ՀԱՅԱՍՏԱՆԻ ՀԱՆՐԱՊԵՏՈՒԹՅԱՆ ԿՐԹՈՒԹՅԱՆ, ԳԻՏՈՒԹՅԱՆ,
ՄՇԱԿՈՒՅԹԻ ԵՎ ՍՊՈՐՏԻ ՆԱԽԱՐԱՐՈՒԹՅՈՒՆ

ՀԱՅԱՍՏԱՆԻ ԱԶԳԱՅԻՆ ՊՈԼԻՏԵԽՆԻԿԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

Հարությունյան Էդուարդ Անդրանիկի

ՏԵՍԱՆՅՈՒԹՈՒՄ ԶԳԱՑՄՈՒՆՔՆԵՐԻ ՀԱՅՏՆԱԲԵՐՄԱՆ ԱՎՏՈՄԱՏԱՑՄԱՆ ՄԻՋՈՑՆԵՐԻ ՄՇԱԿՈՒՄԸ

Ե.13.02 «Ավտոմատացման համակարգեր» մասնագիտությամբ
տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի
հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

Երևան 2025

МИНИСТЕРСТВО ОБРАЗОВАНИЯ, НАУКИ, КУЛЬТУРЫ И СПОРТА
РЕСПУБЛИКИ АРМЕНИЯ

НАЦИОНАЛЬНЫЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ АРМЕНИИ

Арутюнян Эдуард Андраникович

РАЗРАБОТКА СРЕДСТВ АВТОМАТИЗАЦИИ ОБНАРУЖЕНИЯ ЭМОЦИЙ В ВИДЕОМАТЕРИАЛЕ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата
технических наук по специальности 05.13.02-
“Системы автоматизации”

Ереван 2025

Ատենախոսության թեման հաստատվել է Հայաստանի ազգային պոլիտեխնիկական համալսարանում (ՀԱՊՀ)։

Գիտական ղեկավար՝ տ.գ.դ. Վազգեն Շավարշի Մելիքյան

Պաշտոնական ընդդիմախոսներ՝ տ.գ.դ. Հայկ Ստեփանի Սուքիասյան
ֆ.-մ.գ.թ. Աշոտ Վալերիի Մինասյան

Առաջատար կազմակերպություն՝ ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտ

Ատենախոսության պաշտպանությունը կայանալու է 2025թ. հուլիսի 16-ին, ժամը 10^{00}-ին, ՀԱՊՀ-ում գործող «Կառավարման և ավտոմատացման» 032 մասնագիտական խորհրդի նիստում (հասցեն՝ 0009, Երևան, Տերյան փ., 105, 17 մասնաշենք)։

Ատենախոսությանը կարելի է ծանոթանալ ՀԱՊՀ– ի գրադարանում։

Սեղմագիրն առաքված է 2025թ. հունիսի 13-ին։

032 Մասնագիտական խորհրդի գիտական քարտուղար, տ.գ.թ. Անուշ Վազգենի Մելիքյան

Тема диссертации утверждена в Национальном политехническом университете Армении (НПУА)

Научный руководитель: д.т.н. Вазген Шаваршович Меликян

Официальные оппоненты: д.т.н. Айк Степанович Сукиасян
к.ф.-м.н. Ашот Валерьевич Минасян

Ведущая организация: Институт проблем информатики и автоматизации НАН РА

Защита диссертации состоится 16-го июля 2025г. в 10^{00} ч. на заседании Специализированного совета 032 — "Управления и автоматизации", действующего при Национальном политехническом университете Армении, по адресу: 0009, г. Ереван, ул. Теряна, 105, корпус 17.

С диссертацией можно ознакомиться в библиотеке НПУА.

Автореферат разослан 13-го июня 2025 г.

Ученый секретарь
Специализированного совета 032, к.т.н. Ануш Вазгеновна Меликян

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В условиях стремительного развития цифровых технологий и беспрецедентного роста видеоматериалов на медиа-платформах задача автоматического обнаружения эмоций из видеоматериалов приобрела особую важность. Автоматическое обнаружение и анализ эмоциональных проявлений в настоящее время занимает центральное место в научных исследованиях, что обусловлено большим потенциалом их применения для решения многочисленных практических задач. В этой области существует ряд фундаментальных сложностей, включая культурные и индивидуальные особенности, технические ограничения и трудности понимания контекста.

Несмотря на значительные инвестиции в этой области, существующие решения все еще не полностью удовлетворяют современным требованиям. Методы, основанные на визуальных характеристиках (ВХ), часто ограничиваются только анализом выражений лица; методы звуковых характеристик (ЗХ) нестабильны по отношению к шумам, а текстовой анализ (ТХ) затрудняется при обнаружении многозначных выражений. Существующие мультимодальные решения либо игнорируют важные модальности, уступая в точности, либо требуют чрезмерно больших вычислительных ресурсов.

Диссертация посвящена разработке такого средства автоматизации обнаружения эмоций из видеоматериалов, которое обеспечит высокую точность путем генерации описательных текстов из визуальных и звуковых характеристик, их совместного анализа с текстом диалога, одновременно снижая затраты вычислительных ресурсов. Предлагаемые решения направлены на преодоление ограничений существующих методов и обеспечение создания эффективного средства обнаружения эмоций.

Объект исследования. Средства автоматизации обнаружения эмоций в видеоматериалах и методы получения текстовых описаний из аудиовизуальных характеристик.

Цель работы. Разработка способов и средств повышения эффективности обнаружения эмоций языковыми моделями посредством совместного анализа аудиовизуальных характеристик и текста диалога.

Методы исследования. В ходе выполнения диссертации были использованы алгоритмы глубокого обучения, методы мультимодального анализа, языковые модели, технологии искусственного интеллекта (ИИ) и специально разработанные программные инструменты для обнаружения эмоций из видеоматериалов. Также применялись современные методы извлечения признаков, предварительной обработки данных и генерации текста.

Научная новизна:

- Предложены подходы к разработке средств автоматизации обнаружения эмоций из видеоматериалов, которые благодаря генерации описательных текстов из визуальных и звуковых характеристик, совместному анализу этих текстов и текста диалога, а также применению эффективной языковой модели удовлетворяли бы современным требованиям с точки зрения точности результатов и затрат вычислительных ресурсов.

- Разработан метод генерации описательного текста из визуальных характеристик видеоматериалов, обеспечивающий в результате применения предварительных этапов обработки по очистке шумов и выбору ключевых кадров сокращение количества рассматриваемых кадров на 96,6%, а также ускорение этого процесса в два раза за счет потери некоторых деталей (BLEU-4: 0,46 и METEOR: 0,38), сохраняя при этом общую семантическую точность описаний (CIDEr-D: 0,89 и Wu-P: 0,81).
- Создан метод генерации описательного текста из звуковых характеристик видеоматериалов, благодаря которому путем применения этапов предварительной обработки по очистке звуковых шумов и повышению качества в среднем улучшены основные показатели оценки качества описаний на 13,5% за счет дополнительных временных затрат в 9,42%.
- Предложен механизм совместного анализа описательных текстов, сгенерированных из различных модальностей, и текста диалога, обеспечивающий в результате применения оптимизации TensorRT и стратегии последовательной загрузки ускорение модели Gemma в 3,2 раза, сокращение использования памяти в 2,7 раза за счет снижения точности всего на 0,8% и показателя w-F1 на 1,2%. Данный механизм превзошел существующие решения в среднем на 8,15% за счет вычислительной сложности параллельного применения трех отдельных моделей.

Практическая ценность работы. Разработано программное средство (ПС) обнаружения эмоций в видеоматериала “ERC System”, которое внедрено в ООО “Тутор Платформ” и успешно применяется для эмоционального анализа аудиовизуального контента. Благодаря внедрению микросервисной архитектуры и потоковой обработки обеспечены модульность системы и возможность легкой замены отдельных компонентов без необходимости полного перепроектирования системы. Оценка эффективности программного средства в реальных условиях показала, что благодаря автоматическому распознаванию речи и автоматической идентификации говорящих ПС применимо без наличия эталонных данных за счет потери точности предложенного метода на 9,5%.

На защиту выносятся:

- Метод генерации описательного текста из визуальных характеристик видеоматериалов.
- Метод генерации описательного текста из звуковых характеристик видеоматериалов.
- Метод совместного анализа описательных текстов, сгенерированных из различных модальностей, и текста диалога.
- Программный инструмент автоматизированного обнаружения эмоций из видеоматериала “ERC System”.

Достоверность научных положений. Достоверность научных результатов подтверждена экспериментальными результатами программной реализации средств, представленных в диссертации, и математическими обоснованиями.

Внедрение. Программный инструмент “ERC System” внедрен в ООО “Тутор Платформ” и используется для эмоционального анализа содержания различных видеоматериалов с целью многослойного представления данных на основе

различных модальностей, что позволило упростить процесс анализа видеоматериалов и обнаружения эмоций.

Апробация работы. Основные научные и практические результаты диссертации докладывались на:

- Международном симпозиуме "IEEE East-West Design & Test Symposium (EWDTS)" (Батуми, Грузия, 2023 г.);
- Международном симпозиуме "IEEE East-West Design & Test Symposium (EWDTS)" (Ереван, Армения, 2024 г.);
- научных семинарах кафедры "Информационные технологии и автоматизация" НПУА (Ереван, Армения, 2022–2024 гг.);
- научных семинарах кафедры "Микроэлектронные схемы и системы" НПУА (Ереван, Армения, 2022 - 2024 гг.).

Публикации. Положения, представленные в диссертации, обобщены в девяти научных статьях, список которых представлен в конце автореферата.

Структура и объем диссертации. Работа состоит из введения, трех глав, основных выводов, списка литературы из 127 наименований и трех приложений. В первом приложении представлен акт внедрения диссертации, во втором - дополнительные описания наборов данных, промежуточные результаты и фрагменты функциональной структуры реализованного программного средства “ERC System”, в третьем - списки использованных рисунков, таблиц и сокращений. Объем диссертации составляет 132 страниц, а вместе с приложениями - 154 страниц. Диссертация написана на армянском языке.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, сформулированы цель и основные задачи исследования, представлены разработанные методы, научная новизна, практическое значение и основные научные положения, выносимые на защиту.

В первой главе представлены основные методы автоматизированного обнаружения эмоций в видеоматериалах, рассмотрены различные модальности анализа эмоциональных проявлений, изучены существующие подходы к извлечению визуальных, звуковых и текстовых характеристик (ВХ, ЗХ и ТХ).

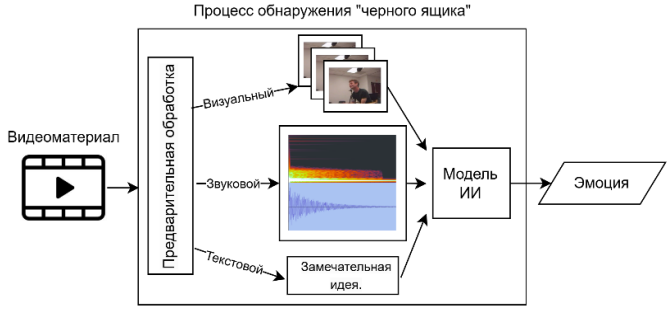

Рис. 1. Схема мультимодального метода автоматического обнаружения эмоций в видео с помощью ИИ

Проанализированы современные мультимодальные системы, объединяющие разные источники данных. Исследованы недостатки существующих методов, включая ограниченность одномодальных подходов, вычислительную сложность и недостаточную точность в реальных условиях. Исходя из этих проблем, эффективность обнаружения эмоций в значительной степени зависит от правильного выбора предварительной обработки данных, отбора модальностей и стратегий их интеграции (рис. 1).

Визуальные характеристики играют ключевую роль в выражении и восприятии эмоций в процессе человеческой коммуникации. Методы обнаружения эмоций из визуальных характеристик основаны на нескольких основных источниках: макро-выражения лица, микровыражения лица и язык тела (рис. 2).

Рис. 2. Источники обнаружения эмоций из ВХ

Микровыражения представляют собой непроизвольные, кратковременные движения лица, которые могут указывать на скрытые эмоции. Автоматическое обнаружение этих выражений является сложной задачей из-за их кратковременности и трудной различимости. В разных исследованиях были использованы гибридные архитектуры MobileViT, объединяющие легкие сверточные нейронные сети (СНС) с преимуществами визуальных трансформеров (ВТ).

Язык тела также является важным индикатором эмоционального состояния. Для автоматического распознавания движений тела предложены методы, основанные на трехмерных СНС и рекуррентных нейронных сетях для анализа временных характеристик.

Наиболее часто исследуются макровыражения лица, так как они четче выражают эмоциональное состояние человека и сравнительно легко поддаются автоматическому анализу. Предложенные методы сосредоточены на различных архитектурах глубокого обучения с широким применением СНС, эффективных для извлечения признаков из изображений. Часто используются предварительно обученные сети, такие как ResNet и VGG, которые переобучаются для задачи классификации эмоций.

Более современным подходам является применение ВТ для обнаружения эмоций. В отличие от СНС, которые анализируют локальные участки, трансформер исследует взаимосвязь всех элементов выражения лица, разбивая изображение на мелкие части и изучая связи между ними с помощью механизма внимания.

Аудиальные характеристики играют важную роль в процессе выражения и восприятия эмоций в человеческой коммуникации. Методы обнаружения эмоций из звуковых характеристик основаны на нескольких основных признаках: просодических, акустических и спектральных (рис. 3).

Просодические признаки считаются одним из основных компонентов систем обнаружения эмоций. Эти признаки характеризуют интонационные особенности речи, такие как тон, интенсивность (громкость голоса), энергия и длительность.

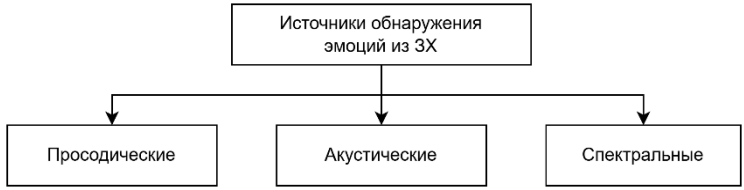

Рис. 3. Подходы к обнаружению эмоций из ЗХ

Спектральные признаки вычисляются с помощью системы голосового тракта и представляют краткосрочные характеристики речевого сигнала. Наиболее распространенными в этой группе являются мел-частотные кепстральные коэффициенты и коэффициенты линейного предсказания, которые эффективны для анализа эмоциональной речи, моделируя особенности человеческой слуховой системы.

Акустические признаки дополняют две вышеуказанные группы, включая характеристики качества голоса. В этой группе особенно важны нестабильность колебаний голосовых связок и флуктуации амплитуды, которые позволяют обнаружить нюансы, характерные для эмоциональной речи.

Текстовые характеристики играют важную роль в процессе выражения и восприятия эмоций в человеческой коммуникации. Методы обнаружения эмоций из текстовых характеристик основаны на нескольких основных подходах: лексических, методах ИИ и гибридных подходах (рис. 4).

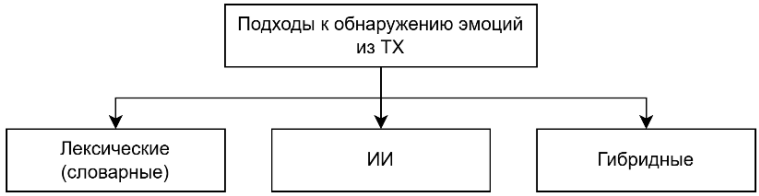

Рис. 4. Признаки обнаружения эмоций из ТХ

Лексические подходы состоят из списка слов, каждому из которых присваивается числовое значение, соответствующее эмоции. Эти подходы делятся на две подгруппы: словарные и основанные на корпусе. В словарном подходе поддерживается словарь ключевых слов, в то время как подход, основанный на корпусе, представляет собой коллекцию текстов данного языка.

В отличие от лексических методов, подходы, основанные на ИИ, могут автоматически изучать сложные закономерности из больших объемов данных. Задача обнаружения эмоций в диалогах (ОЭД) является одним из наиболее сложных проявлений задачи обнаружения эмоций, требующей не только отдельных эмоциональных выражений, но и учета полного контекста разговора анализа.

Мультимодальные характеристики (МХ) обеспечивают более полные и точные результаты в процессе выражения и восприятия эмоций в человеческой коммуникации. Методы обнаружения эмоций из мультимодальных характеристик основаны на комбинированном анализе визуальных, звуковых и текстовых характеристик (рис. 5).

Рис. 5. Подходы к обнаружению эмоций из МХ

Предложены различные инновационные подходы, такие как механизм совместного перекрестного внимания, комбинация графовых нейронных сетей и архитектуры трансформеров. Эти методы позволили одновременно моделировать как временные зависимости во время разговора, так и взаимодействия между различными модальностями, обеспечивая более высокую точность обнаружения эмоций.

Современные методы обнаружения эмоций, несмотря на значительный прогресс последних лет, всё ещё сталкиваются с рядом существенных ограничений. Методы, основанные на анализе визуальных характеристик, в основном ограничиваются изучением выражений лица, часто игнорируя движения тела. Методы анализа звуковых характеристик показывают высокую чувствительность к окружающим шумам и часто не способны всесторонне анализировать все типы характеристик. В области текстового анализа существующие методы затрудняются обнаруживать двусмысленные выражения, а мультимодальные подходы сталкиваются с проблемами вычислительной сложности.

Во второй главе представлены разработанные методы и даются решения проблем, описанных в первой главе.

Метод сокращения объема обрабатываемых данных путем предварительной очистки шумов и выбора ключевых кадров за счет генерации описательного текста из визуальных характеристик видеоматериалов

Разработанный метод основан на последовательной обработке видеоматериалов с целью выделения ключевой информации и сокращения объема анализируемых данных. Алгоритм выбора ключевых кадров (КК) (рис. 6) начинается с предварительного выбора кадров, включающего расчет номер анализируемого кадра в секунду (НАКС) (3 кадров/с) и получение кадров с интервалами НАКС. Далее производится расчет гистограммы, который включает преобразование кадров в пространство HSV, расчет гистограмм в 32-битном интервале и расчет разницы между гистограммами. После этого выполняется расчет порогового значения (ПЗ), основанный на вычислении среднего значения различий и стандартного отклонения. Определение ключевых кадров происходит путем

сравнения разницы с ПЗ: если разница превышает ПЗ, кадр добавляется в список КК, в противном случае происходит пропуск кадра. Процесс очистки шумов (рис. 7) включает настройку параметров (выбор размера фильтра, определение окна поиска и адаптацию параметра h), а затем непосредственную очистку шумов с применением нелокального усреднения, поиском цветовых волн и сохранением контуров изображения. Финальный этап обработки включает сегментацию видео, преобразование в цифровые векторы и формирование команды для модели Apollo-1.5B, которая генерирует текстовое описание визуальных характеристик (рис. 8).

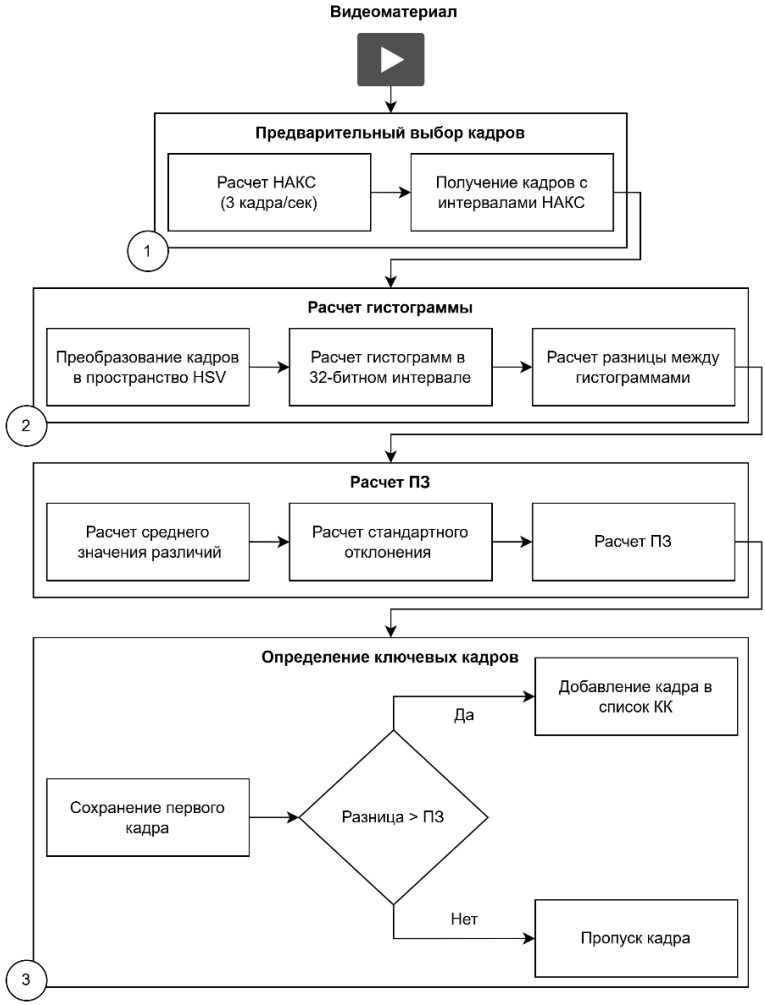

Рис. 6. Последовательность шагов по определению КК из видеоматериалов

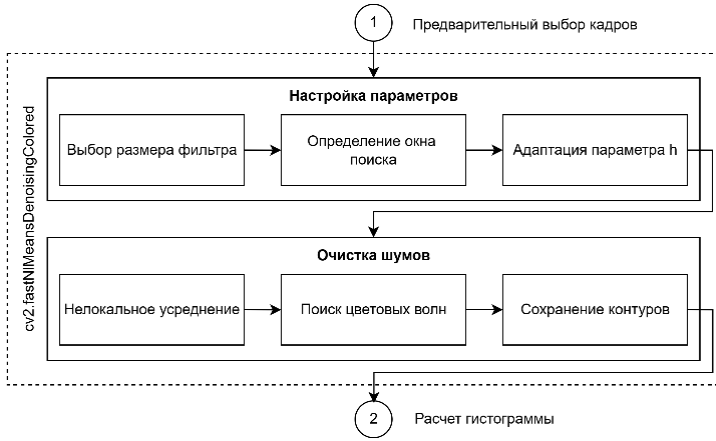

Рис. 7. Процесс очистки шумов в кадрах

Эксперименты на дата-сетах SIDD и MELD подтвердили эффективность предложенного метода. Анализ результатов показал значительное сокращение количества обрабатываемых кадров на 96,61% (табл. 1) при сохранении высокого качества описаний (CIDEr-D: 0,89, Wu-P: 0,81) и двукратное ускорение процесса генерации текстовых описаний.

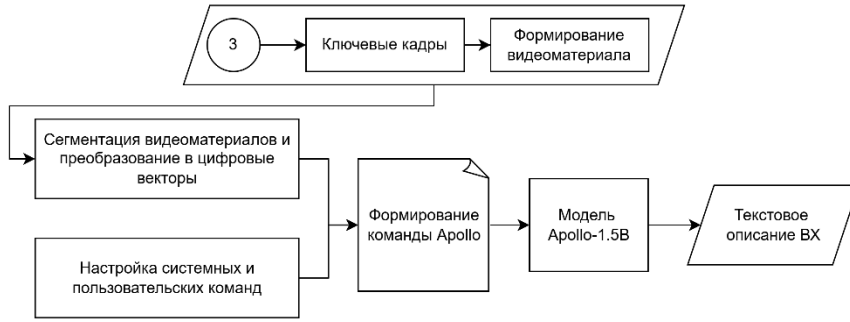

Рис. 8. Получение текстовых описаний визуальных характеристик с помощью модели Apollo

Таблица 1

Результаты сокращения количества кадров

Кадры	Количество	Сокращение (%)
Все	83117	-
Предварительно выбранные	12361	85,13
Ключевые (без очистки шумов)	2862	96,56
Ключевые (с очисткой шумов)	2820	96,61

Метод повышения качества текстовых описаний путем предварительной обработки и очистки шумов за счет генерации описательного текста из звуковых характеристик видеоматериалов

Метод обработки звуковых характеристик включает две основные фазы: очистку аудиошумов и повышение качества сигнала. Как показано на рис. 9, процесс начинается с выделения аудиосигнала из видеоматериалов, затем производится очистка шумов, включающая обнаружение голосовой активности, широкополосное шумоподавление и устранение эха. Следующий этап – повышение качества сигнала путем расширения частотного диапазона, изменения частоты дискретизации и компрессии динамического диапазона.

Финальный этап (рис. 10) включает экспорт характеристик, генерацию мел-спектрограммы, использование большой языковой модели (БЯМ) Whisper, которая после декодирования цифровых векторов генерирует текстовое описание звуковых характеристик.

Эффективность метода подтверждена экспериментами на дата-сетах SpEAR и Clotho. Результаты показали значительное улучшение качества описаний по всем метрикам в среднем на 13,5% (табл. 2) при увеличении времени обработки на 9,42%.

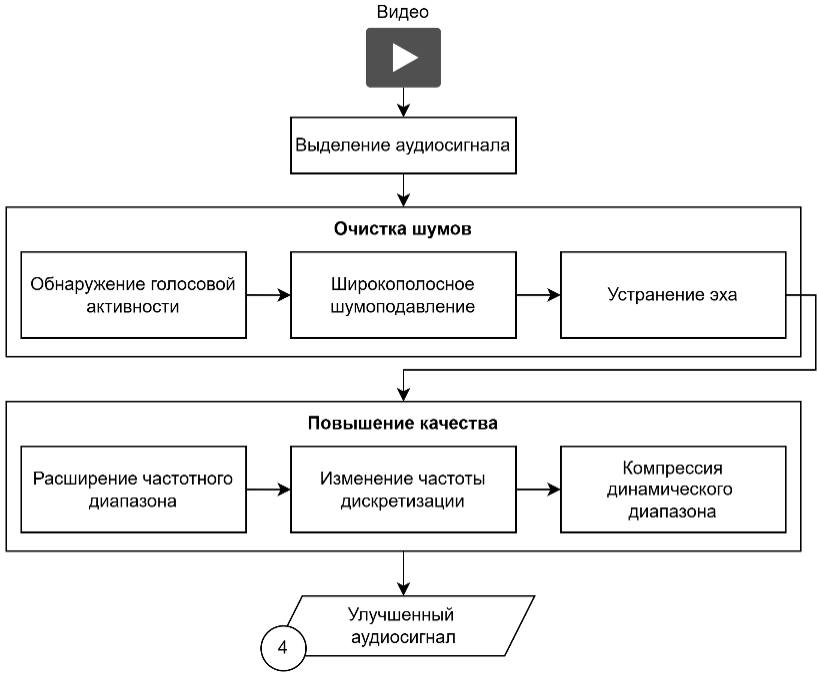

Рис. 9. Процесс очистки шумов и повышения качества в звуковом сигнале

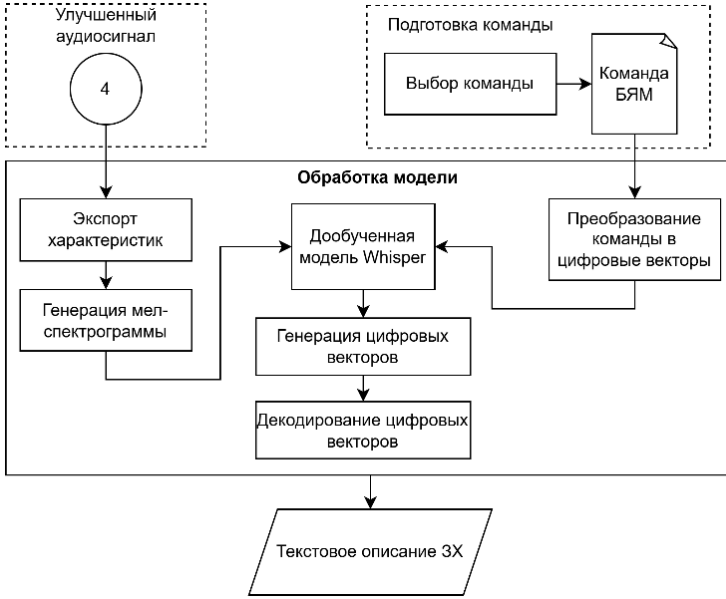

Рис. 10. Способ получения текстового описания звуковых характеристик из видеоматериалов

Таблица 2

Сравнительные результаты качества текстовых описаний

Метрика	Исходный		С предобработкой	
	Среднее	Макс.	Среднее	Макс.
CIDEr-D	0,41	0,45	0,46	0,51
BLEU-4	0,32	0,35	0,36	0,4
METEOR	0,37	0,41	0,42	0,46
ROUGE-L	0,48	0,53	0,54	0,6
SPICE	0,12	0,13	0,13	0,15
Wu-P	0,72	0,79	0,82	0,9

Метод оптимизации вычислительных ресурсов с применением TensorRT и стратегии последовательной загрузки путем совместного анализа мультимодальных описательных текстов

Предложенный механизм основан на комплексном анализе описательных текстов, полученных из различных модальностей видеоматериалов, с применением оптимизированных языковых моделей. Как показано на рис. 11, метод объединяет

три основных источника информации: визуальный описательный текст, аудио-описательный текст и текст диалога.

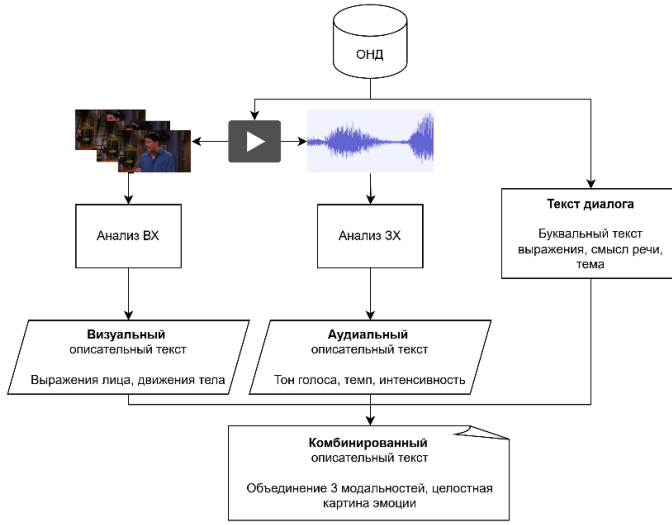

Рис. 11. Комбинирование описаний различных модальностей

Для создания объединенного набора данных (ОНД) произведено слияние двух стандартных наборов данных - MELD и IEMOCAP, с гармонизацией их структуры и содержания. Процесс унификации включал стандартизацию форматов данных, согласование классов эмоций и сбалансированное распределение данных по разделам обучения, валидации и тестирования. Особое внимание уделялось равномерному распределению эмоциональных классов между разделами набора данных.

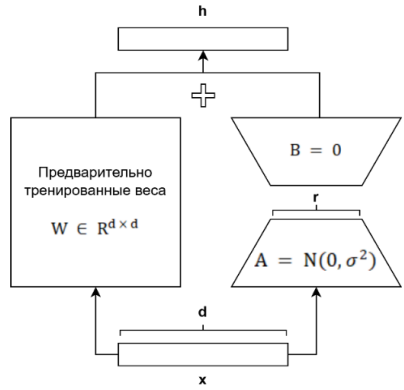

Рис. 12. Метод дообучения LORA

Для анализа мультимодальных описательных текстов проведено исследование трех малых БЯМ: Gemma-2B, Llama-3.2-1B и Phi-2. Сравнительный анализ показал, что Gemma-2B демонстрирует наилучшие результаты благодаря своей архитектуре и большому контекстному окну (8192 токена). Дальнейшая оптимизация параметров модели LORA (r=32, α=64) позволила снизить значение функции потери до 0,468 (рис. 12).

Для повышения производительности модели Gemma применена оптимизация с использованием TensorRT (рис. 13), включающая конвертацию в формат ONNX, квантизацию параметров до FP16 и построение оптимизированного движка. Это позволило ускорить модель в 3,2 раза (с 112 мс до 35 мс на вывод) и сократить использование памяти в 2,7 раза (с 4,6 ГБ до 1,7 ГБ) при минимальной потере точности.

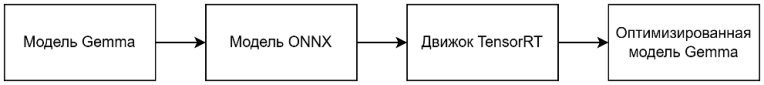

Рис. 13. Оптимизация модели Gemma с помощью TensorRT

Оценка эффективности предложенного метода проводилась с использованием стандартных метрик: точность и взвешенная F1-мера (w-F1-score). Результаты показали, что базовая модель Gemma достигает точности 78,2% и w-F1 76,5%, а оптимизированная версия - 77,4% и 75,3% соответственно (табл. 3).

Таблица 3

Показатели предложенной модели

Модель	**ВХ**	**ЗХ**	**ТХ**	**ОНД (%)**	
				Точность	**w-f1**
Gemma	✓	✓	✓	78,2	76,5
Оптимизированная Gemma	✓	✓	✓	77,4	75.3

Таблица 4

Сравнение с известными моделями

Модель	**ВХ**	**ЗХ**	**ТХ**	**MELD (%)**		**IEMOCAP (%)**	
				Точность	**w-f1**	**Точность**	**w-f1**
DialogueLLM	✓	✗	✓	**71,9**	**71,9**	**70,6**	69,9
InstructERC	✗	✗	✓	-	69,1	-	**71,3**
MMATERIC	✗	✓	✓	66,6	65,2	70,5	70,5
CORECT-6	✓	✓	✓	-	-	69,9	70
ChatGPT	✗	✗	✓	-	63	-	58

Сравнение с существующими решениями показало, что предложенный метод превосходит аналоги в среднем на 8,15% (табл. 4), несмотря на вычислительную сложность параллельного применения трех отдельных моделей. Для решения

проблемы ресурсоемкости внедрена стратегия последовательной загрузки моделей (Apollo → Whisper → оптимизированная Gemma), что позволило значительно снизить требования к вычислительным ресурсам при сохранении высокой точности анализа.

В третьей главе представлено разработанное программно-инструментальное средство “ERC System”, реализующее методы, предложенные во второй главе. Система автоматизирует весь процесс обнаружения эмоций от загрузки видеоматериалов до представления результатов.

Программное обеспечение имеет модульную архитектуру, основанную на шаблоне проектирования “модель-представление-контроллер”, что обеспечивает четкое разделение между данными, бизнес-логикой и пользовательским интерфейсом. Такой подход повышает гибкость, модульность и расширяемость системы. В основе системы лежат три ключевые модели: Gemma-2B для совместного анализа мультимодальных данных и обнаружения эмоций, Apollo-1.5B для генерации описательного текста из визуальных характеристик и Whisper для анализа звуковых характеристик.

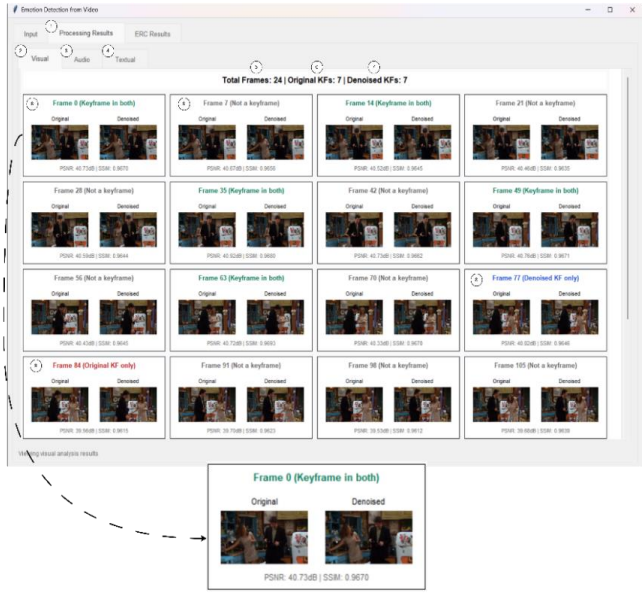

Рис. 14. Подраздел визуальной обработки видеоматериалов

Процесс работы системы включает четыре основных этапа. Первый этап включает обнаружение и сегментацию выражений после предоставления входного файла. На этом этапе выполняется обнаружение речевой активности с помощью модели Whisper, что позволяет выделить участки видеоматериалов, содержащие речь, и разделить их на отдельные выражения. Также с помощью той же модели

выполняется идентификация говорящих, что важно для задачи распознавания эмоций в диалогах (РЭД). Второй этап фокусируется на генерации описательных текстов из визуальных и звуковых характеристик для всех сегментированных выражений. Третий этап объединяет выходные данные предыдущих этапов для РЭД с мультимодальной командой. Последний этап включает реализацию РЭД с мультимодальной командой с использованием дообученной модели Gemma-2B.

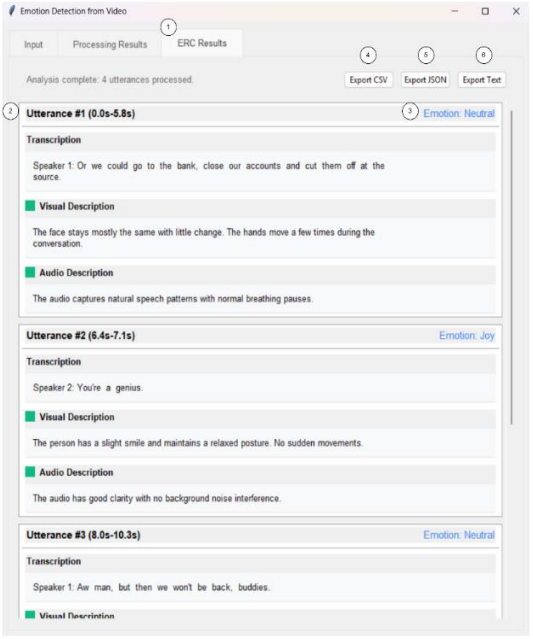

Рис. 15. Раздел результатов РЭД

Интуитивно понятный графический интерфейс системы разработан с учетом удобства использования и включает три основных раздела: “Ввод”, “Результаты обработки” и “Результаты РЭД”. В разделе “Ввод” пользователь может загружать видеофайлы и просматривать их технические характеристики (рис. 14). Раздел “Результаты обработки” содержит три подраздела: “Визуальный”, “Звуковой” и “Текстовой”, которые позволяют оценить качество обработки соответствующих характеристик. В разделе “Результаты РЭД” отображаются итоговые результаты системы, предоставляя для каждого разговорного элемента эмоциональную классификацию и соответствующие описания (рис. 15). Пользователь может экспортировать результаты в различных форматах, включая CSV, JSON и TXT.

Для оценки эффективности проведено тестирование на объединенном наборе данных, где система самостоятельно выполняла все этапы обработки (табл. 5). Снижение эффективности программного средства (с 77,4% до 67,8% по точности) обусловлено неточностями в определении границ выражений и ошибками при идентификации говорящих.

Таблица 5

Оценка программного средства

Метод	ВХ	ЗХ	ТХ	ОНД (%)	
				Точность	w-f1
Применение предложенного метода	✓	✓	✓	77,4	75,3
Применение программного средства	✓	✓	✓	67,8	65,4

Несмотря на это, система демонстрирует достаточный уровень точности для практического применения в реальных условиях.

ОСНОВНЫЕ ВЫВОДЫ ПО ДИССЕРТАЦИОННОЙ РАБОТЕ

1. Предложены подходы к разработке средств автоматизации обнаружения эмоций из видеоматериалов, которые благодаря генерации описательных текстов из визуальных и звуковых характеристик, совместному анализу этих текстов и текста диалога, а также применению эффективной языковой модели удовлетворяли бы современным требованиям с точки зрения точности результатов и затрат вычислительных ресурсов. [1, 6]
2. Разработан метод генерации описательного текста из визуальных характеристик видеоматериалов, обеспечивающий в результате применения предварительных этапов обработки по очистке шумов и выбору ключевых кадров сокращение количества рассматриваемых кадров на 96,6%, а также ускорение этого процесса в два раза за счет потери некоторых деталей (BLEU-4: 0,46 и METEOR: 0,38), сохраняя при этом общую семантическую точность описаний (CIDEr-D: 0,89 и Wu-P: 0,81). [4]
3. Создан метод генерации описательного текста из звуковых характеристик видеоматериалов, благодаря которому путем применения этапов предварительной обработки по очистке звуковых шумов и повышению качества в среднем улучшены основные показатели оценки качества описаний на 13,5% за счет дополнительных временных затрат в 9,42%. [3, 7]
4. Предложен механизм совместного анализа описательных текстов, сгенерированных из различных модальностей, и текста диалога, обеспечивающий в результате применения оптимизации TensorRT и стратегии последовательной загрузки ускорение модели Gemma в 3,2 раза, сокращение использования памяти в 2,7 раза за счет снижения точности всего на 0,8% и показателя w-F1 на 1,2%. Данный механизм превзошел существующие решения в среднем на 8,15% за счет вычислительной сложности параллельного применения трех отдельных моделей. [5, 8, 9]
5. Разработано программное средство обнаружения эмоций в видео «ERC System», которое внедрено в ООО «Тутор Платформ» и успешно применяется для эмоционального анализа аудиовизуального контента, и благодаря внедрению микросервисной архитектуры и потоковой

обработки обеспечило модульность системы и возможность легкой замены отдельных компонентов без необходимости полного перепроектирования системы. Оценка эффективности программного средства в реальных условиях показала, что благодаря автоматическому распознаванию речи и автоматической идентификации говорящих оно применимо без наличия эталонных данных, за счет потери точности предложенного метода на 9,5%. [2]

Основные результаты диссертации опубликованы в следующих работах:

1. **Harutyunyan E.A.** Forming the requirements for emotion detection methods // Proceedings of the RA NAS and NPUA. Series of Technical Sciences. - 2022. - Vol. 75, - no. 4. - P. 508-518, doi: 10.53297/0002306X-2022.v75.4-508
2. **Khachatryan T., Galstyan D., Harutyunyan E.** A Comprehensive Approach for Enhancing Deep Learning Datasets Quality Using Combined SSIM Algorithm and FSRCNN // 2023 IEEE East-West Design & Test Symposium (EWDTS-2023). - 2023. - P. 1-4, doi: 10.1109/EWDTS59469.2023.10297040
3. **Nikoghosyan K.H., Harutyunyan E.A., Galstyan D.M.** Improving the image-to-speech system accuracy through integration of optical character recognition and language processing techniques // Proceedings of NPUA: Information technologies, Electronics, Radio engineering. - 2023. - no. 1. - P. 44-50, doi: 10.53297/18293336-2023.1-44
4. **Galstyan D.M., Harutyunyan E.A., Nikoghosyan K.H.** Human action recognition: Improving the accuracy of deep conv-lstm architecture through noise cleaning prior to key frames selection // Proceedings of the RA NAS and NPUA. Series of Technical Science - 2023. - Vol. 76, - no. 2. - P. 202-209, doi: 10.53297/0002306X-2023.v76.2-202
5. **Nikoghosyan K.H., Khachatryan T.B., Harutyunyan E.A., Galstyan D.M.** Acceleration of transformer architectures on Jetson Xavier using TensorRT // Proceedings of NPUA: Information Technologies, Electronics, Radio Engineering. - 2023. - no. 2. - P. 30-40, doi: 10.53297/18293336-2023.2-30
6. **Nikoghosyan K.H., Khachatryan T.B., Harutyunyan E.A., Galstyan D.M.** A Comprehensive System for Detecting Deepfake Videos and AI-Generated Text // Proceedings of NPUA: Information Technologies, Electronics, Radio Engineering. - 2024. - no. 1. - P. 37-44, doi: 10.53297/18293336-2024.1-37
7. **Nikoghosyan K.H., Khachatryan T.B., Harutyunyan E.A., Galstyan D.M.** Evaluating Open-Source Image Captioning Models with Multiple Metrics on the IAPR TC-12 Dataset // Bulletin of NPUA. Collection of Scientific Papers. - 2024. - no. 1. - P. 164-172
8. **Melikyan V., Khachatryan T., Galstyan D., Harutyunyan E.** Efficient Vision Transformer Deployment on Google Coral Through Knowledge Distillation // 2024 IEEE East-West Design & Test Symposium (EWDTS-2024). - 2024. - P. 1-3, doi: 10.1109/EWDTS63723.2024.10873747
9. **Harutyunyan E.A.** Modality-specific domain adaptation for multimodal emotion detection // Proceedings of NPUA: Information technologies, Electronics, Radio engineering. - 2024. - no. 2. - P. 57-64, doi: 10.53297/18293336-2024.2-57

ԱՄՓՈՓԱԳԻՐ

Ժամանակակից թվային դարաշրջանում տեսանյութերը դարձել են հաղորդակցության և տեղեկատվության փոխանցման հիմնական միջոցներից մեկը: Տեսահոլովակների քանակի անընդհատ աճը սոցիալական մեդիայում, կրթական հարթակներում և անվտանգային ոլորտում առաջացրել է դրանց ավտոմատ մշակման և վերլուծության անհրաժեշտություն: Հատկապես կարևոր է զգացմունքների հայտնաբերման խնդիրը, որի լուծումը կիրառման լայն հնարավորություններ կընձեռի՝ սկսած սպառողների վարքի վերլուծությունից մինչև հոգեկան առողջության մոնիթորինգը և այլն:

Զգացմունքների հայտնաբերման գոյություն ունեցող մեթոդները բաժանվում են 3 հիմնական խմբերի՝ հիմնված տեսողական, ձայնային կամ տեքստային բնութագրիչների վերլուծության վրա: Տեսողական մեթոդները հիմնականում կենտրոնանում են դեմքի արտահայտությունների վրա՝ հաճախ անտեսելով մարմնի լեզուն: Ձայնային մեթոդները, չնայած իրենց արդյունավետությանը, խիստ զգայուն են արտաքին աղմուկների նկատմամբ և ունեն սահմանափակումներ խառը ձայնային միջավայրերում: Տեքստային վերլուծության մեթոդները հաճախ անարդյունավետ են համատեքստային նրբություններ և բազմակի իմաստային շերտեր պարունակող արտահայտությունների մշակման դեպքում, բայց սահմանափակվում են բազմիմաստ և երկիմաստ արտահայտությունների հայտնաբերման դժվարություններով:

Վերջին տարիներին առաջարկված բազմամոդալ մեթոդները, որոնք փորձում են համակցել տարբեր բնութագրիչներ, բախվում են մի շարք խնդիրների: Մեծ մոդելները պահանջում են հսկայական հաշվողական ռեսուրսներ, ինչը դժվարացնում է դրանց կիրառումը իրական ժամանակում: Ի հակադրություն դրանց՝ ավելի թեթև մոդելները զիջում են ճշգրտությամբ և հաճախ չեն կարողանում բավարար ձևով ընկալել մոդալությունների միջև փոխկապակցությունները:

Արհեստական բանականության վերջին զարգացումները, հատկապես մեծ լեզվական մոդելների ոլորտում, բացում են նոր հնարավորություններ զգացմունքների հայտնաբերման համար: Այս մոդելները կարող են միավորել տարբեր մոդալություններից ստացված տեղեկատվությունը՝ ավելի համապարփակ և ճշգրիտ վերլուծություն ապահովելով: Այնուամենայնիվ, այս մոտեցումները դեռևս բախվում են մուտքային տվյալների ֆորմատավորման և վերլուծության արդյունավետության հետ կապված խնդիրներին:

Ատենախոսությունը նվիրված է տեսանյութերից զգացմունքների հայտնաբերման ավտոմատացմանն առնչվող հիմնահարցերի ուսումնասիրությանը, այդ ոլորտում առկա մարտահրավերների լուծմանը և

ժամանակակից տեխնոլոգիաների հիման վրա նոր, առավել արդյունավետ ալգորիթմների ու մեթոդների մշակմանը:

Առաջարկվել են տեսանյութից զգացմունքների հայտնաբերման ավտոմատացման միջոցների մշակման մոտեցումներ, որոնք, տեսողական և ձայնային բնութագրիչներից նկարագրական տեքստերի գեներացման, այդ տեքստերի և երկխոսության տեքստի համատեղ վերլուծության, ինչպես նաև արդյունավետ լեզվական մոդելի կիրառման շնորհիվ, արդյունքների ճշտության և հաշվողական ռեսուրսների ծախսի տեսանկյունից բավարարում են ժամանակակից պահանջները:

Մշակվել է տեսանյութերի տեսողական բնութագրիչներից նկարագրական տեքստի գեներացման մեթոդ, որն աղմուկների մաքրման և առանցքային կադրերի ընտրության նախնական մշակման քայլերի շնորհիվ՝ ապահովել է դիտարկման ենթակա կադրերի քանակի 96,6% կրճատում, ինչպես նաև այդ գործընթացի երկու անգամ արագացում՝ որոշ մանրամասների (BLEU-4՝ 0,46 և METEOR՝ 0,38) կորստի հաշվին, պահպանելով նկարագրությունների ընդհանուր իմաստային (CIDEr-D՝ 0,89 և Wu-P՝ 0,81) ճշտությունը:

Ստեղծվել է տեսանյութերի ձայնային բնութագրիչներից նկարագրական տեքստի գեներացման մեթոդ, որը, ձայնային աղմուկների մաքրման և որակի բարձրացման նախնական մշակման քայլերի կիրառման շնորհիվ միջինը 13,5%-ով բարելավել է նկարագրությունների որակի հիմնական գնահատման ցուցանիշները՝ 9,42% լրացուցիչ ժամանակային ծախսի հաշվին:

Առաջարկվել է տարբեր մոդալություններից գեներացված նկարագրական տեքստերի և երկխոսության տեքստի համատեղ վերլուծության մեխանիզմ, որը, TensorRT օպտիմալացման և հաջորդական բեռնման ռազմավարության կիրառման շնորհիվ, ապահովել է Gemma մոդելի 3,2 անգամ արագացում, հիշողության օգտագործման 2,7 անգամ կրճատում՝ ընդամենը 0,8% ճշտության և 1,2% w-F1 ցուցանիշի նվազման հաշվին: Այն երեք առանձին մոդելների զուգահեռ կիրառման հաշվողական բարդության հաշվին, միջինը 8,15%-ով գերազանցել է առկա լուծումները:

Մշակվել է տեսանյութում զգացմունքների հայտնաբերման «ERC System» ծրագրային միջոցը, որը ներդրված է «Տուտոր Պլատֆորմ» ՍՊԸ-ում, հաջողությամբ կիրառվում է տեսաձայնային բովանդակության զգացմունքային վերլուծության համար և ՄՏԿ ճարտարապետության ու հոսքային մշակման ներդրման շնորհիվ՝ ապահովել է համակարգի մոդուլայնությունը և առանձին բաղադրիչների հեշտ փոխարինման հնարավորությունը՝ առանց ամբողջական համակարգի վերանախագծման անհրաժեշտության: Ծրագրային միջոցի՝ իրական պայմաններում արդյունավետության գնահատումը ցույց է տվել, որ, խոսքի ավտոմատ ճանաչման և խոսակիցների ավտոմատ տարբերակման շնորհիվ, այն կիրառելի է՝ առանց էտալոնային տվյալների առկայության, առաջարկված մեթոդի ճշտության 9,5% կորստի հաշվին:

EDUARD ANDRANIK HARUTYUNYAN

DEVELOPMENT OF AUTOMATION TOOLS FOR EMOTION DETECTION IN VIDEO

SUMMARY

In the modern digital era, videos have become one of the main means of communication and information transfer. The continuous growth in the number of videos on social media, educational platforms, and security sectors has created a need for their automatic processing and analysis. Particularly important is the problem of emotion detection, which has wide application possibilities, ranging from consumer behavior analysis to mental health monitoring and more.

The existing emotion detection methods are divided into 3 main groups, based on visual, audio, or textual characteristics analysis. Visual methods mainly focus on facial expressions, often ignoring body language. Audio methods, despite their effectiveness, are highly sensitive to external noise and have limitations in mixed audio environments. Text analysis methods are often ineffective when processing expressions containing contextual nuances and multiple semantic layers but are limited by difficulties in detecting ambiguous and double-meaning expressions.

Multimodal methods proposed in recent years, which try to combine different characteristics, face a number of problems. Large models require enormous computational resources, which makes their application difficult in real-time. In contrast, lighter models compromise on accuracy and often cannot adequately perceive interconnections between modalities.

Recent developments in artificial intelligence, especially in the field of large language models, open new possibilities for emotion detection. These models can integrate information received from different modalities, providing more comprehensive and accurate analysis. However, these approaches still encounter problems related to input data formatting and analysis efficiency.

The dissertation is dedicated to the analysis of issues related to the automation of emotion detection from videos, solving challenges in this field, and developing new, more effective algorithms and methods based on modern technologies.

Approaches have been proposed for developing automation tools for emotion detection from video, which, through generating descriptive texts from visual and audio characteristics, joint analysis of these texts and dialogue text, as well as the application of an effective language model, satisfy modern requirements in terms of result accuracy and computational resource costs.

A method has been developed for generating descriptive text from visual characteristics of videos, which, thanks to preliminary processing steps of noise cleaning and key frame selection, has provided a 96,6% reduction in the number of frames to be examined, as well as a 2-fold acceleration of this process at the cost of losing some details (BLEU-4: 0,46 and METEOR: 0,38), while maintaining the overall semantic accuracy of descriptions (CIDEr-D: 0,89 and Wu-P: 0,81).

A method has been created for generating descriptive text from audio characteristics of videos, which, through the application of preliminary processing steps for cleaning

audio noise and improving quality, has improved the main quality assessment indicators of descriptions by an average of 13,5% at the cost of an additional 9,42% time expenditure.

A mechanism has been proposed for joint analysis of descriptive texts generated from different modalities and dialogue text, which, through the application of TensorRT optimization and sequential loading strategy, has provided a 3,2-fold acceleration of the Gemma model, a 2,7-fold reduction in memory usage at the cost of only a 0,8% decrease in accuracy and a 1,2% decrease in the w-F1 indicator. It has surpassed existing solutions by an average of 8,15% at the cost of computational complexity of parallel application of 3 separate models.

The "ERC System" software tool for detecting emotions in video has been developed, which has been implemented in "Tutor Platform" LLC and is successfully applied for emotional analysis of audiovisual content, and thanks to the implementation of MSC architecture and stream processing, has ensured the modularity of the system and the possibility of easy replacement of individual components without the need for complete system redesign. Evaluation of the software tool's effectiveness in real conditions has shown that, thanks to automatic speech recognition and automatic speaker identification, it is applicable without the availability of reference data, at the cost of a 9,5% loss in accuracy of the proposed method.