МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра математической кибернетики и компьютерных наук

РАЗРАБОТКА СИСТЕМЫ КОМПЬЮТЕРНОГО ЗРЕНИЯ ДЛЯ РАСПОЗНАВАНИЯ АНИМАЦИОННЫХ ПЕРСОНАЖЕЙ И ИЗВЛЕЧЕНИЯ МЕТАДАННЫХ НА ОСНОВЕ МУЛЬТИМОДАЛЬНОГО АНАЛИЗА ВИДЕО

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 5 курса 551 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Новоселова Александра Константиновича

Научный руководитель
к. ф.-м. н., доцент _____ А. С. Иванов

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. В последние годы технологии компьютерного зрения активно применяются в прикладных задачах анализа изображений и видеоданных. Наиболее заметный прогресс в этой области связан с развитием глубоких нейронных сетей и свёрточных архитектур, позволяющих решать задачи классификации изображений, локализации объектов и детекции на сложных сценах [1]. Это делает возможным создание программных систем, которые не только отображают визуальный контент, но и автоматически интерпретируют его содержимое.

Для современных мультимедийных приложений всё более востребованными становятся сценарии, в которых пользователь получает дополнительную информацию о содержимом кадра непосредственно во время просмотра видео. В случае анимационного видеоконтента такой задачей является распознавание персонажа на стоп-кадре с последующим выводом связанных сведений. Подобный подход объединяет методы компьютерного зрения, обработку видеоданных и извлечение метаданных, что соответствует современному направлению развития интеллектуальных пользовательских приложений [2].

Задача распознавания анимационных персонажей имеет ряд особенностей. В отличие от фотографических изображений реальных объектов, анимационные персонажи являются стилизованными визуальными образами. Их внешний вид зависит от художественного стиля, ракурса, позы, масштаба, освещения, композиции кадра и степени детализации. Дополнительную сложность создают визуальная близость классов, ограниченность специализированных датасетов и различие между подготовленными изображениями из выборки и кадрами реального видео [3].

В связи с этим разработка системы, способной по стоп-кадру обнаруживать персонажа, распознавать его и выводить связанные метаданные, представляет практический интерес. Такая система демонстрирует возможность перевода методов компьютерного зрения из исследовательского формата в форму прикладного программного продукта, ориентированного на понятный пользовательский сценарий.

Степень разработанности темы определяется активным развитием методов классификации изображений, объектной детекции, локализации и мультимодального анализа данных. Методологические основы работы представлены в

исследованиях I. Goodfellow, Y. Bengio, A. Courville, C. M. Bishop, A. Krizhevsky, K. He, а также в работах, посвящённых современным архитектурам компьютерного зрения и мультимодальному обучению. Среди русскоязычных источников использовались работы Э. Д. Шакирьянова и С. И. Николенко, рассматривающие практические аспекты компьютерного зрения и глубокого обучения.

Цель бакалаврской работы — разработка системы компьютерного зрения для распознавания анимационных персонажей и извлечения метаданных на основе мультимодального анализа видео.

Поставленная цель определила следующие задачи:

1. проанализировать основные подходы к распознаванию изображений, локализации объектов и мультимодальному анализу данных;
2. рассмотреть особенности распознавания анимационных персонажей как отдельного класса визуальных объектов;
3. спроектировать архитектуру системы, обеспечивающей получение стоп-кадра, обнаружение персонажа, его классификацию и извлечение метаданных;
4. выбрать технологии и средства программной реализации системы;
5. подготовить и исследовать датасеты для задач классификации и детекции объектов;
6. разработать и исследовать собственную свёрточную нейронную сеть для распознавания персонажей;
7. выполнить сравнительный анализ собственной модели и современных предобученных архитектур классификации;
8. реализовать модуль обнаружения персонажа на стоп-кадре;
9. реализовать механизм извлечения метаданных о персонаже и актёре озвучивания с учётом выбранной аудиодорожки;
10. разработать графический пользовательский интерфейс приложения;
11. провести экспериментальное исследование качества работы разработанной системы.

Объектом исследования являются методы и программные средства анализа визуального содержимого анимационного видеоконтента.

Предметом исследования являются алгоритмы распознавания анимационных персонажей, методы выделения области интереса на изображении и способы извлечения связанных метаданных на основе результатов классификации

и выбранной аудиодорожки.

В работе используются методы компьютерного зрения, машинного обучения, обработки изображений и видеоданных, проектирования программных систем и экспериментальной оценки моделей. В качестве инструментальной базы применяются Python, PyTorch, torchvision, ultralytics, OpenCV, FFmpeg/ffprobe, pandas, matplotlib, scikit-learn, JSON-хранилище метаданных и PySide6 для разработки графического интерфейса.

Теоретическая значимость бакалаврской работы заключается в систематизации подходов к распознаванию анимационных персонажей и описании связки локализации, классификации и извлечения метаданных как единого пользовательского сценария. В работе сопоставлены особенности собственной свёрточной модели и современных предобученных архитектур применительно к задаче распознавания стилизованных персонажей.

Практическая значимость бакалаврской работы состоит в создании действующего программного прототипа, позволяющего пользователю открыть видео, выбрать интересующий кадр, выполнить обнаружение и распознавание персонажа, а затем получить связанную информацию, включая имя персонажа и актёра озвучивания в зависимости от выбранной аудиодорожки. Полученные результаты могут быть использованы при разработке обучающих, мультимедийных и справочных приложений.

Структура и объём работы. Бакалаврская работа состоит из введения, трёх разделов, заключения, списка использованных источников и трёх приложений. Общий объём работы — 77 страниц, из них 66 страниц — основное содержание, включая 10 рисунков и 8 таблиц. Список использованных источников информации содержит 27 наименований.

0.1 КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретические основы распознавания анимационных персонажей и извлечения метаданных из видеоконтента» посвящён анализу предметной области. В нём рассмотрены основные задачи компьютерного зрения при обработке изображений и видеоданных: классификация изображений, локализация объектов, детекция и работа со стоп-кадром как самостоятельной единицей анализа.

В подразделе 1.1 описаны задачи компьютерного зрения, применимые к обработке изображений и видеоданных. Показано, что для разрабатываемой

системы наиболее важны выделение области интереса, классификация визуального объекта и обработка текущего кадра видео как отдельного изображения.

В подразделе 1.2 рассмотрены свёрточные нейронные сети и их роль в задачах распознавания изображений. Описана идея автоматического выделения признаков, операция свёртки, иерархическое представление признаков, а также значение современных глубоких архитектур для задач компьютерного зрения. Отдельно отмечено, что CNN хорошо подходят для распознавания анимационных персонажей, поскольку способны учитывать форму, контуры, цветовые сочетания и устойчивые детали визуального образа.

В подразделе 1.3 проанализированы особенности распознавания анимационных персонажей. Показано, что такие изображения отличаются высокой степенью стилизации, вариативностью ракурса и позы, визуальной близостью классов, сложностью фона и ограниченностью специализированных наборов данных. Сделан вывод, что распознавание персонажа на стоп-кадре должно рассматриваться не как изолированная классификация изображения, а как часть более широкого пользовательского сценария.

В подразделе 1.4 рассмотрены подходы к локализации объектов на изображении. Описано различие между локализацией и детекцией, а также роль ограничивающей рамки как способа выделения области интереса. Приведена метрика Intersection over Union, используемая для оценки качества совпадения предсказанной и эталонной рамок.

В подразделе 1.5 описан мультимодальный подход к извлечению информации из видеоконтента. В рамках работы мультимодальность реализуется через объединение визуального распознавания персонажа и выбора связанных метаданных с учётом аудиодорожки. Подчёркнуто, что аудиодорожка используется как параметр контекста, а не как объект анализа звукового сигнала.

Второй раздел «Проектирование системы распознавания анимационных персонажей и извлечения метаданных» посвящён постановке задачи и проектированию программной системы. В нём формализован пользовательский сценарий, описана архитектура приложения, выбран технологический стек, подготовлены данные и определены критерии оценки качества.

В подразделе 2.1 сформулирована задача разработки системы. Входными данными являются текущий видеокадр, выбранная аудиодорожка и локальный набор метаданных. Выходными данными являются ограничивающая рамка

найденного персонажа, имя распознанного класса и связанные сведения о персонаже и актёре озвучивания. Задача описана как последовательность этапов: получение кадра, обнаружение персонажа, выделение области интереса, классификация и извлечение метаданных.

В подразделе 2.2 выполнена формализация сценария работы системы. Сценарий представлен как цепочка преобразований F ? B ? C ? K ? D, где F — текущий видеокадр, B — ограничивающая рамка, C — вырезанная область интереса, K — класс персонажа, D — итоговые данные, отображаемые пользователю. Такая схема позволяет явно показать зависимость результата от корректности каждого этапа обработки.

В подразделе 2.3 описана архитектура системы. Она построена как модульная структура, включающая подсистему работы с видео, модуль детекции, модуль классификации, модуль метаданных, управляющий контур и графический интерфейс. Такое разделение позволяет независимо развивать отдельные компоненты и при необходимости заменять модели без полной переработки приложения.

В подразделе 2.4 обоснован выбор технологий. В качестве базового языка реализации выбран Python. Для обучения и применения моделей используются PyTorch и torchvision, для детекции — ultralytics и модели YOLOv8, для работы с видео — OpenCV, для получения информации об аудиодорожках — ffprobe, для хранения метаданных — JSON, для анализа результатов — pandas, matplotlib и scikit-learn, для графического интерфейса — PySide6.

В подразделе 2.5 рассмотрена подготовка данных. Для классификации использовался набор изображений покемонов, организованный по классам персонажей. Для детекции применялся отдельный датасет с разметкой ограничивающих рамок. В работе учитывалось, что классификация и детекция требуют различной структуры данных: в первом случае важна метка класса, во втором — координаты объекта на изображении.

В подразделе 2.6 спроектирована модель метаданных. Она предназначена для хранения сведений о персонажах и вариантах озвучивания. Результат классификации используется как ключ для обращения к метаданным, а выбранная аудиодорожка уточняет, какая именно запись должна быть показана пользователю.

В подразделе 2.7 определены критерии оценки качества. Для классифи-

кации используются accuracy, top-k accuracy и матрица ошибок. Для детекции применяются precision, recall, mAP50 и mAP50-95. Также отмечено, что итоговая система должна оцениваться не только по отдельным моделям, но и по успешности всей цепочки обработки: обнаружение персонажа, правильная классификация и корректное извлечение метаданных.

Третий раздел «Реализация, обучение и экспериментальное исследование системы» посвящён практической разработке приложения, обучению моделей и проверке работоспособности итогового решения.

В подразделе 3.1 описана структура программной системы. Проект включает скрипты подготовки данных, обучения и оценки классификационных моделей, обучения детекторов, конфигурационные файлы, сохранённые модели, модуль метаданных и графическое приложение. Такая структура позволяет отделить исследовательские эксперименты от итоговой прикладной реализации.

В подразделе 3.2 реализован модуль распознавания персонажа. Были исследованы собственные свёрточные архитектуры и предобученные модели. В сравнении участвовали custom CNN, ResNet50, EfficientNet-B0, ConvNeXt Tiny, улучшенная собственная CNN и EfficientNet-B1. Наилучший результат показала EfficientNet-B1: точность на тестовой выборке составила 0,9626, а top-3 accuracy — 0,9951. Улучшенная собственная CNN также показала работоспособность и достигла test accuracy 0,6325 и top-3 accuracy 0,8081, однако уступила предобученной модели.

В подразделе 3.3 реализован модуль обнаружения персонажа на кадре. В ходе экспериментов были протестированы YOLOv8s, YOLOv8m и Faster R-CNN ResNet50 FPN. Лучший результат показала YOLOv8s: precision 0,9784, recall 0,9773, mAP50 0,9882 и mAP50-95 0,6932. На основании этих метрик и практической пригодности для пользовательского приложения YOLOv8s была выбрана как финальный детектор.

В подразделе 3.4 описана организация конвейера обработки видеокадра. После остановки видео пользователь запускает распознавание. Система получает текущий кадр, передаёт его в детектор, вырезает область интереса, выполняет классификацию, обращается к метаданным и формирует результат для интерфейса. Такая схема превращает отдельные модели машинного обучения в связанный прикладной механизм.

В подразделе 3.5 представлена разработка графического интерфейса при-

ложения. Интерфейс позволяет открыть видеофайл, управлять воспроизведением, выбрать аудиодорожку, запустить распознавание, увидеть ограничивающую рамку найденного персонажа, имя класса, confidence, top-3 варианты и сведения об озвучивании. Это делает систему пригодной для демонстрации и практического использования без обращения к отдельным скриптам.

В подразделе 3.6 выполнен анализ работы разработанной системы. Показано, что итоговое приложение объединяет детектор, классификатор, модуль метаданных и видеоподсистему в единый программный прототип. Система способна распознавать персонажа по стоп-кадру, выделять область его расположения и выводить связанную информацию. В качестве сильных сторон отмечены модульная архитектура, возможность независимого тестирования компонентов и понятный пользовательский сценарий. К ограничениям отнесены обработка одного кадра по запросу пользователя, ограниченность базы метаданных и необходимость дальнейшего развития сценариев с несколькими объектами на одном кадре.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы были получены следующие основные результаты:

1. исследованы теоретические основы компьютерного зрения применительно к распознаванию анимационных персонажей: классификация изображений, локализация объектов, детекция, особенности стилизованных изображений и мультимодальное извлечение информации из видеоконтента;
2. сформулирована постановка задачи и разработана архитектура системы, в которой обработка стоп-кадра выполняется как последовательность этапов: получение кадра, обнаружение персонажа, выделение области интереса, классификация, извлечение метаданных и отображение результата;
3. подготовлены и использованы отдельные наборы данных для классификации и детекции, что позволило разделить задачи распознавания класса персонажа и выделения его положения на полном кадре;
4. разработана и исследована собственная свёрточная нейронная сеть, а также выполнено сравнение с современными предобученными архитектурами. Наилучшее качество классификации показала EfficientNet-B1 с test accuracy 0,9626 и top-3 accuracy 0,9951;
5. реализован и исследован модуль детекции персонажа. По результатам сравнения YOLOv8s, YOLOv8m и Faster R-CNN ResNet50 FPN в качестве финальной модели выбрана YOLOv8s, показавшая mAP50 0,9882 и mAP50-95 0,6932;
6. реализован модуль метаданных, позволяющий сопоставлять распознанного персонажа с дополнительной информацией и учитывать выбранную пользователем аудиодорожку;
7. разработано настольное приложение на PySide6, объединяющее работу с видео, запуск распознавания, визуализацию ограничивающей рамки, вывод имени персонажа, оценки уверенности и сведений об озвучивании.

Таким образом, цель выпускной квалификационной работы достигнута. Разработана программная система, способная обнаруживать анимационного персонажа на видеокадре, определять его класс и отображать связанные с ним метаданные в удобной для пользователя форме. Полученное решение представляет собой не только набор обученных моделей, но и завершённый программ-

ный прототип, демонстрирующий применение методов компьютерного зрения в мультимедийном пользовательском приложении.

Перспективы дальнейшего развития связаны с расширением набора распознаваемых персонажей, пополнением базы метаданных, поддержкой нескольких объектов на одном кадре, улучшением устойчивости на сложных сценах, добавлением анализа последовательности кадров и более глубокой интеграцией аудио- и текстовых модальностей.

1 Основные источники информации:

1. A review of convolutional neural networks in computer vision [Электронный ресурс]. — URL: https://link.springer.com/article/10.1007/s10462-024-10721-6 (дата обращения: 04.01.2026).
2. A Survey of Multimodal Learning: Methods, Applications, and Future [Электронный ресурс]. — URL: https://dl.acm.org/doi/10.1145/3713070 (дата обращения: 12.04.2026).
3. Goodfellow I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. — Cambridge: The MIT Press, 2015. — 801 p.
4. Bishop C. M. Pattern Recognition and Machine Learning. — New York: Springer, 2006. — 758 p.
5. Khan S. A Guide to Convolutional Neural Networks for Computer Vision / S. Khan, H. Rahmani, A. A. Shah. — San Rafael: Morgan and Claypool Publishers, 2018. — 160 p.
6. Krizhevsky A. ImageNet Classification with Deep Convolutional Neural Networks [Электронный ресурс]. — URL: https://proceedings.neurips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (дата обращения: 05.01.2026).
7. He K. Deep Residual Learning for Image Recognition [Электронный ресурс]. — URL: https://arxiv.org/abs/1512.03385 (дата обращения: 12.01.2026).
8. A Survey of Modern Deep Learning based Object Detection Models [Электронный ресурс]. — URL: https://arxiv.org/abs/2104.11892 (дата обращения: 14.04.2026).
9. Шакирьянов Э. Д. Компьютерное зрение на Python. Первые шаги. — М.: Лаборатория знаний, 2021. — 160 с.
10. Николенко С. И. Глубокое обучение: погружение в мир нейронных сетей / С. И. Николенко, А. А. Кадурин, Е. О. Архангельская. — СПб.: Питер, 2025. — 480 с.