УДК 004.85:004.93
DOI

ИССЛЕДОВАНИЕ АЛГОРИТМОВ КОМПЬЮТЕРНОГО ЗРЕНИЯ ДЛЯ ПРИМЕНЕНИЯ В АССИСТИВНЫХ СИСТЕМАХ ПОМОЩИ СЛАБОВИДЯЩИМ

[1]Вотякова Л. Р. ORCID ID 0009-0000-6406-9179, [2]Нуретдинов Р. Г.

[1]Федеральное государственное бюджетное образовательное учреждение высшего образования «Казанский национальный исследовательский технологический университет», Казань, Российская Федерация, e-mail: votyakovalr@corp.knrtu.ru;
[2] Нижнекамский химико-технологический институт (филиал) федерального государственного бюджетного образовательного учреждения высшего образования «Казанский национальный исследовательский технологический университет», Нижнекамск, Российская Федерация

Современные методы компьютерного зрения решают широкий спектр задач, от базовой категоризации изображений до комплексного анализа пространственной структуры окружения. Выбор моделей определяется их способностью балансировать между точностью, скоростью обработки и адаптивностью к условиям реального мира. Целью данной работы является анализ применимости алгоритмов компьютерного зрения в ассистивных системах помощи слабовидящим и незрячим пользователям, а также оценка того, какие типы моделей могут быть использованы для первичного описания сцены, обнаружения значимых объектов, распознавания текстовой информации и анализа проходимых поверхностей. В рамках данной работы рассмотрены модели для классификации изображений, детекции объектов, семантической сегментации изображений и распознавания текста из окружения в целях разработки ассистивных систем. Классификация сцен может применяться для первичного определения общего контекста окружающей среды, детекция объектов – для выделения отдельных значимых предметов и потенциальных препятствий, оптическое распознавание символов – для извлечения текстовой информации, а семантическая сегментация – для анализа структуры пространства и проходимых поверхностей. В разработанном прототипе ассистивной системы дополнительным завершающим звеном конвейера выступает языковая модель, используемая для консолидации выходных данных отдельных модулей компьютерного зрения и формирования конечной текстовой подсказки для пользователя.

Ключевые слова: ассистивная система, классификация, детекция, семантическая сегментация, распознавание

RESEARCH OF COMPUTER VISION ALGORITHMS FOR USE IN ASSISTIVE SYSTEMS FOR THE BLIND

[1]Votyakova L. R. ORCID ID 0009-0000-6406-9179, [2]Nuretdinov R. G.

[1]Federal State Budgetary Educational Institution of Higher Education "Kazan National Research Technological University", Kazan, Russian Federation, e-mail: votyakovalr@corp.knrtu.ru;
[2]Nizhnekamsk Institute of Chemical Technology (branch) of the Federal State Budgetary Educational Institution of Higher Education "Kazan National Research Technological University", Nizhnekamsk, Russian Federation

Modern computer vision methods solve a wide range of tasks, from basic image categorization to complex analysis of the spatial structure of the environment. The choice of models is determined by their ability to balance accuracy, processing speed, and adaptability to real-world conditions. The goal of this work is to analyze the applicability of computer vision algorithms in assistive systems for visually impaired and blind users, and to evaluate which types of models can be used for initial scene description, detection of significant objects, recognition of text information, and analysis of walkable surfaces. As part of this work models for image classification, object detection, semantic image segmentation, and text recognition from the environment are considered. Scene classification can be used to initially determine the general context of the environment, object detection can be used to identify individual objects and potential obstacles, optical character recognition can be used to extract text information, and semantic segmentation can be used to analyze the structure of the environment and the surfaces that can be walked on. In the developed prototype of the assistive system, an additional final link in the pipeline is a language model used to consolidate the output data of individual computer vision modules and generate a final text prompt for the user.

Keywords: assistive systems, classification, detection, semantic segmentation, recognition

Введение

По данным Всемирной организации здравоохранения, нарушения зрения остаются одной из значимых медико-социальных проблем: в мире не менее 2,2 млрд чел. имеют нарушения ближнего или дальнего зрения, при этом как минимум в 1 млрд случаев нарушение зрения могло быть предотвращено либо еще не получило необходимой коррекции [1]. Для слабовидящих и незрячих пользователей эта проблема выражается не только в снижении качества

зрительного восприятия, но и в повседневных трудностях: самостоятельном передвижении, поиске ориентиров, чтении информации в окружающей среде и своевременном обнаружении опасных участков пути.

Развитие мобильных устройств, методов компьютерного зрения и искусственного интеллекта привело к появлению ассистивных приложений, которые частично компенсируют недостаток зрительной информации. К таким решениям относятся Microsoft Seeing AI, Google Lookout и Envision AI. Эти приложения используют камеру смартфона для чтения текста, описания объектов и сцен, распознавания отдельных элементов окружения и предоставления пользователю голосовой информации [2]. Вместе с тем для задачи безопасного движения одного общего описания сцены часто недостаточно. Пользователю важно понимать не только то, какие объекты находятся перед ним, но и какая часть пространства является проходимой, где расположены препятствия, ступени, границы пола, дороги или тротуара.

В связи с этим в данной работе рассматривается не отдельная модель компьютерного зрения, а сочетание нескольких подходов: классификации сцены, детекции объектов, распознавания текста и семантической сегментации. Отдельное внимание уделяется сегментации, поскольку она позволяет перейти от общего описания изображения к анализу структуры пространства. На основе пиксельного выделения пола, дороги, тротуара, стен, лестниц и препятствий можно формировать признаки, необходимые для математической оценки проходимой поверхности: направление свободной зоны, наличие перепадов высоты, границы безопасного движения и возможные препятствия на пути пользователя. Такой подход дополняет существующие ассистивные решения, ориентированные преимущественно на чтение текста, распознавание объектов и описание окружающей сцены [3–5].

Цель исследования – анализ применимости алгоритмов компьютерного зрения в ассистивной системе помощи слабовидящим и незрячим пользователям, а также оценка того, какие типы моделей могут быть использованы для первичного описания сцены, обнаружения значимых объектов, распознавания текстовой информации и анализа проходимых поверхностей.

Материалы и методы исследования

Современные методы компьютерного зрения решают широкий спектр задач, от базовой категоризации изображений до комплексного анализа пространственной структуры окружения. Выбор моделей определяется их способностью балансировать между точностью, скоростью обработки и адаптивностью к условиям реального мира [6–8].

Классификация изображений в предлагаемом конвейере используется как этап первичной оценки сцены [9]. На этом уровне системе не требуется подробно описывать все объекты перед пользователем; важно быстро определить общий контекст: находится ли пользователь в помещении, коридоре, на улице, в торговом пространстве или в домашнем интерьере. Такая информация не заменяет детекцию объектов и семантическую сегментацию, но помогает выбрать дальнейшую стратегию анализа изображения.

В литературе для решения задач классификации изображений широко применяются сверточные и гибридные архитектуры, отличающиеся соотношением точности и вычислительной сложности. В частности, EfficientNet основан на согласованном масштабировании глубины, ширины и разрешения сети, что делает его удобным вариантом для задач, где важен баланс между качеством и скоростью обработки [10]. ConvNeXt, в свою очередь, представляет собой современную переработку сверточной архитектуры с учетом решений, характерных для более поздних моделей компьютерного зрения [11]. Поэтому в экспериментальную часть были включены модели EfficientNet-B0, MobileNetV3-Small, ResNet-50 и ConvNeXt-Tiny.

Для того чтобы оценка моделей не ограничивалась только теоретическим сравнением архитектур, была проведена экспериментальная проверка на локальной выборке изображений. В рамках исследования рассматривались четыре функциональных блока ассистивной системы: классификация сцен, детекция объектов, семантическая сегментация и распознавание текста.

Количественная оценка в данной версии работы была выполнена для классификации сцен и OCR-модуля. Для детекции объектов и семантической сегментации дополнительно были проведены пилотные количественные проверки. Для детекции была подготовлена bbox-разметка на 30 изображениях по пяти укрупненным классам, значимым для ассистивной навигации: person, vehicle, animal, furniture и obstacle. Для семантической сегментации была подготовлена пилотная mask-разметка на 18 изображениях по классам floor, road, sidewalk, stairs и obstacle. Полученные результаты рассматривались как предварительная количественная проверка работоспособности модулей в составе ассистивного конвейера, а не как итоговый независимый benchmark.

Для экспериментальной оценки была сформирована локальная тестовая выборка изображений, отражающих типовые визуальные ситуации, с которыми может сталкиваться слабовидящий пользователь при передвижении в помещении и на улице. Источником изображений выступила авторская выборка, составленная из фотографий типовых сцен окружающей среды. В выборку включались изображения, на которых визуально определялся основной контекст сцены и отсутствовали персональные данные, позволяющие идентифицировать людей.

Выборка для задачи классификации сцен включала 50 изображений, распределенных по пяти категориям: школьный коридор, офисное помещение, торговое пространство, уличная сцена и домашний интерьер. Для сохранения баланса классов в каждую категорию было включено по 10 изображений. Такой состав выборки был выбран для проверки способности моделей различать основные типы пространства, значимые для первичного описания окружающей среды в ассистивной системе.

Критериями включения изображений в выборку являлись: наличие отчетливо различимого основного типа сцены, достаточное качество изображения, отсутствие выраженных артефактов сжатия, наличие элементов, характерных для соответствующего класса, а также практическая значимость сцены для задачи ориентации пользователя. В категорию «школьный коридор» включались изображения протяженных внутренних проходов с дверями, стенами и перспективой движения; в категорию «офисное помещение» – рабочие комнаты, кабинеты и пространства с офисной мебелью; в категорию «торговое пространство» – изображения магазинов, торговых залов и проходов между стеллажами; в категорию «уличная сцена» – изображения тротуаров, дорог, дворовых и городских пространств; в категорию «домашний интерьер» – жилые комнаты и бытовые помещения.

Разметка изображений для классификации выполнялась вручную на уровне класса сцены. Каждому изображению присваивалась одна основная метка, соответствующая преобладающему визуальному контексту. В случаях, когда изображение могло быть отнесено сразу к нескольким категориям, например коридор в торговом центре или офисный холл, оно исключалось из выборки либо относилось к классу, который являлся наиболее выраженным с точки зрения задачи навигации. Такой подход позволил уменьшить неоднозначность при оценке качества классификации.

Для OCR-модуля была сформирована отдельная тестовая выборка из 50 изображений с текстовыми фрагментами. В нее включались изображения вывесок, табличек, указателей, коротких надписей и фрагментов печатного текста. Для каждого изображения вручную задавалась эталонная строка распознавания, с которой затем сравнивался результат OCR-систем. Разметка выполнялась в виде пары «изображение – эталонный текст». Перед расчетом метрик проводилась нормализация текста: приведение к единому регистру, удаление лишних пробелов и незначимых различий в пунктуации.

Для детекции объектов и семантической сегментации использовалась группа изображений окружающей среды, отобранная из набора качественной проверки ассистивного конвейера. Для детекции была сформирована пилотная выборка из 30 изображений, на которых выполнена bbox-разметка объектов классов person, vehicle, animal, furniture и obstacle. Всего было размечено 132 объекта. Для семантической сегментации была сформирована пилотная выборка из 18 изображений с грубой mask-разметкой классов floor, road, sidewalk, stairs и obstacle. Разметка использовалась для предварительного расчета mAP, precision, recall, mIoU, pixel accuracy и per-class IoU.

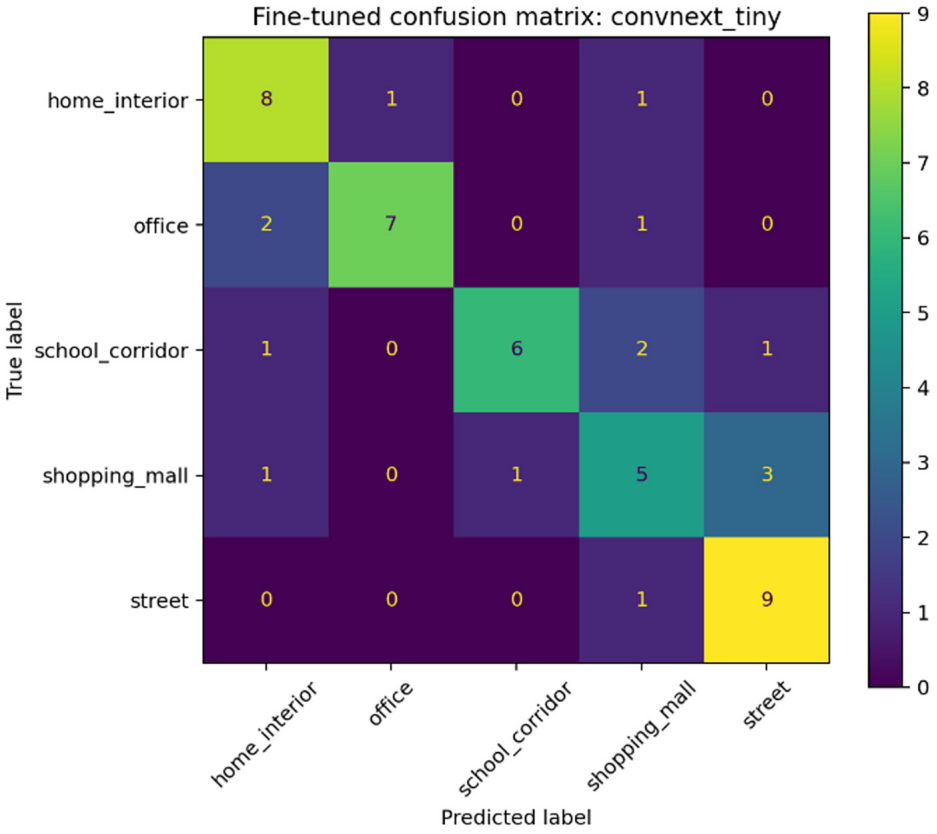
Результаты исследования и их обсуждение

В задаче классификации использовалась независимая тестовая выборка из 50 изображений, включающая 5 категорий: школьные коридоры, офисные помещения, торговые центры, уличные пространства и домашние интерьеры. Для сохранения баланса классов в каждую категорию было включено по 10 изображений. Сравнивались модели EfficientNet-B0, MobileNetV3-Small, ResNet-50 и ConvNeXt-Tiny. Для оценки использовались accuracy, macro-precision, macro-recall, macro-F1, 95 % доверительные интервалы для accuracy, среднее время обработки одного изображения и FPS. Для более детального анализа ошибок классификации для модели с наибольшим значением macro-F1 дополнительно строилась матрица ошибок, отражающая распределение правильных и неправильных предсказаний по каждому классу сцены. Для accuracy рассчитывались 95 % доверительные интервалы по биномиальной модели. Такой подход был выбран из-за малого объема выборки и необходимости учитывать нестабильность метрик при небольшом числе наблюдений в каждом классе. Тестовая выборка не использовалась при обучении и подборе параметров моделей.

Результаты классификации сцен на независимой тестовой выборке

Модель	Accuracy	95 % ДИ для accuracy	Macro-F1	Среднее время обработки, мс	FPS
EfficientNet-B0	0,660	0,522–0,776	0,651	31,60	31,65
MobileNetV3-Small	0,480	0,348–0,615	0,482	10,67	93,76
ResNet-50	0,660	0,522–0,776	0,651	78,60	12,72
ConvNeXt-Tiny	0,700	0,562–0,809	0,699	63,31	15,80

Примечание: составлена авторами на основе полученных данных в ходе исследования.



Матрица ошибок классификации сцен для модели ConvNeXt-Tiny
Примечание: составлен авторами по результатам данного исследования

Эксперименты выполнялись на персональном компьютере без использования CUDA-ускорения: CPU Intel64 Family 6 Model 60, 15,9 ГБ RAM, Windows 10, Python 3.12.5, PyTorch 2.5.1+cpu. Поэтому полученные значения времени инференса следует рассматривать как характеристику работы моделей в условиях CPU-запуска, а не как оценку производительности на мобильном устройстве или GPU.

Результаты классификации показали, что на данной тестовой выборке наибольшее значение macro-F1 получила модель ConvNeXt-Tiny – 0,699 при accuracy 0,700 и среднем времени обработки 63,31 мс на изображение. EfficientNet-B0 и ResNet-50 показали одинаковое значение accuracy – 0,660, однако EfficientNet-B0 обрабатывала изображение быстрее: 31,60 мс против 78,60 мс у ResNet-50. Наиболее быстрой моделью оказалась MobileNetV3-Small – 10,67 мс на изображение, что соответствует 93,76 FPS, однако ее качество классификации было ниже: accuracy 0,480 и macro-F1 0,482 (таблица). С учетом малого объема тестовой выборки и широких доверительных интервалов эти результаты следует рассматривать как предварительное сравнение применимости моделей, а не как статистически доказанное превосходство одной архитектуры.

На рисунке представлена матрица ошибок для модели ConvNeXt-Tiny, показавшей наибольшее значение macro-F1 среди сравниваемых архитектур. Анализ матрицы ошибок показал, что наиболее устойчиво модель распознавала уличные сцены и домашние интерьеры: для классов street и home_interior было получено соответственно 9 и 8 правильных ответов из 10. Более выраженные ошибки наблюдались для классов shopping_mall и school_corridor. Изображения торговых пространств частично классифицировались как street, а изображения школьных коридоров – как shopping_mall, street или home_interior. Это может быть связано с визуальной близостью отдельных категорий: наличием протяженных проходов, дверей, стен, стеллажей, искусственного освещения и перспективы движения.

Следующим этапом обработки является детекция объектов. Если классификатор отвечает на вопрос о типе сцены в целом, то детектор позволяет выделить конкретные элементы окружения: человека, транспортное средство, животное, предмет мебели или иной объект, потенциально влияющий на безопасность движения. Двери, лестницы, границы пола и проходимые поверхности в предлагаемом конвейере целесообразно дополнительно анализировать с использованием семантической сегментации и специализированной разметки.

В качестве модели детекции была рассмотрена архитектура YOLOv8, ориентированная на задачи обнаружения объектов в режиме, близком к реальному времени [12]. Для предварительной количественной оценки детектора была сформирована пилотная bbox-разметка на 30 изображениях, отобранных из набора изображений для качественной проверки детекции. Разметка включала 5 укрупненных классов, значимых для ассистивной навигации: person, vehicle, animal, furniture и obstacle. Всего в пилотной выборке было размечено 132 объекта.

Оценка YOLOv8 на пилотной выборке показала precision 0,663, recall 0,879, mAP@0.5 0,785 и mAP@0.5:0.95 0,620. Наиболее высокое значение AP@0.5 было получено для класса person – 0,940, а также для класса vehicle – 0,853. Более низкие значения наблюдались для классов animal и obstacle, что может быть связано с малым числом объектов этих категорий в пилотной выборке.

Следует учитывать, что данная оценка носит пилотный характер. Bounding boxes были подготовлены как минимальная model-assisted разметка на основе COCO-детекций с последующим сопоставлением с укрупненными ассистивными классами. Поэтому полученные значения mAP, precision и recall следует рассматривать не как итоговый независимый benchmark, а как предварительную количественную проверку возможности включения детектора объектов в общий ассистивный конвейер.

В отличие от классификации и детекции, семантическая сегментация рассматривается в данной работе как основа для анализа проходимой поверхности. Классификация сообщает общий тип сцены, детекция выделяет отдельные объекты, а сегментация позволяет перейти к пиксельной карте пространства. Для ассистивной системы это принципиально важно: пользователю нужно не только знать, что перед ним находится объект, но и знать, где проходит безопасная зона движения, где заканчивается пол или тротуар, где начинается лестница, бордюр, дорога или препятствие.

В предлагаемом подходе сегментационная карта рассматривается как промежуточное представление для последующего математического анализа поверхности. После выделения областей пола, дороги, тротуара, стен, лестниц и препятствий можно рассчитывать признаки, связанные с безопасностью движения: ширину свободной зоны, направление возможного прохода, наличие разрывов поверхности, границы проходимой области и расположение препятствий относительно центральной оси движения пользователя. Такой подход дополняет существующие ассистивные решения, ориентированные преимущественно на чтение текста, распознавание отдельных объектов или общее описание сцены.

С технической точки зрения результат семантической сегментации может быть представлен в виде матрицы классов M(x, y), где каждому пикселю изображения с координатами x и y соответствует определенный класс сцены. В контексте ассистивной навигации интерес представляют не все классы, а прежде всего те, которые связаны с возможностью безопасного движения. Поэтому из общей карты сегментации выделяется множество потенциально проходимых классов P, например floor, sidewalk и road. Пиксели, относящиеся к этим классам, формируют бинарную карту проходимости: B(x , y) (B(x, y) = 1, если M(x, y) принадлежит P; B(x, y) = 0, если M(x, y) не принадлежит P). Далее анализируется связная область карты B(x, y), пересекающая нижнюю центральную часть кадра. Такая область может рассматриваться как ближайшая к пользователю потенциально проходимая зона. Для нее могут рассчитываться ширина свободного прохода, направление центральной линии движения, расстояние до бли-

жайшей границы проходимой области и наличие разрывов, связанных с препятствиями, ступенями или переходом на другой тип поверхности.

Например, если проходимая область сужается в нижней части кадра, это может указывать на наличие препятствий по бокам или на узкий проход. Если область резко прерывается горизонтальной границей, это может соответствовать ступени, бордюру или переходу на непроходимую поверхность. Если центральная линия проходимой области отклоняется вправо или влево, система может использовать это как признак возможного направления безопасного движения. Таким образом, сегментация в предлагаемой системе используется не только для визуального разделения изображения на классы, но и как основа для получения числовых навигационных признаков.

Для семантической сегментации была выбрана модель SegFormer-B0, дообученная на ADE20K. Архитектура SegFormer сочетает иерархический Transformer-энкодер и простой MLP-декодер, что позволяет использовать ее для задач сегментации сцен. Датасет ADE20K содержит плотную разметку сцен и включает 150 классов объектов и областей, что делает его полезной основой для распознавания элементов городской и внутренней среды [13]. В рамках проекта использовалась модель nvidia/segformer-b0-finetuned-ade-512-512.

Для предварительной количественной оценки сегментационного модуля была подготовлена пилотная mask-разметка на 18 изображениях. Разметка включала классы background, floor, road, sidewalk, stairs и obstacle; при расчете mIoU основное внимание уделялось пяти foreground-классам: floor, road, sidewalk, stairs и obstacle. Предсказания SegFormer-B0, дообученной на ADE20K, были сопоставлены с укрупненными ассистивными классами, при этом нерелевантные классы ADE20K относились к background.

На пилотной выборке pixel accuracy составила 0,675, mIoU по foreground-классам – 0,294, а mIoU с учетом background – 0,351. Наиболее высокие значения IoU были получены для классов floor – 0,470, sidewalk – 0,391 и road – 0,374. Более низкие значения наблюдались для obstacle – 0,195 и stairs – 0,039, что может быть связано с визуальной неоднородностью препятствий, малым числом пикселей лестниц и грубым характером пилотной масочной разметки. Отдельным компонентом ассистивной системы является распознавание текстовой информации. Для слабовидящего пользователя текст в окружающей среде может иметь разную значимость: номер кабинета, название улицы, предупреждающая табличка или указатель маршрута могут быть важнее рекламной надписи. Поэтому OCR-модуль в такой системе должен не только извлекать текст, но и передавать его в общий блок приоритизации сообщений. Для обнаружения текстовых строк может применяться CTPN, где текстовая линия формируется как последовательность мелких текстовых предложений на карте признаков [14]. Для последующего распознавания используется CRNN-подход, объединяющий сверточное извлечение признаков и рекуррентную обработку последовательности без необходимости предварительно разделять строку на отдельные символы. Такая схема хорошо согласуется с задачей чтения вывесок, табличек, указателей и предупреждающих надписей в естественной среде.

В экспериментальной части работы для оптического распознавания символов (OCR) были проверены два практических инструмента: EasyOCR и Tesseract. Выбор этих решений связан с их доступностью и возможностью использования в воспроизводимом экспериментальном конвейере [15]. Tesseract является открытым OCR-движком, подробно описанным в литературе. EasyOCR представляет собой готовую библиотеку для извлечения текста из изображений и поддерживает более 80 языков, включая письменности на основе кириллицы. Оценка проводилась на тестовой выборке из 50 изображений с текстовой разметкой. В качестве метрик использовались CER, WER, exact match accuracy, среднее время обработки изображения и FPS.

Для дополнительной проверки устойчивости OCR-модуля была сформирована расширенная мультискриптовая выборка из 50 изображений, включающая кириллические, латинские и смешанные русско-английские надписи. В выборку вошли 17 изображений с кириллическим текстом, 16 изображений с латинским текстом и 17 изображений со смешанными надписями. Для каждого изображения были сформированы варианты с типовыми затрудняющими условиями распознавания: сниженным освещением, бликами, наклоном текста, размытием и уменьшенным размером шрифта. Всего оценивались шесть условий: исходное изображение, сниженное освещение, блики, наклон текста, размытие и малый шрифт. Для EasyOCR использовались языковые режимы en, ru и ru + en, для Tesseract – eng, rus и eng + rus.

Результаты дополнительной проверки показали, что для обеих OCR-систем исходные изображения, сниженное освещение, блики, размытие и уменьшение размера

шрифта не приводили к ошибкам распознавания после нормализации текста. Наиболее сложным условием оказался наклон текста. Для EasyOCR при наклоне CER составил 0,415, WER – 0,420, exact match accuracy – 0,540. Для Tesseract при наклоне CER составил 0,780, WER – 0,900, exact match accuracy – 0,060. При анализе по типу письменности EasyOCR показал CER 0,020 для кириллицы, 0,028 для латиницы и 0,157 для смешанных русско-английских надписей. Для Tesseract соответствующие значения CER составили 0,088; 0,135 и 0,167. Полученные результаты показывают, что смешанный русско-английский текст и геометрический наклон являются наиболее сложными условиями для OCR-модуля.

Интеграция отдельных моделей в ассистивную систему должна строиться с учетом того, что разные типы информации имеют разную ценность для пользователя. Классификатор сцены дает общий контекст, детектор объектов выделяет конкретные предметы, сегментация показывает структуру проходимой поверхности, а OCR-модуль извлекает текст. В простейшем варианте эти результаты могут объединяться в общий список сообщений, где каждому сообщению присваивается приоритет. Например, предупреждение о лестнице, автомобиле или препятствии должно иметь больший приоритет, чем описание удаленной вывески или второстепенного объекта.

Такой подход позволяет избежать избыточного потока голосовой информации. Для пользователя важно не получить максимально подробное описание всей сцены, а своевременно услышать то, что влияет на ориентацию и безопасность движения. Поэтому в дальнейшем целесообразно развивать блок фильтрации сообщений: учитывать тип объекта, расстояние до него, положение в кадре, направление движения пользователя и наличие текстовых предупреждений.

В прототипной реализации ассистивной системы LLM-компонент использовался как завершающий этап обработки данных, предназначенный для объединения результатов отдельных модулей компьютерного зрения в краткую текстовую подсказку для пользователя. В рамках данной работы собственная языковая модель не обучалась, не дообучалась и не разворачивалась локально. Взаимодействие с языковой моделью осуществлялось через внешний API, что позволило проверить сам принцип консолидации данных без разработки отдельной инфраструктуры для локального инференса.

На вход LLM-компонента передавалось не исходное изображение, а структурированное текстовое описание результатов работы других модулей: предполагаемый тип сцены, перечень обнаруженных объектов, сведения о проходимой области, наличие потенциальных препятствий, направление свободного прохода и распознанные OCR-фрагменты. Таким образом, языковая модель не выполняла первичное распознавание визуальной информации, а использовалась как модуль постобработки и формирования итогового сообщения.

Основная задача LLM-компонента заключалась в снижении информационной нагрузки на пользователя. Без этапа консолидации система могла бы последовательно озвучивать все найденные объекты, классы сегментации и текстовые фрагменты, что затрудняло бы восприятие. Использование языковой модели позволяло преобразовать разрозненные результаты в короткую подсказку, ориентированную на безопасность движения. Приоритет в такой подсказке отдавался препятствиям, лестницам, транспортным средствам, изменению поверхности, направлению свободного прохода и предупреждающим надписям.

Например, при наличии результатов классификации сцены, сегментации и OCR система могла сформировать сообщение следующего вида: «Перед вами коридор, свободный проход находится прямо, справа расположена дверь, препятствий впереди не обнаружено». В ситуации потенциальной опасности подсказка должна быть более краткой и приоритетной: «Впереди препятствие на пути, требуется остановиться или обойти слева».

Ограничением данного подхода является зависимость от внешнего API, сетевой задержки, версии используемой языковой модели и корректности входных данных, поступающих от модулей компьютерного зрения. Кроме того, языковая модель может некорректно интерпретировать неполные или противоречивые результаты. Поэтому в ассистивной системе LLM-компонент не должен самостоятельно формировать предупреждения, не подтвержденные результатами детекции, сегментации или OCR. Его роль ограничивается объединением, сокращением и ранжированием уже полученной информации.

Для предварительной оценки практической применимости предложенного конвейера было проведено сценарное тестирование безопасности. Данный этап не являлся полноценным пользовательским исследованием с участием слабовидящих или незрячих пользователей, однако позволил проверить, насколько формируемые системой подсказки соответствуют типовым ситуациям, важным для безопасного передвижения.

Сценарии тестирования были сформированы с учетом повседневных ситуаций, в которых слабовидящему пользователю требуется быстро получить краткую и приоритетную информацию об окружающей среде. Рассматривались следующие группы ситуаций: движение по коридору без препятствий, наличие препятствия в центральной зоне кадра, приближение к лестнице или ступеням, движение вдоль тротуара или дороги, наличие транспортного средства в зоне обзора, распознавание таблички или указателя, а также сцены с несколькими одновременно обнаруженными объектами.

Для каждого сценария задавалось ожидаемое содержание подсказки: указание свободного направления движения, предупреждение о потенциально опасном объекте, сообщение о препятствии, ступенях, дороге, транспортном средстве или значимом тексте. После этого результат, сформированный прототипом, сопоставлялся с ожидаемой подсказкой. Подсказка считалась удовлетворительной, если она не противоречила визуальной ситуации, указывала на основной риск и не содержала избыточного перечисления второстепенных объектов.

Особое внимание уделялось приоритизации сообщений. В сценариях с потенциальной опасностью система должна была формировать краткую предупреждающую подсказку, например: «Впереди препятствие, требуется остановиться» или «Впереди ступени». В менее опасных ситуациях допустимым считалось более описательное сообщение: «Перед вами коридор, свободный проход прямо». Такой подход позволяет оценивать не только факт распознавания отдельных объектов, но и пригодность итоговой подсказки для использования в условиях движения.

Ограничением данного этапа является отсутствие полноценного пользовательского тестирования с участием слабовидящих и незрячих пользователей. Поэтому результаты сценарной проверки не позволяют оценить субъективную понятность сообщений, скорость реакции пользователя, уровень когнитивной нагрузки и удобство системы в реальных условиях. На следующем этапе планируется проведение пользовательского тестирования с участием представителей целевой группы, включающего оценку понятности подсказок, времени реакции, числа ошибочных решений и субъективной удовлетворенности пользователей.

Заключение

Проведенное исследование предварительно показало возможность использования нескольких типов алгоритмов компьютерного зрения в структуре ассистивной системы помощи слабовидящим пользователям. Классификация сцен может применяться для первичного определения общего контекста окружающей среды, детекция объектов – для выделения отдельных значимых предметов и потенциальных препятствий, OCR – для извлечения текстовой информации, а семантическая сегментация – для анализа структуры пространства и проходимых поверхностей. При этом полученные результаты следует рассматривать как предварительную оценку применимости выбранных подходов, поскольку количественная проверка была выполнена не для всех компонентов конвейера.

Количественная экспериментальная оценка в данной версии работы выполнена для двух блоков: классификации сцен и OCR. В задаче классификации наиболее высокое значение macro-F1 на тестовой выборке показала ConvNeXt-Tiny, однако EfficientNet-B0 может рассматриваться как более сбалансированный вариант для первичного этапа ассистивного конвейера за счет меньшего времени обработки при приемлемом качестве классификации. В задаче OCR обе проверенные системы, EasyOCR и Tesseract, не допустили ошибок на исходных изображениях после нормализации текста. Дополнительная проверка на расширенной мультискриптовой выборке показала, что наибольшее снижение качества распознавания связано с наклоном текста, особенно для смешанных русско-английских надписей. Детекция объектов и семантическая сегментация в данной версии исследования были дополнительно проверены на пилотных размеченных выборках. Для детекции была выполнена bbox-разметка и рассчитаны precision, recall и mAP, а для сегментации – mask-разметка, mIoU, pixel accuracy и per-class IoU. При этом именно семантическая сегментация является ключевым направлением дальнейшего развития системы, поскольку она позволяет перейти от простого распознавания объектов к математическому анализу проходимой поверхности.

Кроме того, в программной реализации предусмотрен LLM-компонент, выполняющий консолидацию результатов классификации, детекции, семантической сегментации и OCR в единую краткую подсказку для пользователя. Его использование позволяет перейти от разрозненных выходных данных отдельных моделей к более понятному сообщению, ориентированному на практическую помощь при движении и ориентации в пространстве.

Список литературы

1. Официальный сайт «Всемирная организация здравоохранения». Слепота и нарушение зрения. Данные от 10.02.2026. [Электронный ресурс]. URL: https://www.who.int/ru/news-room/fact-sheets/detail/blindness-and-visual-impairment (дата обращения: 16.06.2026).

2. Болотский Н. Н. Технологии искусственного интеллекта как способ улучшения качества жизни людей с ограниченными возможностями здоровья по зрению // Известия ТулГУ. Технические науки. 2024. № 7. С. 374–376. URL: https://tidings.tsu.tula.ru (дата обращения: 16.06.2026). DOI: 10.24412/2071-6168-2024-7-374-375.

3. Лебедев О. Б., Черкасов Р. И. Применение технологий компьютерного зрения в системах обработки визуальной информации // Известия ЮФУ. Технические науки. 2025. № 5 (247). С. 254–276. URL: https://izv-tn.tti.sfedu.ru (дата обращения: 16.06.2026). DOI: 10.18522/2311-3103-2025-5-254-276.

4. Зубова Д. П., Семенист С. А., Широбокова С. Н. Проектирование компонента классификации объектов и интерпретации их действий с использованием методов компьютерного зрения и машинного обучения // Инженерный Вестник Дона. 2025. № 6 (126). С. 245–252. URL: https://www.ivdon.ru/ru/magazine/archive/n6y2025/10148 (дата обращения: 16.06.2026).

5. Mohammed Fayiz Ferosh. Assistive Technology for Navigation of Visually Impaired People // Journal of Electrical Systems. 2024. Vol. 20. Is. 7. P. 989–996. DOI: 10.52783/jes.3479.

6. Корыткин Н. Г. Применение нейросетевых методов семантической сегментации в системах компьютерного зрения реального времени в условиях ограниченных ресурсов // Труды МАИ. 2026. № 146. URL: https://trudymai.ru/published.php?ID=187461 (дата обращения: 16.06.2026).

7. Pydala B., Kumar P., Baseer K. Kh. VisiSense: A Comprehensive IOT-based Assistive Technology System for Enhanced Navigation Support for the Visually Impaired // Scalable Computing. 2024. Vol. 25. Is. 2. P. 1134–1151. DOI: 10.12694/scpe.v25i2.2619.

8. Messaoudi M. D., Menelas B. A. J., Mcheck H. Review of Navigation Assistive Tools and Technologies for the Visually Impaired // Sensors. 2022. Vol. 22. Is. 20. Art. 7888. URL: https://www.academia.edu/114359090/Review_of_Navigation_Assistive_Tools_and_Technologies_for_the_Visually_Impaired (дата обращения: 16.06.2026). DOI: 10.3390/S22207888.

9. Либерман А. И. Решение задачи классификации объектов на изображении методами компьютерного зрения в различных сферах человеческой деятельности // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2024. № 3. С. 74–91. DOI: 10.17308/sait/1995-5499/2024/3/74-91. EDN: UTQZPZ.

10. Zia Q., Jan A., Yang D., Zhang H., Li Y. Optimized Real-Time Decision Making with EfficientNet in Digital Twin-Based Vehicular Networks // Electronics. 2025. Vol. 14. Is. 6. P. 1084. DOI: 10.3390/electronics14061084.

11. Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., Xie S. A ConvNet for the 2020s // IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. P. 11976–11986. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html (дата обращения: 16.06.2026).

12. Цинь Х., Янь Я., Цзинь Ц., Чье Е. У., Воронин В. В. Алгоритм обнаружения дефектов дорожного полотна на основе модели YOLOV8S // Вестник ТОГУ. 2024. № 2(73). С. 51–62. DOI: 10.38161/1996-3440-2024-2-51-62.

13. Zhou B., Zhao H., Puig X., Fidler S., Barriuso A., Torralba A. Scene Parsing Through ADE20K Dataset // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 633–641. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html (дата обращения: 16.06.2026).

14. Качалин В. С., Панов Ю. Н., Попов Н.-Л. Э. Сравнительный анализ различных систем оптического распознавания символов при работе с текстом, написанным с помощью кириллического алфавита // Современные наукоемкие технологии. 2022. № 8. С. 65–70. URL: https://top-technologies.ru/ru/article/view?id=39268 (дата обращения: 16.06.2026). DOI: 10.17513/snt.39268.

15. Федотова А. М., Романов А. С. Методика идентификации текстов, сгенерированных большими языковыми моделями // Информатика и автоматизация. 2025. Т. 24. С. 1444–1470. URL: https://ia.spcras.ru/index.php/sp/issue/view/836/ia.2025.24.5 (дата обращения: 16.06.2026). DOI: 10.15622/ia.24.5.7.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.