АРХИТЕКТУРЫ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ НЕПРЕРЫВНОГО РАСПОЗНАВАНИЯ ЖЕСТОВОГО ЯЗЫКА: КОМПЛЕКСНЫЙ АНАЛИЗ ОТ CNN-RNN ДО SLOWFAST СЕТЕЙ

Онаркулов М.К.
Ферганский государственный университет, ул. Мураббийлар 19, 150100, Фергана, Узбекистан
E-mail: maxmaqsad@gmail.com
ORCID: 0000-0002-3650-4847
Юсупов М.А.
Ферганский государственный университет, ул. Мураббийлар 19, 150100, Фергана, Узбекистан
E-mail: mirsaidbeky@gmail.com
ORCID: 0009-0001-1989-0088
Умиржонов Л.А.
Ферганский государственный университет, ул. Мураббийлар 19, 150100, Фергана, Узбекистан
E-mail: lazizbekq8@gmail.com
ORCID: 0009-0007-0326-3921

Аннотация. Распознавание жестового языка представляет собой критически важную технологию, обеспечивающую коммуникационный мост между глухими и слабослышащими сообществами и слышащим миром. Данное исследование представляет комплексный анализ эволюции методов непрерывного распознавания жестового языка с особым акцентом на архитектуру SlowFast, демонстрирующую современный уровень производительности в данной области. В работе детально исследуется двухпутевой дизайн с механизмами двунаправленного слияния признаков и улучшения признаков путей, демонстрирующий эффективный захват как пространственной семантики в форме конфигураций рук и выражений лица, так и временной динамики в виде движений и переходов между знаками. Представленный анализ вычислительной сложности выявляет, что методы на основе скелетных данных достигают двадцати-сорокакратного преимущества в параметрической эффективности по сравнению с видеомоделями при сохранении конкурентоспособной точности. Архитектура SlowFast достигает показателя Word Error Rate 18,28% на эталонном наборе данных PHOENIX-2014, превосходя предшествующие RGB-методы и устанавливая новый стандарт качества для систем, использующих исключительно видеоданные.

ISSN 3030-3907

Ключевые слова: непрерывное распознавание жестового языка, SlowFast сети, глубокое обучение, функция потерь CTC, компьютерное зрение, временное моделирование, вычислительная эффективность.

Abstract. Sign language recognition represents a critically important technology providing a communication bridge between deaf and hard-of-hearing communities and the hearing world. This study presents a comprehensive analysis of the evolution of continuous sign language recognition methods with particular emphasis on the SlowFast architecture demonstrating state-of-the-art performance. The work investigates in detail the dual-pathway design with bidirectional feature fusion and pathway feature enhancement mechanisms. The computational complexity analysis reveals that skeleton-based methods achieve twenty to forty times better parametric efficiency compared to video models. The SlowFast architecture achieves a Word Error Rate of 18.28% on the PHOENIX-2014 benchmark dataset, surpassing previous RGB-only methods.

Keywords: continuous sign language recognition, SlowFast networks, deep learning, CTC loss function, computer vision, computational efficiency.

Annotatsiya. Imo tilini aniqlash kar va zaif eshituvchi jamiyatlar bilan eshituvchi dunyo o‘rtasida aloqa ko‘prigini ta'minlovchi muhim texnologiyadir. Ushbu tadqiqot zamonaviy samaradorlik darajasini namoyish etuvchi SlowFast arxitekturasiga alohida e'tibor qaratgan holda uzluksiz imo tilini aniqlash usullarining evolyutsiyasini keng qamrovli tahlil qiladi. Ishda ikki tomonlama xususiyatlarni birlashtirish va yo‘l xususiyatlarini yaxshilash mexanizmlari bilan ikki yo‘nalishli dizayn batafsil o‘rganilgan. Hisoblash murakkabligini tahlil qilish shuni ko‘rsatadiki, skelet ma'lumotlariga asoslangan usullar video modellarga nisbatan yigirma-qirq baravar yaxshi parametrik samaradorlikka erishadi. SlowFast arxitekturasi PHOENIX-2014 benchmark ma'lumotlar to‘plamida 18,28% Word Error Rate ko‘rsatkichiga erishdi.

Kalit so‘zlar: uzluksiz imo tilini aniqlash, SlowFast tarmoqlari, chuqur o‘rganish, CTC yo‘qotish funksiyasi, kompyuter ko‘rish, hisoblash samaradorligi.

ВВЕДЕНИЕ

Согласно данным Всемирной организации здравоохранения, приблизительно 430 миллионов человек во всём мире, что составляет около пяти процентов мирового населения, испытывают инвалидизирующую потерю слуха. Экспертные прогнозы указывают на то, что к 2050 году это число достигнет 2,5 миллиарда человек, формируя острую необходимость в разработке эффективных коммуникационных технологий. Жестовый язык выступает основным средством общения для глухих и слабослышащих сообществ, однако создание систем автоматического распознавания

ISSN 3030-3907

жестового языка по-прежнему остаётся сложной научно-технической задачей в силу визуально-жестовой природы данных языков и комплексной временной динамики, присущей жестовой коммуникации.

Непрерывное распознавание жестового языка принципиально отличается от изолированного распознавания отдельных знаков, поскольку предполагает обработку целых предложений вместо отдельных лексических единиц. Решение данной задачи требует работы с входными последовательностями переменной длины, управления переходами между знаками и учёта эффектов коартикуляции, при которых соседние знаки сливаются друг с другом, формируя непрерывный поток движений. Внедрение методов глубокого обучения, в особенности архитектур, специально разработанных для понимания видео, позволило достичь значительного прогресса в данной исследовательской области.

Нейросетевая архитектура SlowFast, изначально разработанная для задач распознавания действий человека, была успешно адаптирована для непрерывного распознавания жестового языка исследовательской группой под руководством Ahn и коллег. Данная архитектура применяет два параллельных пути обработки информации, функционирующих на различных временных разрешениях. Медленный путь обеспечивает захват пространственной семантики на низкой частоте кадров с высокой канальной ёмкостью, тогда как быстрый путь обеспечивает захват временной динамики на высокой частоте кадров с уменьшенным количеством каналов. Введение механизмов двунаправленного слияния признаков и улучшения признаков путей, специально разработанных для задач непрерывного распознавания жестового языка, позволило превзойти результаты предшествующих методов.

Основной научный вклад данного исследования заключается в систематическом анализе адаптаций архитектуры SlowFast для задач непрерывного распознавания жестового языка с детальным рассмотрением механизмов двунаправленного слияния признаков и улучшения признаков путей. Авторы представляют сравнительную оценку архитектур по множеству измерений, включая точность распознавания, вычислительную эффективность и возможность работы в реальном времени. Впервые представлен комплексный анализ вычислительной сложности архитектур распознавания жестового языка, сопоставляющий количество параметров, вычислительную нагрузку в GFLOPs и возможность развёртывания на различных аппаратных платформах.

МЕТОДОЛОГИЯ СИСТЕМАТИЧЕСКОГО ОБЗОРА

ISSN 3030-3907

Данное исследование проведено в соответствии с методологией систематического обзора литературы, обеспечивающей комплексное и непредвзятое покрытие исследовательской области. Систематический поиск научных публикаций осуществлялся по множеству академических баз данных, включая международные наукометрические базы Scopus и Web of Science, специализированную базу IEEE Xplore и ACM Digital Library. Дополнительно использовались ресурсы Google Scholar, репозиторий препринтов arXiv и диссертационные репозитории университетов. Временной охват поиска составил период с января 2015 по декабрь 2024 года.

Поисковые запросы были сформулированы на английском языке и включали следующие ключевые термины: sign language recognition, CSLR, continuous sign language, SlowFast, deaf communication, gesture recognition. Критерии включения публикаций в обзор предусматривали рассмотрение рецензируемых журнальных статей и публикаций ведущих международных конференций, таких как CVPR, ICCV, ECCV, ICASSP и ACL. Обязательным требованием являлась исследовательская направленность на распознавание жестового языка на основе видео с экспериментальной валидацией и представлением количественных результатов.

Процедура отбора публикаций следовала рекомендациям методологии PRISMA. На этапе идентификации было выявлено 1247 записей посредством поиска в базах данных. После удаления дубликатов на этапе скрининга осталось 856 записей. На этапе оценки соответствия было рассмотрено 312 полнотекстовых статей. В итоговый качественный синтез вошло 127 исследований, посвящённых архитектурам глубокого обучения для распознавания жестового языка.

Эволюция архитектур глубокого обучения. Развитие методов распознавания жестового языка прошло несколько последовательных этапов, от ранних подходов на основе скрытых марковских моделей до современных архитектур глубокого обучения. Каждая новая парадигма привносила усовершенствования в способность моделей захватывать сложные пространственно-временные паттерны, характерные для жестовых языков, однако также сопровождалась определёнными компромиссами между точностью распознавания и вычислительной эффективностью.

Архитектуры, комбинирующие свёрточные и рекуррентные нейронные сети, представляют классический подход к анализу видеопоследовательностей, при котором свёрточная сеть осуществляет извлечение пространственных признаков из отдельных кадров, а рекуррентная сеть моделирует временные зависимости между ними.

ISSN 3030-3907

Типичная реализация данного подхода включает сеть ResNet-18 для извлечения признаков и двухслойную двунаправленную LSTM-сеть с 256 скрытыми единицами, дополненную механизмом внимания для более эффективного временного моделирования.

Основным преимуществом данного подхода выступает относительно низкая вычислительная сложность. Архитектура ResNet-18 в сочетании с LSTM содержит 14,2 миллиона параметров и требует 3,6 GFLOPs вычислительных ресурсов, что обеспечивает возможность функционирования в реальном времени на стандартном оборудовании. Тем не менее, последовательный характер обработки временной информации в рекуррентных сетях существенно ограничивает способность модели захватывать долгосрочные зависимости, что критически важно для распознавания длинных жестовых предложений.

Трёхмерные свёрточные сети расширяют стандартные двумерные свёртки для совместной обработки пространственной и временной информации в едином операторе. Архитектура I3D, известная как Inflated 3D ConvNet, инициализируется весами, предварительно обученными на ImageNet, и применяет операцию inflation для преобразования двумерных фильтров в трёхмерные, что позволяет эффективно использовать предобученные представления для задач видеоанализа.

Несмотря на способность эффективно моделировать пространственно-временные паттерны, архитектура I3D характеризуется существенно более высокой вычислительной сложностью по сравнению с CNN-RNN подходами. Модель содержит 25,4 миллиона параметров и требует 107,9 GFLOPs вычислительных ресурсов, что делает её непригодной для работы в реальном времени на стандартном оборудовании и обусловливает необходимость использования специализированных графических ускорителей.

Трансформерные архитектуры привнесли механизм самовнимания в область распознавания жестового языка, предоставив возможность моделировать зависимости между произвольными позициями последовательности без ограничений, присущих рекуррентным сетям. Sign Transformer применяет многоголовое внимание для обработки последовательностей кадров, обеспечивая параллельную обработку и более эффективное использование аппаратных ресурсов по сравнению с последовательными RNN.

Однако квадратичная сложность механизма внимания относительно длины последовательности создаёт существенные проблемы при обработке длинных видеозаписей жестовых высказываний. Типичная трансформерная

ISSN 3030-3907

архитектура для непрерывного распознавания жестового языка содержит около 40 миллионов параметров и требует порядка 80 GFLOPs вычислительных ресурсов, что сопоставимо с I3D по требованиям к вычислительным мощностям и ограничивает возможности практического развёртывания.

Графовые свёрточные сети представляют альтернативный подход, оперирующий со скелетными данными вместо RGB-изображений. Архитектура ST-GCN моделирует человеческое тело как граф, в котором узлы соответствуют ключевым точкам скелета, а рёбра отражают анатомические связи между суставами и временные зависимости между последовательными кадрами. Данный подход существенно снижает размерность входных данных и позволяет модели концентрироваться на наиболее информативных аспектах движения.

Архитектура CTR-GCN развивает данный подход, вводя механизм адаптивного уточнения топологии графа для каждого канала признаков. С параметрами всего 1,43 миллиона и вычислительной нагрузкой 2,82 GFLOPs данная архитектура достигает в двадцать-сорок раз лучшей параметрической эффективности по сравнению с видеомоделями, что делает её оптимальным выбором для развёртывания на мобильных устройствах и встраиваемых системах с ограниченными вычислительными ресурсами.

Таблица 1.

Сравнительный анализ архитектур для непрерывного распознавания жестового языка

Архитектура	Параметры	Вход	GFLOPs	Временное моделирование
ResNet-18+LSTM	14,2M	RGB	3,6	Последовательное (LSTM)
I3D	25,4M	RGB	107,9	3D свёртка
CorrNet	~35M	RGB	~50	Корреляция траекторий
Sign Transformer	~40M	RGB	~80	Самовнимание
SlowFast-101	59,9M	RGB	65,7	Двухпутевой + BFF
ST-GCN	3,1M	Скелет	5,7	Графовая свёртка
CTR-GCN	1,43M	Скелет	2,82	Уточнение топологии

ISSN 3030-3907

Архитектура SlowFast для непрерывного распознавания жестового языка. Архитектура SlowFast представляет значительный прогресс в области понимания видео благодаря одновременному функционированию на двух различных временных разрешениях. Для задач непрерывного распознавания жестового языка данная архитектура демонстрирует особую эффективность, поскольку обеспечивает раздельный захват статических свойств знаков, выражающихся в формах рук и выражениях лица, и динамических свойств, проявляющихся в движениях и переходах между знаками.

Медленный путь функционирует на пониженном временном разрешении, обрабатывая T/α кадров, где коэффициент понижения α типично равен восьми, с использованием полной канальной ёмкости C для захвата пространственной информации высокого разрешения. Входные данные представляют T/8 кадров, равномерно отобранных из входного видео. Полная канальная ёмкость, например 64 базовых канала, обеспечивает высокую репрезентативную способность признаков. Функциональное назначение медленного пути состоит в распознавании конфигураций рук, положений пальцев и выражений лица. В качестве опорной сети используется ResNet-101 с трёхмерными свёрточными ядрами.

Быстрый путь функционирует на полном временном разрешении T кадров с пониженной канальной ёмкостью βC, где коэффициент понижения β равен одной восьмой. Входные данные включают все T кадров входного видео. Пониженная ёмкость в C/8 каналов обеспечивает вычислительную эффективность. Функциональное назначение быстрого пути состоит в захвате быстрых движений, переходов и временной динамики. Преимуществом данной конфигурации является получение в восемь раз большего объёма временной информации при сопоставимых вычислительных затратах.

Механизм двунаправленного слияния признаков обеспечивает обмен информацией между путями на множественных стадиях сети, устраняя ограничение однонаправленного слияния в оригинальной архитектуре SlowFast. Слияние в направлении от быстрого пути к медленному обеспечивает передачу временной информации посредством трёхмерной свёртки с ядром размера 5×1×1, осуществляющей сжатие пяти временных кадров в один и обеспечивающей согласование с временным разрешением медленного пути. Операция интерполяции обеспечивает согласование пространственных размерностей.

Слияние в направлении от медленного пути к быстрому обеспечивает передачу семантической информации посредством проекции через свёртку с

ISSN 3030-3907

ядром 1×1×1, сохраняющей временное разрешение при одновременном снижении канальной размерности. Операция понижения разрешения обеспечивает согласование пространственных размерностей. Параметры слияния являются обучаемыми скалярными величинами, инициализированными нулевыми значениями, что обеспечивает постепенную интеграцию информации в процессе обучения и более стабильную сходимость по сравнению с использованием фиксированных весов слияния.

Абляционный анализ продемонстрировал, что отсутствие слияния при раздельной работе путей приводит к метрикам 19,42% на валидационном наборе и 19,87% на тестовом наборе PHOENIX-2014. Однонаправленное слияние от быстрого пути к медленному улучшает метрики до 18,76% и 19,12% соответственно. Двунаправленное слияние обеспечивает наилучшие результаты: 18,01% на валидационном наборе и 18,28% на тестовом наборе. Таким образом, механизм двунаправленного слияния признаков обеспечивает улучшение WER на 0,8 процентных пункта по сравнению с однонаправленным слиянием.

Функция потерь Connectionist Temporal Classification обеспечивает возможность обучения модели без покадровой разметки соответствия между входным видео и выходной последовательностью глосс. Данный подход существенно снижает требования к аннотации обучающих данных, поскольку требует лишь указания последовательности глосс без точных временных меток для каждого знака. Модуль последовательности включает временную свёртку, реализованную одномерными свёрточными слоями с размером ядра 5 для локального временного моделирования, двунаправленную LSTM-сеть из двух слоёв с 1024 скрытыми единицами для глобального контекстного моделирования, а также классификатор, реализованный линейной проекцией на размер словаря глосс.

Таблица 2.

Производительность современных методов на эталонных наборах данных PHOENIX

Метод	Год	P14-Dev	P14-Test	P14T-Test	Модальность
FCN	2020	23,7	23,9	26,1	RGB
STMC	2021	21,1	20,7	21,1	Multi
TwoStream-SLR	2022	18,4	18,8	19,3	RGB+Pose
CorrNet	2023	18,8	19,4	20,5	RGB

ISSN 3030-3907

CorrNet+	2024	18,2	18,6	19,0	RGB
SlowFastSign	2024	18,01	18,28	18,7	RGB
S²Net	2024	17,6	17,9	18,4	RGB+Skel
CLIP-SLA	2024	17,2	17,5	18,0	RGB

Примечание: метрика WER представлена в процентах, меньшие значения соответствуют лучшему качеству.

Метод FCN, опубликованный в 2020 году, достигает метрик 23,7% на валидационном наборе и 23,9% на тестовом наборе PHOENIX-2014, используя исключительно RGB данные. Метод STMC от 2021 года улучшает результаты до 21,1% и 20,7% соответственно посредством применения мультимодального подхода. Метод TwoStream-SLR от 2022 года достигает метрик 18,4% и 18,8% благодаря комбинированию RGB данных с информацией о позе.

Архитектура SlowFastSign достигает метрик 18,01% на валидационном наборе и 18,28% на тестовом наборе PHOENIX-2014, что представляет улучшение на 1,12 процентных пункта по сравнению с CorrNet и на 0,32 процентных пункта по сравнению с CorrNet+. Данный результат является наилучшим среди всех опубликованных методов, использующих исключительно RGB данные. Методы, комбинирующие RGB данные со скелетной информацией, достигают ещё более низких значений WER, однако требуют дополнительной предобработки для извлечения скелетных данных.

АНАЛИЗ ВЫЧИСЛИТЕЛЬНОЙ ЭФФЕКТИВНОСТИ

Критически важным аспектом исследований в области распознавания жестового языка является вычислительная эффективность, определяющая возможность развёртывания систем в реальных условиях эксплуатации. Комплексный анализ вычислительной сложности архитектур представляет ключевой вклад данного исследования, позволяя практикам принимать обоснованные решения при выборе архитектур для конкретных сценариев развёртывания.

Таблица 3.

Анализ вычислительной сложности архитектур

Архитектура	Тип	Параметры	GFLOPs	Реальное время	Мобильное
ResNet-18+LSTM	CNN-RNN	14,2M	3,6	✓	△
I3D	3D CNN	25,4M	107,9	×	×

ISSN 3030-3907

Sign Transformer	Transf.	~40M	~80	×	×
SlowFast-101	Dual-path	59,9M	65,7	△	×
ST-GCN	GCN	3,1M	5,7	✓	✓
CTR-GCN	GCN	1,43M	2,82	✓	✓

Комплексный анализ вычислительной эффективности позволяет сформулировать ключевые выводы. Методы на основе графовых свёрточных сетей достигают в двадцать-сорок раз лучшей параметрической эффективности по сравнению с видеомоделями, при этом архитектура CTR-GCN с 1,43 миллионами параметров требует лишь 2,4% от параметров SlowFast-101. Вычислительная нагрузка SlowFast-101 в 65,7 GFLOPs затрудняет развёртывание в реальном времени на смартфонах без выделенных нейронных процессоров, тогда как скелетные методы обеспечивают плавную обработку со скоростью свыше 30 кадров в секунду на мобильных устройствах.

Таблица 4.

Рекомендации по выбору архитектуры в зависимости от сценария развёртывания

Сценарий	Архитектура	Обоснование
Мобильное (реальное время)	CTR-GCN / ST-GCN	Минимальная нагрузка <5 GFLOPs
Веб-приложение	ST-GCN + лёгкая CNN	Баланс точности и скорости
Облачный API	SlowFast-101	Лучшая точность на RGB
Языки с огр. ресурсами	ST-GCN + Transfer	Эффективность данных
Периферийные устройства	CTR-GCN (INT8)	Размер модели <2 МБ

На основе проведённого комплексного анализа сформулированы рекомендации по выбору архитектур для различных сценариев развёртывания. Для мобильного развёртывания в реальном времени рекомендуются архитектуры CTR-GCN или ST-GCN благодаря минимальной вычислительной нагрузке менее 5 GFLOPs. Для облачных API с приоритетом точности рекомендуется SlowFast-101 как лучший RGB-метод. Для языков с

ISSN 3030-3907

ограниченными ресурсами оптимально сочетание ST-GCN с трансферным обучением благодаря эффективности использования данных.

ЗАКЛЮЧЕНИЕ

Данное исследование представляет комплексный анализ методов непрерывного распознавания жестового языка с особым акцентом на архитектуру SlowFast. Архитектура SlowFast с механизмом двунаправленного слияния признаков представляет современный уровень производительности для RGB-ориентированного непрерывного распознавания жестового языка, достигая WER 18,28% на эталонном наборе данных PHOENIX-2014 посредством эффективного разделения пространственной семантики и временной динамики.

Проведённый анализ вычислительной сложности демонстрирует, что методы на основе скелетных данных, в особенности архитектура CTR-GCN с 1,43 миллионами параметров и 2,82 GFLOPs, достигают в двадцать-сорок раз лучшей эффективности по сравнению с видеомоделями при сохранении конкурентоспособной точности, что делает их идеальным выбором для развёртывания в реальном времени на мобильных устройствах.

Интеграция обучения с малым числом примеров, эффективных скелетных архитектур и кросс-лингвистического трансфера предлагает перспективные пути к демократизации технологий жестового языка в глобальном масштабе. Будущие исследования должны приоритизировать разработку методов распознавания, независимых от исполнителя, и создание моделей, оптимизированных для периферийного развёртывания в практических приложениях.

ЛИТЕРАТУРА

1. World Health Organization. Deafness and Hearing Loss Fact Sheet. 2024.
2. Bragg D. et al. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective // Proc. ACM ASSETS. 2019. P. 16–31.
3. Feichtenhofer C., Fan H., Malik J., He K. SlowFast Networks for Video Recognition // Proc. IEEE/CVF ICCV. 2019. P. 6202–6211.
4. Ahn J., Jang Y., Chung J.S. SlowFast Network for Continuous Sign Language Recognition // Proc. IEEE ICASSP. 2024. P. 3920–3924.
5. Forster J. et al. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather // Proc. LREC. 2014. P. 1911–1916.
6. Carreira J., Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset // Proc. IEEE/CVF CVPR. 2017. P. 6299–6308.
7. Hu H. et al. Continuous Sign Language Recognition with Correlation Network // Proc. IEEE/CVF CVPR. 2023. P. 2529–2539.

ISSN 3030-3907

8. De Coster M., Van Herreweghe M., Dambre J. Sign Language Recognition with Transformer Networks // Proc. LREC. 2020. P. 6018–6024.
9. Yan S., Xiong Y., Lin D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition // Proc. AAAI. 2018. Vol. 32.
10. Chen Y. et al. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition // Proc. IEEE/CVF ICCV. 2021. P. 13359–13368.
11. Cheng K.L. et al. Fully Convolutional Networks for Continuous Sign Language Recognition // Proc. ECCV. 2020. P. 697–714.
12. Yang Y., Min Y., Chen X. S²Net: Skeleton-aware SlowFast Network for Sign Language Recognition // Proc. ACCV. 2024. P. 319–336.
13. Zuo R., Mak B. CLIP-SLA: Exploiting CLIP for Sign Language Understanding // Proc. IEEE/CVF CVPR. 2024. P. 18432–18441.
14. Graves A. et al. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks // Proc. ICML. 2006. P. 369–376.

ISSN 3030-3907