hafner D., Schindler K., Koltun V.* Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer // IEEE Trans. on Pattern Analysis and Machine Intelligence. 2022. Vol. 44, N 03. P. 1623—1637.

4. *Sinha A., Murez Z., Bartolozzi J., Badrinarayanan V., Rabinovich A.* DELTAS: Depth Estimation by Learning Triangulation and Densification of Sparse Points // Computer Vision – ECCV. 2020. Vol. 12366. P. 104—121. DOI: 10.1007/978-3-030-58589-1_7.

5. *Duzceker A., Galliani S., Vogel C., Speciale P., Dusmanu M., Pollefeys M.* DeepVideoMVS: Multi-View Stereo on Video with Recurrent Spatio-Temporal Fusion // Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. 2021. June. P. 15324—15333. DOI: 10.1109/CVPR46437.2021.01507.

6. *Wofk D., Ma F., Yang T.-J., Karaman S., Sze V.* FastDepth: Fast monocular depth estimation on embedded systems // Proc. IEEE Intern. Conf. on Robotics and Automation. 2019. Vol. 2019, May. P. 6101—6108. DOI: 10.1109/ICRA.2019.8794182.

7. *Poggi M., Aleotti F., Tosi F., Mattoccia S.* Towards Real-Time Unsupervised Monocular Depth Estimation on CPU // IEEE Intern. Conf. on Intelligent Robots and Systems. 2018. P. 5848—5854. DOI: 10.1109/IROS.2018.8593814.

8. *Bokovoy A., Muravyev K., Yakovlev K.* Real-time Vision-based Depth Reconstruction with NVidia Jetson // European Conf. on Mobile Robots (ECMR). 2019. P. 1—6. DOI: 10.1109/ECMR.2019.8870936.

9. *Smolyanskiy N., Kamenev A., Birchfield S.* On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach // IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops. 2018. Vol. 2018, June. P. 1120—1128. DOI: 10.1109/CVPRW.2018.00147.

10. *Tonioni A., Tosi F., Poggi M., Mattoccia S., L. Stefano D.* Real-time self-adaptive deep stereo // Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2019. Vol. 2019. June. P. 195—204. DOI: 10.1109/CVPR.2019.00028.

11. *Xu H., Zhang J.* AANET: Adaptive aggregation network for efficient stereo matching // Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2020. P. 1956—1965. DOI: 10.1109/CVPR42600.2020.00203.

12. *Xia L., Cui J., Shen R., Xu X., Gao Y., Li X.* A survey of image semantics-based visual simultaneous localization and mapping: Application-oriented solutions to autonomous navigation of mobile robots // Intern. Journal of Advanced Robotic Systems. 2020. Vol. 17. P. 1—17. DOI: 10.1177/1729881420919185.

13. *Chen L.-C., Zhu Y., Papandreou G., Schroff F., Adam H.* Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation // Computer Vision – ECCV. 2018. Vol. 11211. P. 833—851. DOI: 10.1007/978-3-030-01234-2_49.

14. *Ronneberger O., Fischer P., Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation // Medical Image Computing and Computer-Assisted Intervention – MICCAI. 2015. Vol. 9351. P. 234—241. DOI: 10.1007/978-3-319-24574-4_28.

15. *Wang J., Sun K., Cheng T., Jiang B., Deng C., Zhao Y., Liu D., Mu Y., Tan M., Wang X., Liu W., Xiao B.* Deep High-Resolution Representation Learning for Visual Recognition // IEEE Trans. on Pattern Analysis and Machine Intelligence. 2021. Vol. 43, N 10. P. 3349—3364. DOI: 10.1109/TPAMI.2020.2983686.

16. *Badrinarayanan V., Kendall A., Cipolla R.* SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation // IEEE Trans. on Pattern Analysis and Machine Intelligence. 2017. Vol. 39, N 12. P. 2481—2495. DOI: 10.1109/TPAMI.2016.2644615.

17. *Seichter D., Köhler M., Lewandowski B., Wengefeld T., Gross H. -M.* Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis // IEEE Intern. Conf. on Robotics and Automation (ICRA). 2021. P. 13525—13531. DOI: 10.1109/ICRA48506.2021.9561675.

18. *He K., Zhang X., Ren S., Sun J.* Deep Residual Learning for Image Recognition // IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2016. P. 770—778. DOI: 10.1109/CVPR.2016.90.

19. *Oleynikova H., Taylor Z., Fehr M., Siegwart R., Nieto J.* Voxblox: Incremental 3D Euclidean Signed Distance Fields for on-board MAV planning // IEEE/RSJ Intern. Conf. on Intelligent Robots and Systems (IROS). 2017. P. 1366—1373. DOI: 10.1109/IROS.2017.8202315.

20. *Rosinol A., Abate M., Chang Y., Carlone L.* Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping // IEEE Intern. Conf. on Robotics and Automation (ICRA). 2020. P. 1689—1696. DOI: 10.1109/ICRA40945.2020.9196885.

21. *Xiang Y., Fox D.* DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks // Robotics: Science and Systems. 2017. Vol. 13. DOI: 10.15607/RSS.2017.XIII.013.

22. *Grinvald M.* et al. Volumetric Instance-Aware Semantic Mapping and 3D Object Discovery // IEEE Robotics and Automation Letters. 2019. Vol. 4, N 3. P. 3037—3044. DOI: 10.1109/LRA.2019.2923960.

23. *Reijgwart V., Millane A., Oleynikova H., Siegwart R., Cadena C., Nieto J.* Voxgraph: Globally Consistent, Volumetric Mapping Using Signed Distance Function Submaps // IEEE Robotics and Automation Letters. 2020. Vol. 5, N 1. P. 227—234. DOI: 10.1109/LRA.2019.2953859.

24. *Zendel O., Honauer K., Murschitz M., Steininger D., Dominguez G. F.* WildDash — Creating Hazard-Aware Benchmarks // Proc. of the European Conf. on Computer Vision (ECCV). 2018. Sept. P. 402—416.

25. *Lambert J., Liu Z., Sener O., Hays J., Koltun V.* MSeg: A Composite Dataset for Multi-Domain Semantic Segmentation // IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2020. P. 2876—2885. DOI: 10.1109/CVPR42600.2020.00295.

26. *Kirillov A., Wu Y., He K., Girshick R.* PointRend: Image Segmentation As Rendering // IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2020. P. 9796—9805. DOI: 10.1109/CVPR42600.2020.00982.

Сведения об авторах

Амиран Рустамович Беркаев — студент; Университет ИТМО, факультет систем управления и робототехники, лаборатория биомехатроники и энергоэффективной робототехники; E-mail: berkaevamiran@mail.ru

Малик Мохрат — студент; Университет ИТМО, факультет систем управления и робототехники, лаборатория биомехатроники и энергоэффективной робототехники; E-mail: malik.mohrat@gmail.com

Алексей Михайлович Бурков — ПАО „Сбербанк“, лаборатория робототехники; ведущий инженер-разработчик; E-mail: amburkoff@gmail.com

Сергей Алексеевич Колюбин — д-р техн. наук, доцент; Университет ИТМО, факультет систем управления и робототехники, лаборатория биомехатроники и энергоэффективной робототехники; вед. научный сотрудник; E-mail: s.kolyubin@itmo.ru

Поступила в редакцию 28.12.21; одобрена после рецензирования 12.01.22; принята к публикации 18.01.22.

REFERENCES

1. Zhao C., Sun Q., Zhang C., Tang Y., and Qian F. *Science China Technological Sciences*, 2020, no. 9(63), pp. 1612–1627, DOI: 10.1007/s11431-020-1582-8.
2. Yin W., Liu Y. and Shen C. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/TPAMI.2021.3097396.
3. Ranftl R., Lasinger K., Hafner D., Schindler K., and Koltun V. *IEEE Transactions on Pattern Analysis* and *Machine Intelligence*, 2022, no. 03(44), pp. 1623–1637.
4. Sinha A., Murez Z., Bartolozzi J., Badrinarayanan V., and Rabinovich A. *Computer Vision – ECCV 2020, Springer International Publishing, Cham*, 2020, vol. 12366, pp. 104–121, DOI: 10.1007/978-3-030-58589-1_7.
5. Düzçeker A., Galliani S., Vogel C., Speciale P., Dusmanu M., and Pollefeys M. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA*, June 2021, pp. 15324–15333, DOI: 10.1109/CVPR46437.2021.01507.
6. Wofk D., Ma F., Yang T.-J., Karaman S., and Sze V. *Proceedings - IEEE International Conference on Robotics and Automation*, 2019, vol. 2019, May, pp. 6101–6108, DOI: 10.1109/ICRA.2019.8794182.
7. Poggi M., Aleotti F., Tosi F., and Mattoccia S. *IEEE International Conference on Intelligent Robots and Systems*, 2018, pp. 5848–5854, DOI: 10.1109/IROS.2018.8593814.
8. Bokovoy A., Muravyev K., and Yakovlev K. *2019 European Conference on Mobile Robots (ECMR)*, 2019, pp. 1–6, DOI: 10.1109/ECMR.2019.8870936.
9. Smolyanskiy N., Kamenev A., and Birchfield S. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018, vol. 2018-June, pp. 1120–1128, doi: 10.1109/CVPRW.2018.00147.
10. Tonioni A., Tosi F., Poggi M., Mattoccia S., and Stefano L.D. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, vol. 2019-June, pp. 195–204, DOI: 10.1109/CVPR.2019.00028.
11. Xu H. and Zhang J. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1956–1965, DOI: 10.1109/CVPR42600.2020.00203.
12. Xia L., Cui J., Shen R., Xu X., Gao Y., and Li X. Int. J. *Adv. Robot. Syst*., 2020, vol. 17, p. 172988142091918, DOI: 10.1177/1729881420919185.
13. Chen L.-C., Zhu Y., Papandreou G., Schroff F., and Adam H. *Computer Vision – ECCV, Springer International Publishing, Cham*, 2018, vol. 11211, pp. 833–851, DOI: 10.1007/978-3-030-01234-2_49.
14. Ronneberger O., Fischer P., and Brox T. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham*, 2015, vol. 9351, pp. 234–241, DOI: 10.1007/978-3-319-24574-4_28.
15. Wang J., Sun K., Cheng T., Jiang B., Deng C., Zhao Y., Liu D., Mu Y., Tan M., Wang X., Liu W. and Xiao B. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, no. 10(43), pp. 3349–3364, DOI: 10.1109/TPAMI.2020.2983686.
16. Badrinarayanan V., Kendall A., and Cipolla R. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,

2017, no. 12(39), pp. 2481–2495, DOI: 10.1109/TPAMI.2016.2644615.
17. Seichter D., Köhler M., Lewandowski B., Wengefeld T., and Gross H.-M. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13525–13531, DOI: 10.1109/ICRA48506.2021.9561675.
18. He K., Zhang X., Ren S., and Sun J. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 770–778, DOI: 10.1109/CVPR.2016.90.
19. Oleynikova H., Taylor Z., Fehr M., Siegwart R., and Nieto J. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 2017, pp. 1366–1373, DOI: 10.1109/IROS.2017.8202315.
20. Rosinol A., Abate M., Chang Y., and Carlone L. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1689–1696, DOI: 10.1109/ICRA40945.2020.9196885.
21. Xiang Y. and Fox D. *2017 Robotics: Science and Systems*, 2017, july, vol. 13, DOI: 10.15607/RSS.2017.XIII.013.
22. Grinvald M. et al. *IEEE Robot. Autom. Lett*., 2019, no. 3(4), pp. 3037–3044, DOI: 10.1109/LRA.2019.2923960.
23. Reijgwart V., Millane A., Oleynikova H., Siegwart R., Cadena C., and Nieto J. *IEEE Robot. Autom. Lett*., 2020, no. 1(5), pp. 227–234, DOI: 10.1109/lra.2019.2953859.
24. Zendel O., Honauer K., Murschitz M., Steininger D., and Dominguez G.F. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–416.
25. Lambert J., Liu Z., Sener O., Hays J., and Koltun V. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2876–2885, DOI: 10.1109/CVPR42600.2020.00295.
26. Kirillov A., Wu Y., He K. and Girshick R. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2020, pp. 9796–9805, DOI: 10.1109/CVPR42600.2020.00982.

Data on authors

Amiran R. Berkaev — Student; ITMO University, Faculty of Control Systems and Robotics, International Laboratory of Biomechatronic and Energy-Efficient Robotics; E-mail: berkaevamiran@mail.ru

Malik Mohrat — Student; ITMO University, Faculty of Control Systems and Robotics, International Laboratory of Biomechatronic and Energy-Efficient Robotics; E-mail: malik.mohrat@gmail.com

Alexey M. Burkov — Sberbank, Robotics Laboratory; Lead Engineer-Designer; E-mail: amburkoff@gmail.com

Sergey A. Kolyubin — Dr. Sci., Associate Professor; ITMO University, Faculty of Control Systems and Robotics, International Laboratory of Biomechatronic and Energy-Efficient Robotics; Leading Researcher; E-mail: s.kolyubin@itmo.ru

Received 28.12.21; approved after reviewing 12.01.22; accepted for publication 18.01.22.