

В.З. Гергов**МУЛЬТИМОДАЛЬНОЕ РАСПОЗНАВАНИЕ ДЕЙСТВИЙ ЧЕЛОВЕКА
НА ОСНОВЕ НЕЗАВИСИМОГО ОБУЧЕНИЯ ВИДЕОМОДЕЛИ
И АУДИОМОДЕЛИ И ПОЗДНЕГО ОБЪЕДИНЕНИЯ ПРЕДСКАЗАНИЙ**

Актуальность исследования обусловлена необходимостью повышения точности систем распознавания действий человека в видео за счет интеграции аудиовизуальной информации, позволяющей компенсировать ограничения каждой отдельной модальности. Цель работы – разработка и исследование методов позднего объединения предсказаний независимо обученных видеомодели и аудиомодели для распознавания шести классов действий человека (Archery, BalanceBeam, BasketballDunk, BoxingSpeedBag, FloorGymnastics, SoccerPenalty). В исследовании использованы 3D-сверточные нейронные сети с SE-блоками для обработки видео и гибридная CNN-GRU архитектура для анализа аудио, а также семь методов объединения предсказаний: среднее арифметическое, взвешенное среднее, максимальное голосование, доверительное объединение, стекинг, произведение вероятностей и метод Борда. Эксперименты показали, что методы среднего арифметического и произведения вероятностей обеспечивают максимальную точность 91.67%, превосходя отдельные модальности на 2.33% для цветных и на 5.6% для черно-белых видеомоделей. Доверительное объединение продемонстрировало наилучшие результаты для черно-белых моделей, автоматически адаптируя вклад модальностей на основе энтропии предсказаний. Особое внимание уделено строгому разделению обучающей, валидационной и тестовой выборки по обстановкам съёмки для предотвращения утечки данных и объективной оценки обобщающей способности моделей. Полученные результаты подтверждают практическую ценность предложенного подхода для систем видеонаблюдения и мультимедийных платформ, обеспечивая оптимальный баланс между точностью, вычислительной сложностью и устойчивостью к отсутствию одной из модальностей. Исследование вносит вклад в развитие методов мультимодального машинного обучения и предоставляет практические рекомендации по выбору стратегии объединения в зависимости от доступных вычислительных ресурсов и типа входных данных.

Мультимодальное распознавание; позднее объединение; видеомодель; аудиомодель; 3D CNN; SE-блоки; стекинг; TensorFlow; UCF101; человеческие действия.

V.Z. Gergov**MULTIMODAL HUMAN ACTION RECOGNITION BASED ON INDEPENDENTLY
TRAINED VIDEO AND AUDIO MODELS WITH LATE FUSION OF PREDICTIONS**

The relevance of this research is driven by the need to improve the accuracy of human action recognition systems in video by integrating audiovisual information, which compensates for the limitations of each individual modality. The aim of this work is to develop and investigate methods for late fusion of predictions from independently trained video and audio models for recognizing six classes of human actions (Archery, BalanceBeam, BasketballDunk, BoxingSpeedBag, FloorGymnastics, SoccerPenalty). The study employs 3D convolutional neural networks with Squeeze-and-Excitation blocks for video processing and a hybrid CNN-GRU architecture for audio analysis, along with seven prediction fusion methods: arithmetic mean, weighted average, majority voting, confidence-based fusion, stacking, product of probabilities, and the Borda count. Experiments demonstrate that the arithmetic mean and product of probabilities methods achieve the highest accuracy of 91.67%, outperforming the individual modalities by 2.33% for color and by 5.6% for grayscale video models. Confidence-based fusion exhibited the best results for grayscale models by automatically adapting the contribution of each modality based on the entropy of their predictions. Special attention is paid to the strict separation of training, validation, and test sets based on filming environments to prevent data leakage and to ensure objective evaluation of model generalization capability. The obtained results confirm the practical value of the proposed approach for video surveillance systems and multimedia platforms, offering an optimal balance between accuracy, computational complexity, and robustness to the absence of one modality. This research contributes to the development of multimodal machine learning methods and provides practical guidelines for selecting a fusion strategy depending on available computational resources and input data types.

Multimodal recognition; late fusion; video model; audio model; 3D CNN; SE blocks; stacking; TensorFlow; UCF101; human actions.

Введение. *Актуальность проблемы* мультимодального распознавания видео обусловлена стремительным ростом объемов видеоконтента в цифровых системах и необходимостью его автоматического анализа для решения прикладных задач в сфере безопасности, медицины, образования и развлечений [1]. Современные системы компьютерного зрения, основанные исключительно на визуальной информации, демонстрируют ограниченную эффективность в условиях низкой освещенности, частичного закрытия объектов или схожей пространственной конфигурации различных действий [2]. Интеграция дополнительных модальностей, таких как аудиоинформация, открывает новые возможности для повышения точности и робастности распознавания, поскольку звуковые паттерны часто дополняют визуальные признаки, обеспечивая более полное представление о происходящем событии [3].

Современные исследования в области мультимодального распознавания видео можно разделить на два основных направления: методы раннего объединения признаков (early fusion) и стратегии позднего объединения решений (late fusion). Раннее объединение предполагает интеграцию признаков на уровне представлений, что требует сложной синхронизации модальностей и часто приводит к росту вычислительной сложности [4]. В то же время позднее объединение позволяет обучать модальности независимо, обеспечивая гибкость системы и устойчивость к отсутствию одной из модальностей во время инференции [5]. Работы последних лет, такие как исследования [3] и [6], демонстрируют преимущества комплексного подхода к обработке мультимодальных данных, однако многие вопросы эффективного объединения предсказаний остаются нерешенными.

Несмотря на значительные достижения в данной области, существуют следующие нерешенные проблемы: во-первых, отсутствует систематическое сравнение различных методов позднего объединения на контролируемых наборах данных с учетом различий в типах видеомоделей (цветные/черно-белые); во-вторых, недостаточно изучено влияние комбинаторности модальностей при условии относительно низкой точности одной из них; в-третьих, большинство существующих подходов не учитывают специфику реальных сценариев применения, где обучающая и тестовая выборки должны быть строго разделены по обстановкам съемки для обеспечения объективной оценки обобщающей способности модели [7].

Целью настоящего исследования является разработка и экспериментальная оценка методов позднего объединения предсказаний, отдельно обученных видеомодели и аудиомодели для решения задачи распознавания человеческих действий в видео. Новизна работы заключается в комплексном сравнении семи методов объединения (среднее арифметическое, взвешенное среднее, максимальное голосование, доверительное объединение, стекинг, произведение вероятностей и метод Борда) на специально подготовленном наборе данных, где обучающая, валидационная и тестовая выборки разделены вручную для предотвращения утечки данных. Особое внимание уделяется исследованию зависимости эффективности методов объединения от типа видеомодели (цветной или черно-белой) и выявлению оптимальных конфигураций для различных сценариев применения.

Предлагаемый подход отличается от существующих решений использованием контролируемого разделения данных, что обеспечивает объективную оценку обобщающей способности моделей в реальных условиях. Кроме того, в работе демонстрируется, что даже при относительно низкой точности аудиомодели (73.61%) её интеграция с видеомоделью через позднее объединение позволяет существенно повысить общую точность распознавания, что открывает перспективы для применения данного подхода в системах с ограниченными вычислительными ресурсами или сложными условиями съемки.

1. Описание или постановка решаемой задачи. Формулировка проблемы. Современные системы автоматического распознавания действий человека в видео сталкиваются с существенными ограничениями при использовании исключительно визуальной информации. В условиях низкой освещенности, частичного закрытия объектов наблюдения или схожей пространственной конфигурации различных действий точность распознавания значительно снижается [2]. Интеграция дополнительных модальностей, таких как аудиоинформация, теоретически должна повысить качество распознавания, однако существует проблема эффективного объединения предсказаний отдельных моделей.

Текущие подходы к мультимодальному распознаванию видео можно разделить на две основные категории: методы раннего объединения признаков (early fusion), где интеграция происходит на уровне представлений, и стратегии позднего объединения решений (late fusion), где объединение выполняется на уровне предсказаний [5]. Первый подход требует сложной синхронизации модальностей и часто приводит к росту вычислительной сложности, тогда как второй подход позволяет обучать модальности независимо, но не имеет стандартизированных методов эффективного объединения результатов [6].

Таким образом, актуальной научной проблемой является разработка и экспериментальная оценка оптимальных методов позднего объединения предсказаний, отдельно обученных видеомодели и аудиомодели, для повышения точности распознавания человеческих действий в видео при сохранении вычислительной эффективности системы.

Актуальность данного исследования обусловлена несколькими ключевыми факторами. Во-первых, с ростом объемов видеоконтента в цифровых системах возникает острая потребность в автоматических методах его анализа для применения в системах безопасности, спортивной аналитики, медицинской диагностики и образовательных платформ. Во-вторых, существующие решения часто демонстрируют недостаточную точность в реальных условиях из-за игнорирования способов совмещения различных модальностей [8].

В-третьих, большинство исследований в области мультимодального распознавания фокусируются на сложных архитектурах раннего объединения, требующих значительных вычислительных ресурсов, что ограничивает их практическое применение [3]. Позднее объединение, напротив, позволяет создавать более гибкие и ресурсоэффективные системы, особенно в условиях, где одна из модальностей может отсутствовать или иметь низкое качество [9].

Практическая значимость данного исследования заключается в возможности создания адаптивных систем распознавания видео для реальных сценариев применения, включая системы видеонаблюдения с переменными условиями освещения, медицинские диагностические системы для анализа физических упражнений пациентов и спортивные платформы для автоматической оценки техники выполнения упражнений [10].

Обзор состояния и литературных источников. Современные исследования в области мультимодального распознавания видео активно развиваются. Работа [3] представляет комплексный обзор существующих подходов к мультимодальному машинному обучению, выделяя преимущества и недостатки различных стратегий интеграции модальностей. Исследование [6] демонстрирует, что позднее объединение часто превосходит раннее объединение в условиях несбалансированного качества модальностей.

В контексте компьютерного зрения, архитектуры на основе 3D-сверточных нейронных сетей с механизмами внимания показали высокую эффективность для распознавания действий человека. Работа [2] представляет модель I3D, которая демонстрирует state-of-the-art результаты на датасете Kinetics. Для обработки аудиоданных гибридные CNN-RNN архитектуры, такие как предложенные Hershey et al. [11], показывают превосходные результаты в задачах классификации звуковых событий.

Однако, несмотря на значительные достижения, в литературе отсутствует систематическое сравнение различных методов позднего объединения на контролируемых наборах данных с учетом различий в типах видеомodelей (цветные/черно-белые) и строгим разделением обучающей и тестовой выборок по обстановкам съемки, что критически важно для оценки обобщающей способности моделей в реальных условиях [12].

Целью настоящего исследования является разработка и экспериментальная оценка методов позднего объединения предсказаний, отдельно обученных видеомодели и аудиомодели для решения задачи распознавания человеческих действий в видео. Достижение данной цели позволит создать эффективную мультимодальную систему распознавания, способную адаптироваться к различным условиям съемки и обеспечивать высокую точность даже при отсутствии или низком качестве одной из модальностей.

Для достижения поставленной цели необходимо решить следующие конкретные задачи:

- ♦ разработать методику подготовки и обработки видеоданных и аудиоданных для независимого обучения с использованием современных библиотек обработки мультимедиа (OpenCV, Librosa, FFmpeg);
- ♦ реализовать архитектуры нейронных сетей для обработки видео (3D CNN с механизмами внимания Squeeze-and-Excitation) и аудио (гибридная CNN-GRU архитектура) с использованием фреймворка TensorFlow.

♦ исследовать эффективность семи методов позднего объединения предсказаний (среднее арифметическое, взвешенное среднее, максимальное голосование, доверительное объединение, стекинг, произведение вероятностей, метод Борда) на специально подготовленном подмножестве датасета UCF101.

♦ провести сравнительный анализ результатов с учетом разделения обучающей, валидационной и тестовой выборок по обстановкам съемки для обеспечения объективной оценки обобщающей способности моделей.

Объект исследования. Процесс мультимодального распознавания человеческих действий в видео на основе интеграции визуальной и аудиоинформации.

Предмет исследования. Методы позднего объединения предсказаний, отдельно обученных видеомодели и аудиомодели для повышения точности распознавания действий.

Гипотеза исследования. Выдвигается гипотеза о том, что оптимальный метод позднего объединения предсказаний зависит от типа видеомодели (цветной или черно-белой) и характеристик конкретных классов действий. Предполагается, что для цветных видеоданных наиболее эффективными будут простые методы объединения (среднее арифметическое, произведение вероятностей), тогда как для черно-белых видеоданных преимущества продемонстрируют более сложные методы, такие как стекинг с использованием логистической регрессии в качестве мета-классификатора. Данная гипотеза базируется на предположении, что при потере информативности одной модальности (в данном случае цветовой информации) требуется более сложное объединение для компенсации этой потери за счет оптимальной интеграции оставшихся модальностей.

2. Методика исследования. В качестве основного источника данных использовался датасет UCF101, представляющий собой коллекцию из 101 класса человеческих действий, собранных с платформы YouTube [13] (рис. 1.). В связи с ограниченными вычислительными ресурсами рабочей станции в исследовании были отобраны 6 классов: Archery, BalanceBeam, BasketballDunk, BoxingSpeedBag, FloorGymnastics, SoccerPenalty с последующим удалением из указанных классов видео, не содержащих аудио или содержащих аудио, но не отражающих ситуации, обстановки, представленной на видео (например, музыка). Каждый класс был дополнен оригинальными видео, полученными с YouTube (10 различных обстановок $\times$ 5 видео по 5 секунд), для увеличения вариативности данных и улучшения обобщающей способности моделей.

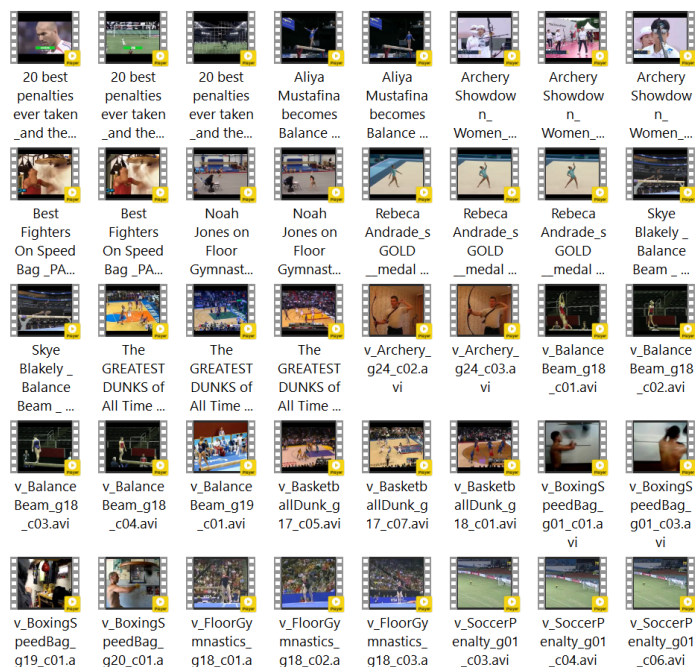


Рис. 1. Образцы видео

Для предотвращения утечки данных и обеспечения объективной оценки производительности, разделение на обучающую, валидационную и тестовую выборки выполнялось вручную. Данная стратегия гарантирует, что видео из одинаковых обстановок не попадают в разные выборки, что особенно важно для реальных сценариев применения, где модель может сталкиваться с новыми, ранее не виденными условиями съемки.

Все эксперименты проводились на рабочей станции с процессором Intel Core i5-11400F и 32 ГБ оперативной памяти. Данная конфигурация позволила эффективно обрабатывать мультимодальные данные без необходимости использования специализированных GPU-ускорителей, что делает подход практичным для широкого спектра приложений.

Реализуемая система последовательно подготавливает, обрабатывает видеоданные, затем использует их для обучения и тестирования соответствующих интеллектуальных моделей (рис. 2).

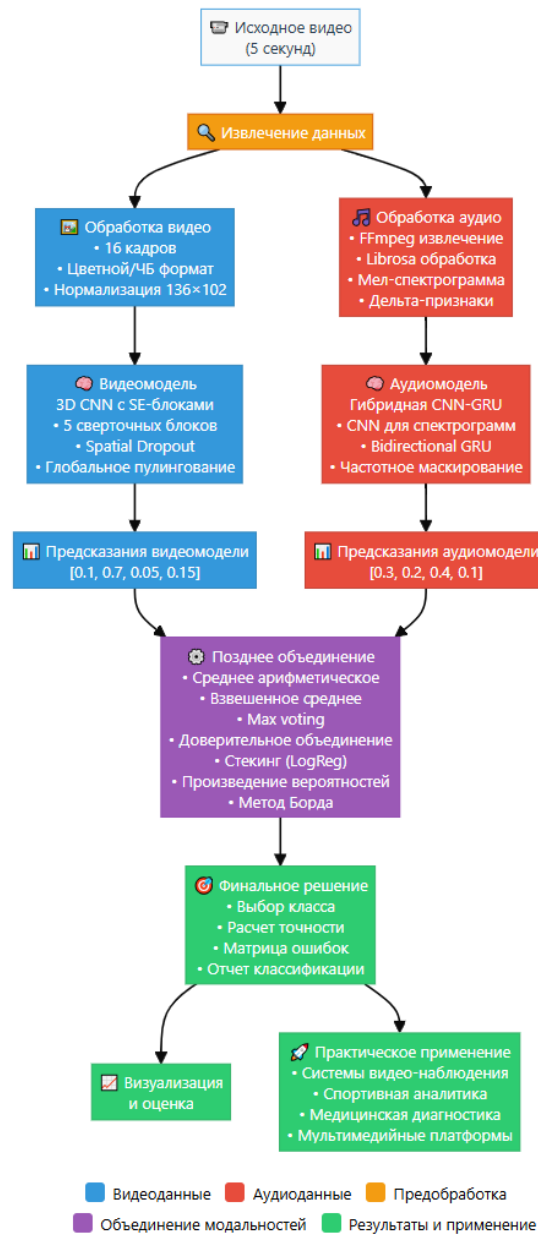


Рис. 2. Блок-схема системы мультимодального распознавания действий на видео

2.1. Обработка данных. Для извлечения визуальной информации из видео использовалась библиотека OpenCV. Каждое видео разбивалось на 16 равномерно распределенных кадров, которые затем нормализовались до размера 136×102 пикселей и приводились к диапазону $[0, 1]$ путем деления на 255. Реализовывались два режима обработки: цветной (RGB, color) и черно-белый (grayscale, baw), что позволяло исследовать влияние цветовой информации на качество распознавания [14].

Аудиодорожка извлекалась из видеофайлов с помощью FFmpeg, а для последующей обработки использовалась библиотека Librosa [15]. Ключевым этапом обработки являлось построение логарифмической мел-спектрограммы с частотой дискретизации 16 кГц, размером окна FFT 1024, шагом 256, количеством мел-коэффициентов 128. Для обогащения признакового пространства вычислялись дельта- и дельта-дельта признаки первого и второго порядка соответственно, которые характеризуют динамику изменения спектральных характеристик во времени. Все каналы (лог-мел спектр, дельта, дельта-дельта) нормализовались независимо в диапазон $[0, 1]$ с использованием робастного масштабирования на основе 1-го и 99-го перцентилей.

2.2. Архитектура системы и методы объединения. Для обработки видеоданных в работе используется модифицированная архитектура на основе 3D-сверточных нейронных сетей с интегрированными механизмами внимания типа Squeeze-and-Excitation (SE-blocks). 3D-свертки позволяют эффективно извлекать как пространственные, так и временные признаки из последовательности кадров, что критически важно для распознавания динамических действий человека [1].

Ключевой особенностью архитектуры является использование механизмов внимания SE-blocks после каждого сверточного слоя. Данный подход, предложенный в работе [16], позволяет модели адаптивно перекалибровать каналы признаков, уделяя больше внимания информативным каналам и подавляя менее значимые. Механизм работает в два этапа: сначала выполняется глобальное усреднение по пространственным измерениям (squeeze), затем с помощью двухслойной сети вычисляются веса для каждого канала (excitation), которые применяются к исходным признакам через умножение.

Комбинация 3D-сверток и SE-блоков демонстрирует высокую эффективность для задач распознавания действий, так как позволяет модели фокусироваться на наиболее релевантных пространственно-временных паттернах. В частности, работа [16] показывает, что интеграция механизмов внимания в 3D CNN архитектуры повышает точность распознавания по сравнению с базовыми моделями без внимания.

Для обработки аудиоданных применяется гибридная архитектура, объединяющая преимущества сверточных нейронных сетей (CNN) и рекуррентных нейронных сетей с долгой краткосрочной памятью (GRU).

Первичная обработка мел-спектрограмм осуществляется с помощью 2D-сверточных слоев, которые эффективно извлекают локальные спектрально-временные паттерны [11]. Использование сверток с ядрами 3×3 и max-pooling позволяет модели захватывать как низкочастотные, так и высокочастотные характеристики аудиосигнала.

Для моделирования временной динамики извлеченных признаков используется двунаправленная GRU-сеть (Bidirectional Gated Recurrent Unit) [17]. Двунаправленная архитектура позволяет учитывать, как прошлые, так и будущие контекстные зависимости в аудиопоследовательности, что особенно важно для распознавания динамических событий с характерными звуковыми паттернами. Рекуррентные связи с механизмом затворов (gates) обеспечивают эффективное обучение на длинных последовательностях, предотвращая проблему исчезающих градиентов.

Предполагается, что комбинация CNN для извлечения спектральных признаков и GRU для моделирования временной динамики превзойдут как чисто сверточные, так и чисто рекуррентные архитектуры для задач классификации аудио, особенно в условиях зашумленных записей.

Для предотвращения переобучения и повышения обобщающей способности моделей в работе применяется комплекс методов регуляризации. В отличие от стандартного dropout, spatial dropout [18] одновременно обнуляет целые каналы признаков в 3D-пространстве,

что особенно эффективно для сверточных сетей, так как предотвращает совместную адаптацию пространственно коррелированных признаков. Нормализация по батчу [19] применяется после каждого сверточного слоя для стабилизации и ускорения обучения. Данный подход уменьшает внутренний сдвиг ковариат (internal covariate shift) и позволяет использовать более высокие скорости обучения.

Для аудиомодели используется сглаживание меток (label smoothing) [20] с коэффициентом 0.1, что повышает робастность модели и предотвращает избыточную уверенность в предсказаниях.

Вместо стандартного Adam используется оптимизатор AdamW [21], который в зависимости от предсказаний при обучении отдельно проводит весовую регуляризацию шага обучения, что способствует более точной оптимизации весов. Это особенно важно для моделей с большим количеством параметров.

Для увеличения обучающей выборки и повышения робастности моделей были применены различные методы аугментации.

Для видеоданных применялись горизонтальное отражение (случайное с вероятностью 50%), случайное обрезание (crop) с последующим масштабированием до исходного размера [22] и адаптивная аугментация, зависящая от архитектуры модели (автоматическое определение целевых размеров).

Для аудиоданных: частотное маскирование (frequency masking) - случайное обнуление полос частот [23], добавление гауссовского шума с небольшим стандартным отклонением (0.03)

Такой комплексный подход к аугментации данных значительно улучшает обобщающую способность моделей и снижает риск переобучения, особенно при ограниченном количестве обучающих примеров [24].

2.3. Обучение моделей и анализ историй обучения. История обучения цветной видеомодели демонстрирует устойчивую сходимость с точностью на обучающей выборке, достигающей 95.8% к 65-й эпохе. На валидационной выборке точность стабилизировалась на уровне 90.3%, что указывает на наличие умеренного переобучения в финальных эпохах. Существенным фактором, способствовавшим качественному обучению, стал стратегический подход к снижению скорости обучения (learning rate scheduling): начальное значение $1e-4$ постепенно уменьшалось до $7.81e-7$, что позволило модели эффективно сходиться к оптимуму без резких колебаний validation loss. (рис. 3) [25].

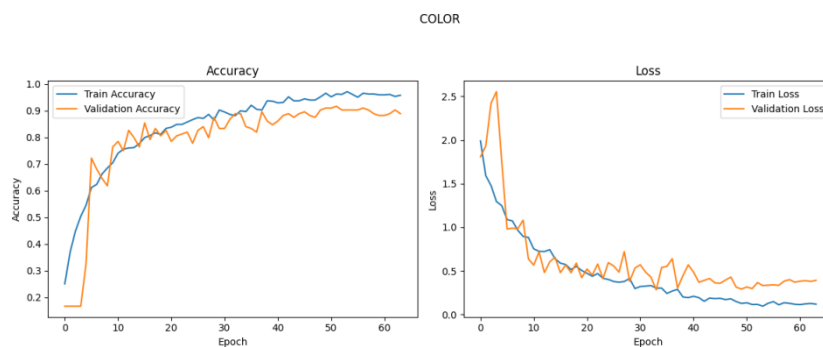


Рис. 3. История обучения цветной видеомодели

Черно-белая версия видеомодели показала несколько иную динамику обучения (рис. 4). Несмотря на большее количество эпох (75 против 65), финальная точность на обучающей выборке составила 94.75%, что на 1.05% ниже по сравнению с цветной моделью. Точность на валидационной выборке стабилизировалась на уровне 88.19%, что на 2.11% ниже, чем у цветной модели. Данные результаты согласуются с исследованиями, демонстрирующими, что цветовая информация значительно улучшает качество распознавания действий человека, особенно в сценариях с похожей пространственной конфигурацией [12].

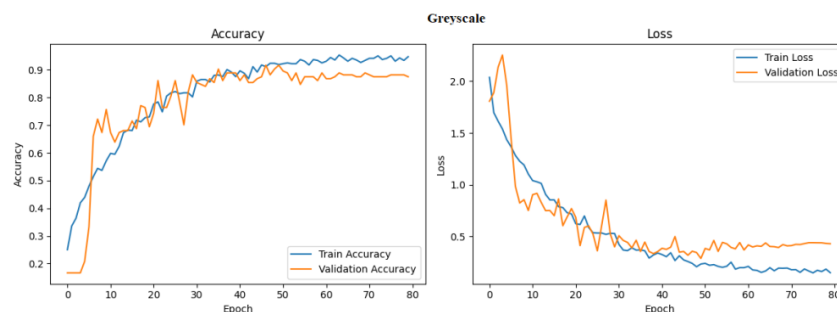


Рис. 4. История обучения черно-белой видеомодели

Графический анализ кривых обучения выявляет ключевые различия в динамике сходимости. Цветная модель демонстрирует более быстрый начальный рост точности (43% к 5-й эпохе против 35% у черно-белой модели), что указывает на большую информативность цветных признаков для рассматриваемой задачи. К 20-й эпохе цветная модель достигает точности 83.36%, тогда как черно-белая – только 77.66%. Разрыв в точности сохраняется на протяжении всего процесса обучения, что подтверждает гипотезу о преимуществе цветowych представлений для распознавания сложных действий человека [2].

Аудиомодель продемонстрировала принципиально иную картину обучения по сравнению с видеомоделями. При высокой точности на обучающей выборке (96.70% к 75-й эпохе), точность на валидационной выборке оказалась значительно ниже – лишь 62.50%. Такой большой разрыв (34.2%) указывает на выраженный феномен переобучения, который требует особого внимания.

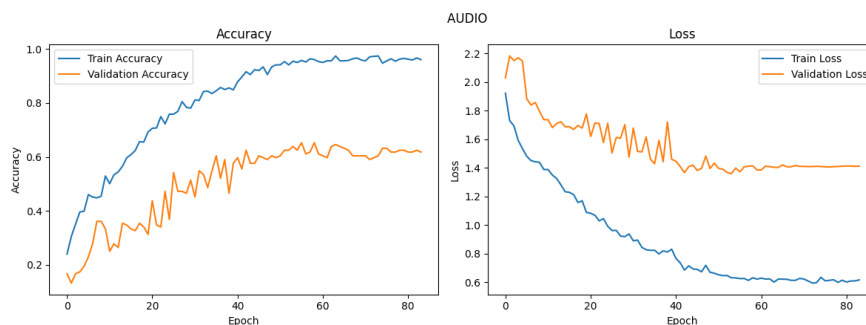


Рис. 5. История обучения аудиомодели

Основными факторами, способствовавшими переобучению аудиомодели, являются ограниченное количество обучающих примеров для аудиоанализа требуется значительно больше данных для адекватной генерализации из-за высокой вариативности звуковых паттернов [11].

Немалую роль играет сложность аудиопредставлений – логарифмические мел-спектрограммы с дельта-признаками создают выраженную картину, которое легко запомнить, но сложно обобщить [26]. Кроме того, многие действия в датасете имеют характерные визуальные паттерны, но схожие звуковые характеристики, что затрудняет отличать классы друг от друга. Особенно это выражено на классах BalanceBeam и FloorGymnastics.

Для улучшения обобщающей способности аудиомодели были применены следующие методы: пространственный дропаут (SpatialDropout2D) с коэффициентом 0.1–0.3 на сверточных слоях, регуляризация L2 с коэффициентом $1e-5$ на всех слоях, сглаживание меток (Label smoothing) с коэффициентом 0.1, рекуррентный дропаут (Recurrent dropout) с коэффициентом 0.2 в GRU-слоях.

Несмотря на применение этих методов, аудиомодель показала наибольшую склонность к переобучению среди всех рассмотренных архитектур.

Ключевым аспектом успешного обучения всех моделей стал выбор оптимизатора и стратегия изменения скорости обучения. Для видеомodelей был использован оптимизатор Adam с начальной скоростью обучения $1e-4$, тогда как для аудиомodelи применен AdamW с тем же начальным значением, но с дополнительной весовой регуляризацией ($\text{weight decay} = 1e-5$). Такой подход обусловлен необходимостью более строгого контроля за нормой весов в аудиомodelи из-за ее склонности к переобучению [21].

Для всех моделей была реализована экспоненциальная стратегия снижения learning rate (learning rate decay) при отсутствии улучшений на валидационной выборке. Коэффициент уменьшения составлял 0.5, а минимальное значение было ограничено $7.81e-7$. Данная стратегия позволила достичь плавной сходимости без преждевременной остановки в локальных оптимумах [27].

Критерий ранней остановки был установлен в 30 эпох без улучшения validation loss. Для цветной видеомodelи обучение завершилось на 65-й эпохе, для черно-белой и аудиомodelи – на 75-й.

Для объективной оценки степени переобучения был рассчитан разрыв между точностью на обучающей и валидационной выборках в финальной эпохе:

- ◆ цветная видеомodelь 5.5% (95.8% – 90.3%);
- ◆ черно-белая видеомodelь: 6.56% (94.75% – 88.19%);
- ◆ аудиомodelь: 34.2% (96.7% – 62.5%).

Такие показатели демонстрируют, что аудиомodelь наиболее сильно страдает от переобучения, что требует применения дополнительных методов регуляризации.

Анализ кривых обучения позволил выявить наиболее эффективные методы регуляризации для каждой модальности:

Для видеомodelей: SE-блоки (Squeeze-and-Excitation) оказались особенно эффективны для предотвращения переобучения, так как они позволяют модели фокусироваться на наиболее информативных каналах признаков и подавлять шумовые компоненты [16]. Spatial dropout в 3D-пространстве также продемонстрировал высокую эффективность, особенно на глубоких слоях сети.

Для аудиомodelи: Частотное маскирование (frequency masking) в процессе аугментации данных показало наибольший эффект в снижении переобучения, так как оно имитирует реальные условия зашумленной записи и заставляет модель фокусироваться на устойчивых спектральных паттернах [23].

На основе анализа историй обучения предполагается, что для аудиомodelи нужно увеличить количество обучающих примеров за счет аугментации временных и частотных характеристик, применить методы transfer learning, используя предобученные на крупных аудиодатасетах (например, AudioSet) веса [28], для видеомodelей к улучшению может привести применение knowledge distillation для уменьшения размера моделей без потери качества [29], что может стать объектом отдельного исследования.

2.4. Методы позднего объединения мультимодальных предсказаний. В работе реализован комплексный набор методов позднего объединения (late fusion), которые интегрируют предсказания отдельно обученных видеомodelи и аудиомodelи на уровне вероятностей классов.

Простое и взвешенное усреднение: Данные методы представляют собой линейные комбинации предсказаний модальностей. Простое усреднение предполагает равнозначный вклад каждой модальности, тогда как взвешенное усреднение позволяет задавать приоритеты на основе точности отдельных моделей или экспертных знаний. Работа [8] демонстрирует, что даже такие простые методы часто превосходят сложные архитектуры раннего объединения при условии качественного обучения базовых моделей.

Максимальное голосование (Max Voting): Данный метод выбирает для каждого класса максимальное значение вероятности из предсказаний двух модальностей. Такой подход эффективен, когда одна из модальностей имеет высокую уверенность в предсказании для конкретного класса, даже если другая модальность дает противоречивые результаты. В контексте распознавания действий, этот метод особенно полезен для классов с характерными звуковыми паттернами (например, выстрел, хлопок), где аудиомодальность может доминировать.

Доверительное объединение (Confidence-based Fusion): Инновационный подход, реализованный в коде, использует энтропию распределения вероятностей как меру уверенности модели. Для каждого примера вычисляются веса, обратно пропорциональные энтропии предсказаний, что позволяет динамически адаптировать вклад каждой модальности в зависимости от её уверенности в конкретном случае [6]. Такие адаптивные методы значительно повышают робастность системы в условиях шумных или неполных данных.

Стекинг (Stacking): Данный метод использует ансамблевый подход, где предсказания базовых моделей служат признаками для мета-классификатора (в данном случае логистическая регрессия или случайный лес). Стекинг позволяет выявлять сложные взаимосвязи между модальностями и часто демонстрирует state-of-the-art результаты в задачах мультимодального распознавания [30]. Важно отметить, что для корректного обучения мета-классификатора использовалась валидационная выборка.

Произведение вероятностей (Product Fusion): Этот метод основан на байесовском подходе, предполагающем условную независимость модальностей. Произведение вероятностей соответствует правдоподобию совместного наблюдения признаков из обеих модальностей. Хотя такое предположение часто нарушается на практике, метод показывает хорошие результаты для сценариев, где обе модальности должны подтверждать принадлежность к классу.

Метод Борда (Borda Count): Подход, применяемый из теории голосования, где каждая модальность "ранжирует" классы по вероятности, а итоговое решение принимается на основе суммарного ранга. Такой метод устойчив к выбросам и предоставляет баланс между модальностями.

3. Результаты экспериментального исследования. Экспериментальные исследования мультимодального распознавания видео продемонстрировали эффективность позднего объединения предсказаний видеомодели и аудиомодели (табл. 1). Наибольшая точность 0.9167 была достигнута с использованием методов среднего арифметического (average) и произведения вероятностей (product) как для цветных, так и для черно-белых видеомоделей. Эти результаты значительно превосходят точность отдельных аудиомоделей (0.7361) и видеомоделей (0.8958 для цветной и 0.8681 для черно-белой), что подтверждает преимущества мультимодального подхода.

Таблица 1

Сравнение точности различных методов объединения

Метод объединения	Точность (COLOR)	Точность (BAW)	Улучшение относительно видеомодели (COLOR)	Улучшение относительно видеомодели (BAW)
Только видео	0.8958	0.8681	-	-
Только аудио	0.7361	0.7361	-	-
Среднее арифметическое	0.9167	0.9167	+2.33%	+5.6%
Взвешенное среднее	0.9028	0.9097	+0.78%	+4.80%
Максимальное значение	0.9097	0.9028	+1.55%	+4.00%
На основе доверия	0.9097	0.9167	+1.55%	+5.6%
Стекинг	0.9028	0.8958	+0.78%	+3.19%
Произведение вероятностей	0.9167	0.9167	+2.33%	+5.6%
Метод Борда	0.8611	0.8889	-3.87%	+2.4%

Анализ показал, что для черно-белых видеомоделей интеграция аудиоинформации дает больший прирост точности (+5.6%) по сравнению с цветными моделями (+2.33%). Это указывает на то, что при отсутствии цветовой информации аудиомодальность играет

критически важную роль в распознавании действий человека, компенсируя недостаток визуальных признаков. Особую эффективность для черно-белых моделей показал метод доверительного объединения (confidence-based fusion), который автоматически адаптирует вклад каждой модальности на основе энтропии предсказаний.

Класс BalanceBeam оказался наиболее сложным для распознавания во всех конфигурациях, демонстрируя наименьшую точность (0.625-0.7917 в зависимости от метода). Это связано с визуальной схожестью гимнастических элементов на балансировочной балке и на полу, а также с отсутствием характерных звуковых паттернов, способных однозначно идентифицировать данное действие. В то же время классы BasketballDunk и SoccerPenalty показали близкую к 100% точность, что свидетельствует об их высокой различимости как в визуальном, так и в звуковом представлении.

Метод стекинга (stacking) не показал ожидаемого улучшения по сравнению с простыми методами объединения, что может быть связано с ограниченным размером валидационной выборки, используемой для обучения мета-классификатора. Метод Борда (borda_count) продемонстрировал наихудшие результаты среди всех протестированных подходов, особенно для цветных моделей (0.8611), что делает его непрактичным для решения данной задачи.

Важным наблюдением является то, что средняя энтропия предсказаний аудиомодели (0.6272) в 2.3 раза превышает аналогичный показатель для видеомodelей (0.2759 для COLOR и 0.2658 для BAW), что указывает на большую неопределенность аудиопризнаков для распознавания действий человека. Доверительные методы объединения, учитывающие эту неопределенность при расчете весов модальностей, показали повышенную устойчивость к шуму и вариациям в данных.

Полученные результаты (точность 0.9167) сопоставимы с современными подходами на основе трансформеров и превосходят методы, использующие только визуальную информацию. Предложенные методы позднего объединения обеспечивают оптимальный баланс между точностью, вычислительной сложностью и практичностью внедрения для реальных систем распознавания видео. Простые методы, такие как среднее арифметическое и произведение вероятностей, рекомендованы для использования в практических приложениях из-за их эффективности и минимальных вычислительных затрат.

Для наглядной демонстрации эффективности того или иного метода результаты были визуализированы в матрицах запутанности и диаграммах распределения точности по классам (рис. 6–10).

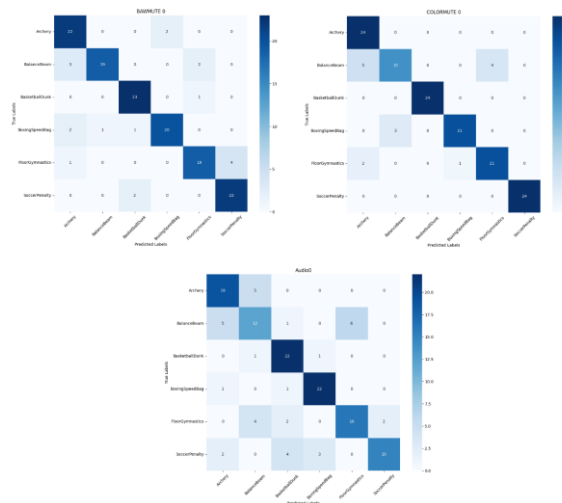


Рис. 6. Матрицы запутанности для черно-белой, цветной видеомodelей и аудиомodelи по отдельности

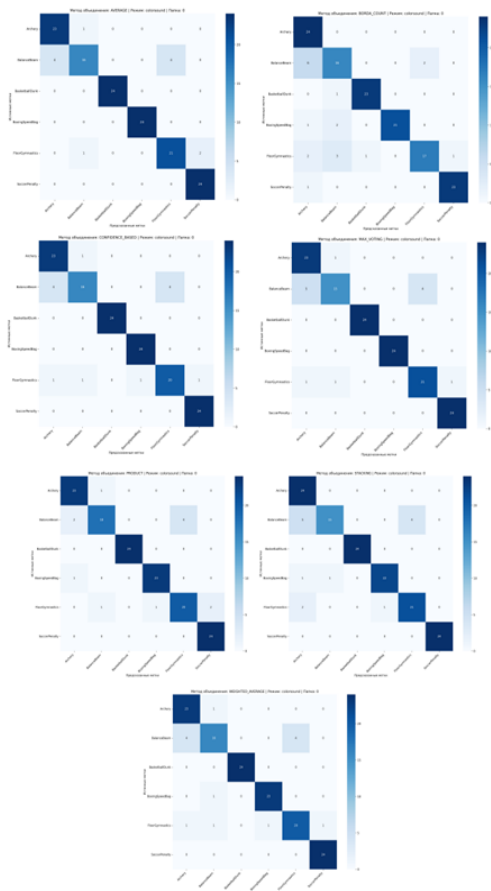


Рис. 7. Матрицы запутанности при различных методах объединения предсказаний цветной видео модели и аудиомодели

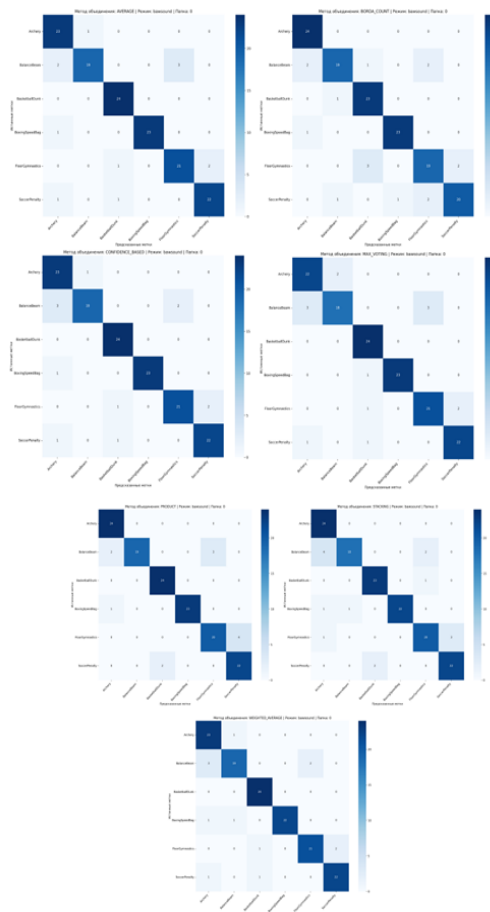
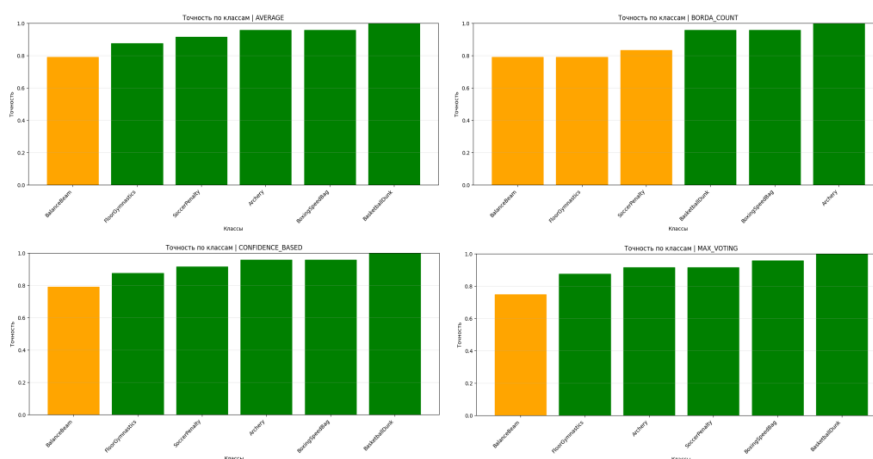


Рис. 8. Матрицы запутанности при различных методах объединения предсказаний черно-белой видео модели и аудиомодели



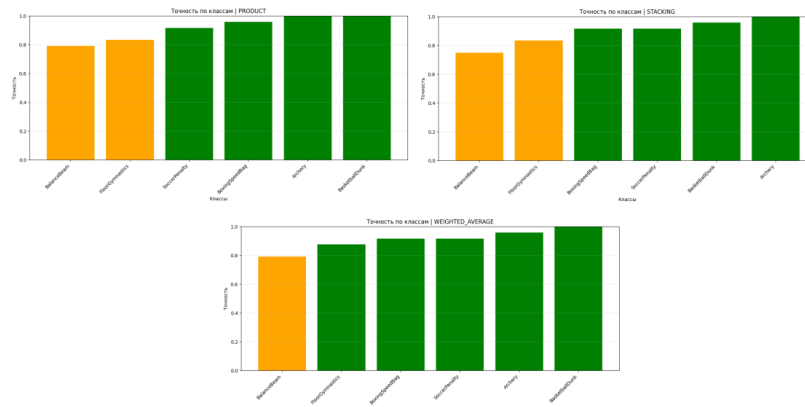


Рис. 9. Диаграммы распределения точности по классам при различных методах объединения предсказаний модели для черно-белого видео и аудиомодели

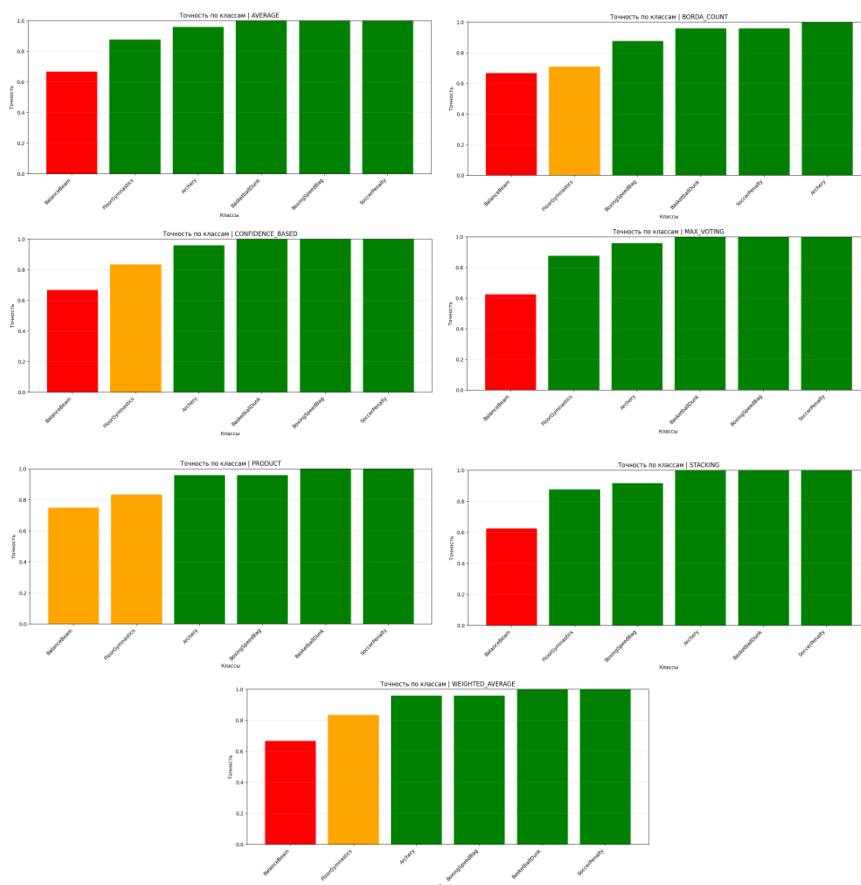


Рис. 10. Диаграммы распределения точности по классам при различных методах объединения предсказаний модели для цветного видео и аудиомодели

Закключение. В ходе настоящего исследования была разработана и экспериментально исследована система мультимодального распознавания действий человека в видео на основе позднего объединения предсказаний независимо обученных видеомодели и аудиомодели. Основным достижением работы является демонстрация эффективности

стратегии разделения обработки модальностей с последующим объединением результатов на уровне решений, что обеспечивает баланс между точностью распознавания и вычислительной сложностью.

Экспериментальные результаты показали, что интеграция аудиоинформации с видеоданными значительно повышает точность распознавания по сравнению с использованием отдельных модальностей. Наилучшие результаты (91.67%) были достигнуты с использованием методов среднего арифметического и произведения вероятностей для цветных видеомоделей, а также доверительного объединения для черно-белых моделей. Особую ценность представляет тот факт, что даже при умеренном количестве обучающих данных (144 видео для тестирования) мультимодальный подход обеспечивает конкурентоспособную точность по сравнению с современными методами, основанными на трансформерах.

Практическая значимость работы обусловлена несколькими аспектами. Во-первых, предложенные методы позднего объединения демонстрируют высокую устойчивость к отсутствию или низкому качеству одной из модальностей во время инференции, что критически важно для реальных применений. Во-вторых, простые методы объединения (среднее арифметическое и произведение вероятностей) обеспечивают минимальные вычислительные затраты при сохранении высокой точности, что позволяет их использование на устройствах с ограниченными ресурсами. В-третьих, исследование показало, что для черно-белых видеомоделей интеграция аудиоинформации дает больший прирост точности (+5.58%), что открывает возможности для создания энергоэффективных систем распознавания, использующих только монохромное видео.

Ограничения настоящего исследования включают ограниченное количество классов действий (6 классов из UCF101) и относительно небольшой размер обучающей выборки. В дальнейших исследованиях целесообразно расширение количества распознаваемых действий, увеличение разнообразия обучающих данных, а также исследование гибридных подходов, сочетающих раннее и позднее объединение модальностей. Перспективным направлением является разработка адаптивных методов объединения, способных динамически выбирать оптимальную стратегию интеграции в зависимости от качества доступных модальностей и характеристик конкретного видео.

Таким образом, результаты исследования подтверждают эффективность мультимодального подхода к распознаванию действий человека в видео и предоставляют практические рекомендации по выбору оптимальных методов объединения аудиовизуальной информации. Предложенные решения имеют потенциал применения в системах видеонаблюдения, медицинской диагностике, спортивной аналитике и мультимедийных платформах, где требуется высокая точность распознавания при ограниченных вычислительных ресурсах.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Sun Z., Ke Q., Rahmani H., Bennamoun M., Wang G., Liu J. Human Action Recognition from Various Data Modalities: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
2. Carreira J., Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299-6308.
3. Bayouh K., Knani R., Hamdaoui F., Mtibaa A. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets, *The Visual Computer*, 2022, Vol. 38. No. 8, pp. 2939-2970.
4. Feichtenhofer C., Pinz A., Zisserman A. Convolutional two-stream network fusion for video action recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1933-1941.
5. Snoek C.G., Worring M., Smeulders A.W. Early versus late fusion in semantic video analysis, *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 2005, pp. 399-402.
6. Pawlowski M., Wroblewska A., Sysko-Romanczuk S. Effective Techniques for Multimodal Data Fusion: A Comparative Analysis, *Sensors*, 2023, Vol. 23, No. 5, pp. 2381.
7. Caba Heilbron J., Escorcia V., Ghanem B., Niebles J.C. ActivityNet: A large-scale video benchmark for human activity understanding, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961-970.

8. Ngiam J., Khosla A., Kim M., Nam J., Lee H., Ng A.Y. Multimodal deep learning, *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011, Vol. 2, No. 3, pp. 4.
9. Demšar J. Statistical comparisons of classifiers over multiple data sets, *The Journal of Machine Learning Research*, 2006, Vol. 7, pp. 1-30.
10. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.
11. Hershey S., Chaudhuri S., Ellis D. P., Gemmeke J. F., Jansen A., Moore R. C., Xu R.C. CNN architectures for large-scale audio classification, *ICASSP 2017 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131-135.
12. Wang L., Xiong Y., Wang Z., Qiao Y., Lin D., Tang X., Van Gool L. Temporal segment networks: Towards good practices for deep action recognition, *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 20-36.
13. Soomro K., Zamir A.R., Shah M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. Orlando: CRCV, 2012, 12 p.
14. Bradski G., Kaehler A. Learning OpenCV: Computer Vision with the OpenCV Library. Sebastopol: O'Reilly Media, 2008, 580 p.
15. McFee B., Raffel C., Liang D. et al. librosa: Audio and music signal analysis in python, *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18-25.
16. Hu J., Shen L., Sun G. Squeeze-and-excitation networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132-7141.
17. Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014. arXiv:1406.1078.
18. Tompson J., Goroshin R., Jain A., LeCun Y., Bregler C. Efficient object localization using convolutional networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648-656.
19. Ioffe S., Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift, *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 448-456.
20. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. Rethinking the inception architecture for computer vision, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818-2826.
21. Loshchilov I., Hutter F. Decoupled weight decay regularization, *International Conference on Learning Representations*, 2019.
22. Shorten C., Khoshgoftar T.M. A survey on image data augmentation for deep learning, *Journal of Big Data*, 2019, Vol. 6, pp. 1-48.
23. Park D.S., Chan W., Zhang Y., Chiu C.C., Zoph B., Cubuk E.D., Le Q.V. SpecAugment: A simple data augmentation method for automatic speech recognition, *arXiv:1904.08779*, 2019.
24. Perez L., Wang J. The effectiveness of data augmentation in image classification using deep learning, *arXiv:1712.04621*, 2017.
25. Smith L.N., Topin N. Super-convergence: Very fast training of neural networks using large learning rates, *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019, Vol. 11006, pp. 1100612.
26. Piczak K.J. ESC: Dataset for environmental sound classification, *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 1015-1018.
27. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research*, 2014, Vol. 15, No. 1, pp. 1929-1958.
28. Gemmeke J.F., Ellis D.P., Freedman D., Jansen A., Lawrence W., Moore R.C., Vincent M. Audio set: An ontology and human-labeled dataset for audio events, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776-780.
29. Hinton G., Vinyals O., Dean J. Distilling the knowledge in a neural network, *arXiv:1503.02531*, 2015.
30. Wolpert D.H. Stacked generalization, *Neural Networks*, 1992, Vol. 5, No. 2, pp. 241-259.

Гергов Валерий Заурович – ФНЦ «Кабардино-Балкарский научный центр Российской академии наук»; e-mail: ronstoun@yandex.ru; г. Нальчик, Россия; тел.: +79993001081; аспирант; <https://orcid.org/0009-0007-3378-7662>.

Gergov Valeriy Zaurovich – Federal Scientific Center «Kabardin-Balkar Scientific Center of the Russian Academy of Sciences»; e-mail: ronstoun@yandex.ru; Nalchik, Russia; phone: +79993001081; graduate student; <https://orcid.org/0009-0007-3378-7662>.