

МУЛЬТИМОДАЛЬНОЕ ОБУЧЕНИЕ СИСТЕМ КОМПЬЮТЕРНОГО ЗРЕНИЯ ДЛЯ КОНТЕКСТНОГО ПОНИМАНИЯ СЦЕН

MULTIMODAL TRAINING OF COMPUTER VISION SYSTEMS FOR CONTEXTUAL SCENE UNDERSTANDING

Морозов В.П., д-р техн. наук, профессор, ФГБОУ ВО ВГТУ (Воронеж, Россия)

Марченко Н.В., студент, ФГБОУ ВО ВГТУ (Воронеж, Россия)

Капацын Д.В., студент, ФГБОУ ВО ВГТУ (Воронеж, Россия)

Morozov V.P., Doctor of Technical Sciences, Professor, Professor, Voronezh State Technical University (Voronezh, Russia)

Marchenko N.V., student, Voronezh State Technical University (Voronezh, Russia)

Kapatsyn D. V., student, Voronezh State Technical University (Voronezh, Russia)

В статье рассматриваются особенности применения мультимодального обучения систем компьютерного зрения (СКЗ) в интересах контекстного понимания сцен. Целью исследования является анализ особенностей применения технологии мультимодального обучения в СКЗ для контекстного понимания сцен. Описаны такие методы мультимодального обучения, как представление модальностей, мультимодальное выравнивание, мультимодальное преобразование, совместное обучение и слияние данных. Представлены результаты экспериментов, подтверждающих более высокую эффективность функционирования СКЗ, в которой для контекстного понимания сцен использовано мультимодальное обучение, по сравнению с унимодальным обучением. Отражены проблемы и современные направления развития мультимодального обучения.

Ключевые слова: аудиовизуальные данные, контекстное понимание сцены, мультимодальное выравнивание, мультимодальное обучение, мультимодальное слияние, представление модальностей, система компьютерного зрения.

Для цитирования: Морозов В.П., Марченко Н.В., Капацын Д.В. Мультимодальное обучение систем компьютерного зрения для контекстного понимания сцен // Информационно-экономические аспекты стандартизации и технического регулирования. 2026. №4 (91). С. 21–26.

The article discusses the features of the application of multimodal learning of computer vision systems (CVS) in the interests of contextual understanding of scenes. The purpose of the study is to analyze the features of using multimodal learning technology in the CVS for contextual understanding of scenes. Such methods of multimodal learning as representation of modalities, multimodal alignment, multimodal transformation, collaborative learning and data fusion are described. The results of experiments confirming the higher efficiency of the functioning of the CVS, in which multimodal learning is used for contextual understanding of scenes, compared with unimodal learning, are presented. The problems and current trends in the development of multimodal learning systems are reflected.

Keywords: audiovisual data, contextual scene understanding, multimodal alignment, multimodal learning, multimodal fusion, presentation of modalities, computer vision system.

For citation: Marchenko N.V., Kapatsyn D.V., Morozov V.P. Multimodal training of computer vision systems for contextual scene understanding. Information and economic aspects of standardization and technical regulation. 2026, 4(91): 21–26. (In Russ.).

ВВЕДЕНИЕ

Основными способами восприятия внешнего мира для каждого человека являются визуальная и слуховая информация. Мозг анализирует информацию, поступающую из внешней асинхронной среды [1], интегрируя разнородные, так называемые, мультимодальные данные (ММД) для достижения определенных целей [2]. Аудиовизуальные способности людей имеют уникальные характеристики, позволяющие из разных сенсорных каналов восприятия объединять ММД об объектах и связях между ними в интересах интерпретации и формирования представлений о действительности, что составляет суть контекстного понимания сцен

(КПС) [3, 4]. В одной ситуации (сцене), человек может услышать звук, определить его приблизительное место возникновения и сопоставить с существующим объектом, заранее не обладая визуальной информацией. В другой ситуации визуальные данные могут быть использованы для предсказания звуковых сигналов. Например, люди способны по движению губ собеседника определить произносимые слова, а по поведению животного – издаваемый им звук.

Эти особенности человека привели к попыткам ученых научить компьютер совершать аналогичные действия – обучить его КПС на основе анализа ММД [5]. Так появилась технология мультимодального обучения (ТММО), одним из важ-

ных направлений которой является мультимодальное обучение (ММО) систем компьютерного зрения (СКЗ) для КПС, выявляемых в ходе анализа видеопотока [6]. Обученные таким образом СКЗ позволяют преодолеть ограничения обработки информации, которые есть у людей. Они связаны с медлительностью подобной обработки и с малыми объемами обрабатываемой информации. Результаты предварительных исследований показали, что СКЗ, использующие ТММО, способны оперативно анализировать большие объемы ММД (например, одновременно обрабатывать видео, аудио и текстовую информацию в системах видеонаблюдения и голосовых ассистентах) [7].

В настоящее время ТММО находится на начальной стадии развития. Число публикаций по применению ТММО в СКЗ весьма ограничено. При этом известно, что сферы, где применяются СКЗ (автопилотный транспорт, медицина и др.), требуют получения высокоточных результатов. В таких областях цена ошибки при распознавании объекта или события бывает достаточно высокой. Эти обстоятельства обуславливают актуальность данной темы.

Цель статьи – раскрыть особенности применения ТММО в системах компьютерного зрения для контекстного понимания сцен.

1. МЕТОДЫ МУЛЬТИМОДАЛЬНОГО ОБУЧЕНИЯ СИСТЕМ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Термин «мультимодальность», как концепция искусственного интеллекта, сформировался в рамках глубокого машинного обучения в 2010 году, а всплеск его бурного развития пришелся на период с 2021 по 2023 год. В эти годы зарубежными компаниями OpenAI, DeepMind и Microsoft Research были разработаны модели CLIP, Flamingo и Kosmos-1, способные анализировать визуально-языковые сцены на основе объединения различных типов данных, таких как текст, изображения, аудио и др. Эти модели могли создавать изображения исходя из их текстового описания, а также описывать в текстовом виде представленные изображения [6]. В интересах обучения искусственных нейронных сетей (ИНС), лежащих в основе этих моделей, впервые было использовано ММО, базирующееся на использовании при обучении ММД.

Применительно к СКЗ процесс ММО нетривиален и базируется на применении в ходе обучения ИНС нескольких методов обработки информации, таких как: представление модальностей данных, мультимодальное выравнивание данных, мультимодальное преобразование данных, совместное обучение и слияние данных. Данные методы обработки информации в совокупности с взаимной дополняемостью ММД обеспечивают более глубокое семантическое понимание сцен по сравнению с унимодальными методами обучения, основанными на одном типе данных.

Суть метода представления модальностей данных заключается в преобразовании (форматировании) данных, поступающих из различных каналов приема (текст, изображение, звук

и др.) в формат, с которым работает (воспринимает) СКЗ [8]. Обработка текстовой информации осуществляется путем векторного представления семантически похожих слов. При этом в качестве обработчика используется система, построенная на предобученной модели типа «трансформер» [9].

На рис. 1 представлена архитектура системы, построенная на модели «Трансформер».

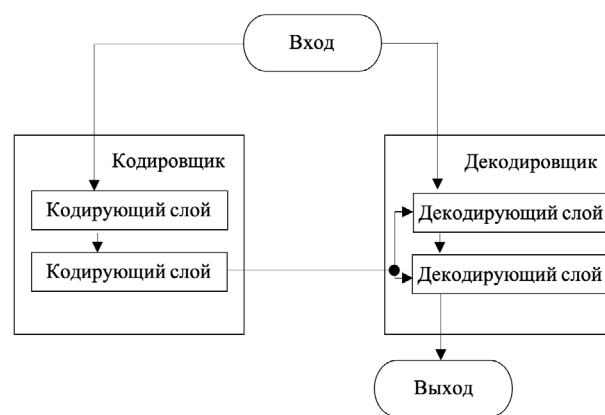


Рис. 1. Архитектура системы, построенная на модели «Трансформер»

Система состоит из двух основных частей: кодировщика и декодировщика.

При обработке текстовой информации кодировщик преобразует входные данные в векторы, формат которых понятен системе, а декодировщик на их основе собирает готовый ответ. Кодировщик и декодировщик работают параллельно, что обеспечивает высокую оперативность обработки информации, а предобученность модели обеспечивает адекватное понимание связи между словами.

Обработка видеoinформации в этой системе осуществляется следующим образом. Видеоданные представляются в виде совокупности пространственных и временных свойств изображения и обрабатываются сверточными нейронными сетями (СНС), а также рекуррентными нейронными сетями (РНС). При этом СНС извлекают пространственные признаки каждого видео, в то время как РНС извлекают информацию между кадрами, получая векторы признаков, представляющих содержимое видео [6]. С помощью сверточных и полносвязных слоев СНС извлекают все данные из изображения – от низкоуровневых признаков, таких как текстура и края, до высокоуровневых семантических признаков, таких как объекты и сцены. Сцена – это функциональное объединение визуальных образов и семантических описаний, где понимание происходящего достигается через интеграцию низкоуровневых признаков внешнего вида с высокоуровневыми представлениями об отношениях и связях между объектами [6]. Содержимое изображения описывается сформированным вектором признаков. Например, при классификации изображений вектор признаков способен

определить категорию изображения, обеспечивая наилучшую точность.

На рис. 2 представлена схема мультимодального совместного представления аудио- и видеoinформации.

На начальном этапе из общего потока аудио- и видеоданных осуществляется извлечение признаков соответствующими методами, изложенными выше. После чего формируется общий мультимодальный вектор, в котором сохраняются данные и устанавливаются связи для каждой модальности. Последующая обработка аудио- и видеоданных позволяет решить несколько задач, например распознать речь и провести анализ настроения человека.

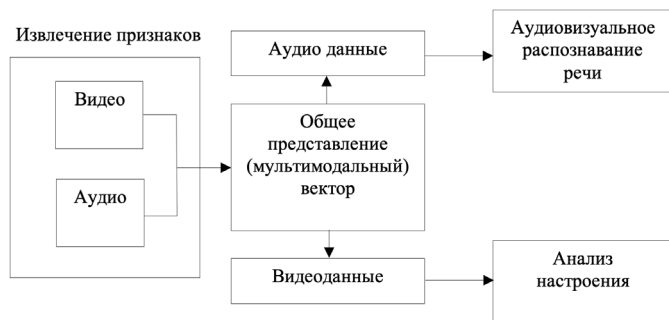


Рис. 2. Схема мультимодального совместного представления аудио- и видеoinформации

Процесс установления соответствия между данными различных типов (например, аудио и видео) называется мультимодальным выравниванием данных. При этом устанавливается, какие элементы одной модальности соответствуют элементам другой. Система распознает, какой звук относится к определенному объекту или действию на видео или изображении. Выравнивание бывает неконтролируемым и контролируемым. В контролируемом выравнивании применяют специально размеченные данные, которые указывают, какой звук соответствует определенному объекту. Данный метод позволяет сделать процесс обучения более легким и улучшить точность определения соответствий, что обуславливает предпочтение его применения в СКЗ. В неконтролируемом выравнивании используется другой метод – модель сама ищет связи данных различных модальностей, опираясь на закономерности или структуру, при этом заранее размеченные данные не используются [6]. При этом сама модель усложняется и становится более ресурсоемкой, требующей дорогостоящего оборудования [10].

Мультимодальное слияние данных представляет собой объединение информации из различных источников (например, аудио и видео) в единое целое. Полученное представление используется моделью для анализа и принятия решений. Основная идея заключается в том, чтобы использовать взаимодополняющие данные разных модальностей для повышения точности и полноты восприятия. Выделяют

три основных уровня мультимодального слияния данных: на уровне данных; на уровне признаков; на уровне решений.

Мультимодальное слияние на уровне данных осуществляется на самой ранней стадии их обработки. В этом случае исходные сведения из различных источников объединяются (например, аудио и видео объединяются в один поток), после чего обрабатываются как единый блок.

Мультимодальное слияние на уровне признаков реализуется после предварительной обработки данных. Каждая модальность обрабатывается отдельно, из нее извлекаются характерные признаки, после чего они образуют единый вектор, который используется для последующего анализа.

Мультимодальное слияние на уровне решений применяется на финальной стадии. Каждая модальность обрабатывается отдельно и формирует собственный результат (например, отдельный прогноз). Затем полученные итоги объединяются, и на их основе принимается окончательное решение. Различие между уровнями слияния заключается в том, на каком этапе происходит объединение информации: до обработки, в процессе извлечения признаков или уже после получения отдельных результатов [6].

В СКЗ находят применение все три уровня мультимодального слияния данных. Однако предпочтение отдается первым двум в силу их более глубокой методической проработки для практического использования.

2. ПРИМЕНЕНИЕ МУЛЬТИМОДАЛЬНОГО ОБУЧЕНИЯ ДЛЯ КОНТЕКСТНОГО ПОНИМАНИЯ СЦЕН В СИСТЕМАХ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Для КПС в СКЗ наиболее важным является семантическое описание изображения, то есть выделение значимых элементов сцены и связей между ними. В прикладных задачах, решаемых СКЗ, этого часто достаточно для выполнения конкретных действий. Например, в автопилотном транспорте для безопасной езды, системе СКЗ нужно не просто восстанавливать геометрию определенной сцены, а выделять дорогу, обочину, пешеходов и препятствия.

На практике семантическое описание изображения предполагает его трехуровневую обработку. Фактически, каждый уровень обработки сцены включает определенный перечень задач и является определенным уровнем ее понимания [11].

Первый уровень охватывает базовую обработку первичных данных. На этом уровне решаются задачи фильтрации шумов, заполнения пропусков в данных, детекции движений и первичной агрегации аудио- и видеосигналов.

На втором уровне осуществляется более глубокий анализ происходящих событий. Система переходит от регистрации сигналов к выделению конкретных объектов и источников звука.

Третий уровень отвечает за когнитивную интерпретацию и установление причинно-следственных связей. На этом

уровне анализируется взаимное влияние объектов, прогнозируется развитие ситуации и формируются ответы на сложные запросы по содержанию сцены (например, как объехать препятствие на дороге, как обогнать автомобиль и др.).

Рассмотрим СКЗ, способную выделять определенные объекты и источники звука, то есть базирующуюся на втором уровне понимания сцены. В ней реализовано раннее лексическое обучение на основе данных, поступающих по двум перекрестным каналам: видео и аудио [6]. Кроме того, в данной системе используется адаптивный метод анализа данных (анализ данных с динамическими весами), позволяющий использовать при обучении данные обоих каналов, отдавая приоритет данным из тех каналов, в которых они менее зашумлены. Объединение данных и возможность переключения (перераспределения весов при анализе) на более помехоустойчивые каналы в зависимости от складывающейся обстановки позволяют формировать устойчивые концепты, в которых один признак (например, звуковой) подкрепляет визуальный образ. Это существенно повышает точность распознавания сцены.

Результаты проведенных экспериментов, которые представлены на рис. 3, показывают, что использование естественных аудиовизуальных пар позволяет идентифицировать около 70–75% объектов даже в условиях высокого зашумления, тогда как работа только с одним типом данных (например, только с видео) приводит к ухудшению результата (точность распознавания сцены менее 40%).

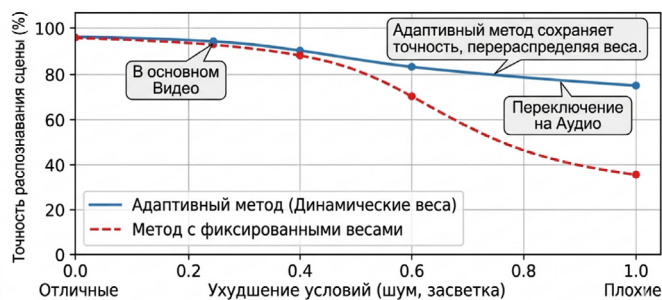


Рис. 3. Сравнение адаптивного метода и метода с фиксированными весами при разных условиях

Статический метод анализа с фиксированными весами (полученные результаты на его основе показаны на рисунке красной пунктирной линией) демонстрирует резкую деградацию точности (в пределах 60%). При достижении критического уровня помех его эффективность падает ниже 40%, так как в системе продолжает использоваться видеоканал со значительными искажениями.

Адаптивный метод с динамическими весами (синяя сплошная линия), напротив, демонстрирует для плохих условий гораздо меньшее ухудшение точности (в пределах 18%). Благодаря перераспределению приоритетов в пользу аудиоканала при зашумлении видеоданных точность распозна-

вания удерживается на уровне выше 70% даже в самых неблагоприятных сценариях.

В рамках практической реализации СКЗ критически важно не просто суммировать данные, а динамически оценивать надежность (качество данных) каждого канала. Например, если изображение «размыто» или ослеплено светом, СКЗ должна автоматически повышать вес аудиоинформации.

Демонстрационный пример функционирования СКЗ, основанной на втором уровне понимания сцены, в плане выбора более надежного канала для предотвращения ошибки в оценке складывающейся дорожной ситуации на автомобильной трассе приведен на рис. 4.

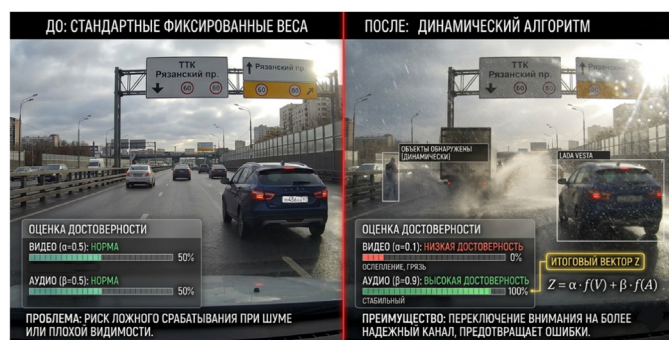


Рис. 4. Пример выбора более надежного канала для предотвращения ошибки в оценке складывающейся дорожной ситуации на автомобильной трассе

На левом снимке данного рисунка показана достаточно простая ситуация, в которой для оценки достоверности дорожной ситуации, видео- и аудиоканалы равнозначны. На правом снимке из-за возникшей аномалии, приведшей к снижению видимости по видеоканалу, для оценки достоверности дорожной ситуации предпочтение отдано аудиоканалу.

ЗАКЛЮЧЕНИЕ

Таким образом, в интересах повышения точности СКЗ в предметных областях (например, в автопилотном транспорте, в медицине и др.), где цена ошибки при распознавании объекта или события достаточно высока, для КПС необходимо использовать ММО.

При этом необходимо учитывать следующие особенности применения ММО.

1. Процесс ММО нетривиален и базируется на применении таких методов обработки информации, как: представление модальностей данных, мультимодальное выравнивание данных, мультимодальное преобразование данных, совместное обучение и слияние данных. Эти методы обеспечивают более глубокое семантическое понимание сцен по сравнению с унимодальными методами обучения, основанными на одном типе данных.
2. В рамках реализации метода представления модальностей данных в качестве обработчика данных целесообразно

использовать предобученную модель типа «трансформер», способную оперативно обрабатывать разнородные данные и обеспечивающую наилучшую точность при работе с зашумленными данными.

3. При выборе вида мультимодального выравнивания данных следует отдать предпочтение контролируемому выравниванию, оперирующему размеченными данными, что обуславливает простоту его реализации и меньшую ресурсоемкость по сравнению с неконтролируемым выравниванием.

4. Предпочтительным результатом мультимодального преобразования данных является вектор, в котором сохраняются данные и устанавливаются связи для каждой модальности.

5. В СКЗ могут использоваться три уровня мультимодального слияния данных (на уровне данных; на уровне признаков; на уровне решений). Однако предпочтение следует отдать первым двум уровням в силу их более глубокой методической проработки для практического использования.

Результаты проведенных экспериментов и демонстрационный пример показывают более высокую эффективность функционирования СКЗ, в которой для анализа КПС использовано ММО, по сравнению с унимодальным обучением.

Основными проблемами в данной предметной области являются: во-первых, проблема точной синхронизации аудио- и видеоданных, во-вторых, сложность выделения причинно-следственных связей между событиями. Именно эти проблемы сейчас остаются наименее проработанными и требуют дальнейшего изучения.

Рецензент: Бурый Алексей Сергеевич, доктор технических наук, заместитель начальника научного отдела, ФГБУ «Институт стандартизации», Москва, Российская Федерация. E-mail: a.s.burij@gostinfo.ru

Список литературы

1. Морозов В.П., Белоусов В.Е., Мистров Л.Е., Сырин А.И. Свойства и информационные процессы мозга человека для разработки нейроморфных систем // Информационно-экономические аспекты стандартизации и технического регулирования. 2023. № 5(74). С. 48–55.
2. Сысков А. М., Борисов В. И., Петренко Т. С. Проектирование мультимодальных интерфейсов мозг-компьютер. – Екатеринбург: Изд-во Урал. ун-та, 2023. – 131 с.
3. Потапов А.С. Системы компьютерного зрения. Учебное пособие. – СПб: Университет ИТМО, 2016. – 161 с.
4. Рой Д., Пентланд А. Обучение словам на основе зрительных и звуковых данных: вычислительная модель // Cognitive Science. 2002. Т. 26, №1. С. 113–146.
5. Крейман Г. Биологическое и компьютерное зрение. – М.: ДМК Пресс, 2021. – 314 с.
6. Дэвис Р., Тёрк М. Компьютерное зрение. Передовые методы и глубокое обучение. – М.: Издательство «ДМК Пресс», 2022. – 690 с.
7. Иванько Д.В., Рюмин Д.А. Обзор нейросетевых методов распознавания и синтеза звучащей и жестовой речи // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 5. С. 669–686.
8. Морозов В.П., Белоусов В.Е., Суханова Д.А., Суханов В.В. Облик информационных систем верификации субъектной информации // Информационно-экономические аспекты стандартизации и технического регулирования. 2025. № 1(82). С. 57–62.
9. Гринчук О.В., Цурков В.И. Обучение мультимодальной нейронной сети для определения подлинности

References

1. Morozov V.P., Belousov V.E., Mistrov L.E., Syrin A.I. Svoystva i informacionnye processy mozga cheloveka dlya razrabotki nejromorfnyh sistem. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya. 2023, no. 5(74), pp. 48–55. (In Russ.).
2. Syskov A. M., Borisov V. I., Petrenko T. S. Proektirovanie mul'timodal'nyh interfejsov mozg-komp'yuter. Ekaterinburg: Izd-vo Ural. un-ta, 2023, 131 p. (In Russ.).
3. Potapov A.S. Sistemy komp'yuternogo zreniya. Uchebnoe posobie. SPb: Universitet ITMO, 2016, 161 p. (In Russ.).
4. Roj D., Pentland A. Obuchenie slovam na osnove zritel'nyh i zvukovyh dannyh: vychislitel'naya model'. Cognitive Science. 2002, vol. 26, no, pp. 113–146. (In Russ.).
5. Krejman G. Biologicheskoe i komp'yuternoe zrenie. Moscow: DMK Press, 2021, 314 p. (In Russ.).
6. Devis R., Tyork M. Komp'yuternoe zrenie. Peredovye metody i glubokoe obuchenie. Moscow: Izdatel'stvo «DMK Press», 2022, 690 p. (In Russ.).
7. Ivan'ko D.V., Ryumin D.A. Obzor nejrosetevykh metodov raspoznavaniya i sinteza zvuchashchej i zhestovoj rechi. Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki. 2024, vol. 24, no. 5, pp. 669–686. (In Russ.).
8. Morozov V.P., Belousov V.E., Suhanova D.A., Suhanov V.V. Oblik informacionnyh sistem verifikacii sub'ektnoj informacii. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya. 2025, no. 1(82), pp. 57–62. (In Russ.).
9. Grinchuk O.V., Curkov V.I. Obuchenie mul'timodal'noj nejronnoj seti dlya opredeleniya podlinnosti

изображений // Известия РАН. Теория и системы управления. 2020. № 4. С. 103–109.

10. Морозов В.П., Белоусов В.Е., Моисеев С.И. Техническое обеспечение информационных систем компьютерного зрения // Информационно-экономические аспекты стандартизации и технического регулирования. 2022. № 4(68). С. 35–41.
11. Борисова И.В. Компьютерное зрение: цифровая обработка и анализ изображений. – Новосибирск: Изд-во НГТУ, 2022. – 163 с.

izobrazhenij. Izvestiya RAN. Teoriya i sistemy upravleniya. 2020, no 4, pp. 103–109. (In Russ.).

10. Morozov V.P., Belousov V.E., Moiseev S.I. Tekhnicheskoe obespechenie informacionnyh sistem komp'yuternogo zreniya. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya. 2022, no. 4(68), pp. 35–41. (In Russ.).
11. Borisova I.V. Komp'yuternoe zrenie: cifrovaya obrabotka i analiz izobrazhenij. Novosibirsk: Izd-vo NGTU, 2022. 163 p. (In Russ.).