ՀԱՅԱՍՏԱՆԻ ՀԱՆՐԱՊԵՏՈՒԹՅԱՆ ԿՐԹՈՒԹՅԱՆ, ԳԻՏՈՒԹՅԱՆ,
ՄՇԱԿՈՒՅԹԻ ԵՎ ՍՊՈՐՏԻ ՆԱԽԱՐԱՐՈՒԹՅՈՒՆ

ՀԱՅԱՍՏԱՆԻ ԱԶԳԱՅԻՆ ՊՈԼԻՏԵԽՆԻԿԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

Գալստյան Դավիթ Մարատի

ՏԵՍԱՆՅՈՒԹՈՒՄ ՇԱՐԺՈՒՄՆԵՐԻ ԼԵԶՎԻՑ ՏԵՔՍՏԻ ՁԵՎԱՎՈՐՄԱՆ 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

ГАЛСТЯН ДАВИД МАРАТОВИЧ

РАЗРАБОТКА СРЕДСТВ АВТОМАТИЗАЦИИ ФОРМИРОВАНИЯ ТЕКСТА ИЗ ЖЕСТОВОГО ЯЗЫКА В ВИДЕОМАТЕРИАЛАХ

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. Աշոտ Գևորգի Հարությունյան տ.գ.թ. Գոռ Արսենի Պետրոսյան
Առաջատար կազմակերպություն՝	ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտ

Ատենախոսության պաշտպանությունը կայանալու է 2025թ. հուլիսի 15-ին, ժամը 12^{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. Ашот Геворкович Арутюнян к.т.н. Гор Арсенович Петросян
Ведущая организация:	Институт проблем информатики и автоматизации НАН РА

Защита диссертации состоится 15-го июля 2025г. в 12^{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. Развитие технологий автоматизации формирования текста из языка жестов (ЯЖ) в видеоматериалах в последние годы стало важным и актуальным направлением, что обусловлено появлением инновационных подходов в областях компьютерного зрения и машинного обучения, а также расширением сфер их применения. В частности, внедрение этих технологий позволило повысить эффективность коммуникации для лиц с нарушениями слуха, а также расширить интеграцию средств коммуникации в виртуальной реальности, медицине и других областях.

Современные системы, основанные на сверточных нейронных сетях и трансформерах, обеспечивают высокую точность распознавания ЯЖ и возможность анализа данных в реальном времени. Однако, несмотря на достигнутый значительный прогресс, существуют также значительные ограничения, связанные с ресурсами, влиянием условий окружающей среды и разнообразием данных. Эти проблемы подчеркивают важность разработки автоматизированных систем и повышения их эффективности.

Диссертация посвящена разработке архитектурных решений и методов, обеспечивающих автоматизированное преобразование языка жестов в текст на основе видеоданных с использованием методов компьютерного зрения и глубинного обучения. Предлагаемые подходы направлены на повышение точности распознавания, снижение вычислительных затрат и обеспечение адаптивности систем в условиях ограниченных ресурсов.

Объект исследования. Архитектурные решения систем автоматизированного формирования текста из ЯЖ в видеоматериалах, изучение факторов, влияющих на эффективность этих архитектур, в частности, методы предварительной обработки данных, техники оптимизации моделей и подходы к управлению ресурсами.

Цель работы. Разработка подходов к повышению эффективности систем распознавания ЯЖ из видеоматериалов, обеспечивающих повышение точности, снижение энергопотребления и эффективное использование ресурсов без значительного ущерба для качества распознавания.

Методы исследования. В процессе выполнения диссертации были использованы современные методы глубокого обучения, трансформерные модели, техники предварительной обработки данных и методы сжатия моделей для уменьшения размеров системы. Кроме того, исследование сопровождалось экспериментальной оценкой систем на основе реальных и синтезированных данных.

Научная новизна:

- В рамках диссертации предложены подходы к разработке автоматизированных средств формирования текста из языка жестов в видеоматериалах, которые благодаря применению современных моделей компьютерного зрения и глубокого обучения обеспечивают высокую точность, низкие требования к вычислительным ресурсам и адаптивность к различным условиям эксплуатации.
- Разработан механизм предварительной обработки видеоданных, который с помощью методов снижения шума, нормализации освещенности и

выбора ключевых кадров позволил сократить объем обрабатываемых данных до 95% и снизить вычислительную нагрузку в три раза при минимальной потере точности в 1,5%.

- Создан метод генерации синтетических данных на основе генеративных состязательных сетей (ГСС), который за счет расширения набора данных на 15000 образцов позволяет увеличить точность распознавания жестов на 7,8%, при этом время обучения увеличивается всего на 2%.
- Предложен подход к сжатию моделей на основе адаптации трансформерных архитектур и применения технологии TensorRT, благодаря чему удалось ускорить выполнение моделей до 2,6 раза при снижении точности распознавания до 1,1%.

Практическая ценность работы. Разработанное программное средство (ПС) „MLT“ внедрено в ООО "ЭЙ АЙ СПЕЙС", которое, применяя модульную архитектуру, техники многомодальной интеграции и оптимизации трансформерной архитектуры, обеспечило возможность распознавания 15000+ жестов с общей точностью 88,8% и средней скоростью 32,7 FPS, однако требуя 512…685 Мб памяти и вызывая в среднем 24,6 мс/к времени задержки передачи данных

На защиту выносятся:

- методы предварительной обработки данных и снижения шума;
- способы сжатия моделей;
- способы архитектуры автоматизированного формирования текста из ЯЖ в видеоматериалах;
- программный инструмент автоматизированного формирования текста из языка жестов в видеоматериалах “MLT System”.

Достоверность научных положений подтверждается экспериментальными исследованиями, проведенными на основе реальных и синтезированных данных, которые показывают эффективность и применимость предложенных методов в практических задачах.

Внедрение. Разработанные программные решения внедрены в ООО "ЭЙ АЙ СПЕЙС", где они применяются для автоматизированного формирования текста из ЯЖ в видеоматериалах, повышая скорость и точность обработки данных и сокращая время проектирования.

Апробация работы. Основные научные и практические результаты диссертации докладывались и обсуждались 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. Основные положения диссертации представлены в девяти научных работах, список которых приведен в конце автореферата.

Структура и объем диссертации. Диссертация состоит из введения, трех глав, основных выводов, списка литературы, включающего 114 наименований, и четырех приложений. В первом приложении представлен акт внедрения диссертации, во втором - фрагменты функциональной структуры реализованного программного средства “MLT”, в третьем – списки рисунков, таблиц и сокращений. Основной объем диссертации составляет 122 страниц, а вместе с приложениями - 136 страниц. Диссертация написана на армянском языке

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, сформулированы цель и задачи исследования, представлены используемые методы, научная новизна, практическое значение и основные положения, выносимые на защиту. **В первой главе** представлены теоретические и практические аспекты построения систем автоматизированного преобразования языка жестов в текст. Обоснована важность применения мультимодальных подходов, сочетающих визуальные и сенсорные данные, а также глубоких нейронных архитектур, обеспечивающих устойчивость к условиям съёмки, индивидуальным особенностям жестов и шуму видеопотока.

Системы распознавания языка жестов (РЯЖ) включают множество модальностей, необходимых для полного понимания. Распознавание жестов состоит из анализа не только движений рук, но и выражений лица, положения головы и движений тела (рис. 1). Интеграция этих модальностей может значительно улучшить работу систем РЯЖ, так как они несут дополнительную лингвистическую и эмоциональную информацию.

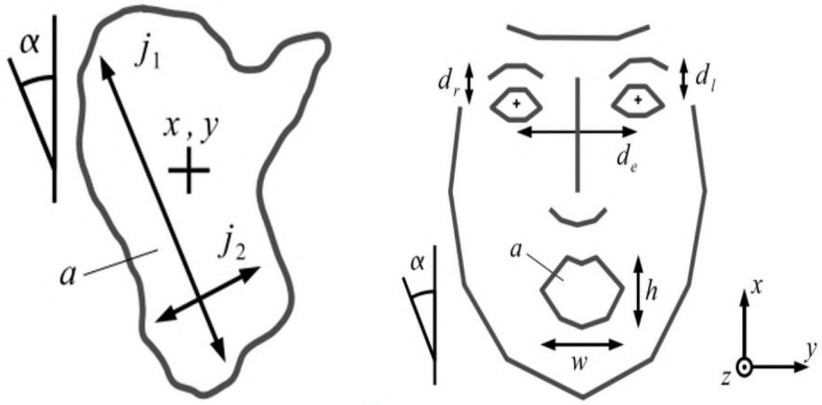

Рис. 1. Особенности лица и руки в распознавания языка жестов

Исследования показывают, что эффективность распознавания ЯЖ существенно зависит от условий окружающей среды, таких как освещение, фоновый шум и другие внешние факторы (рис. 2).

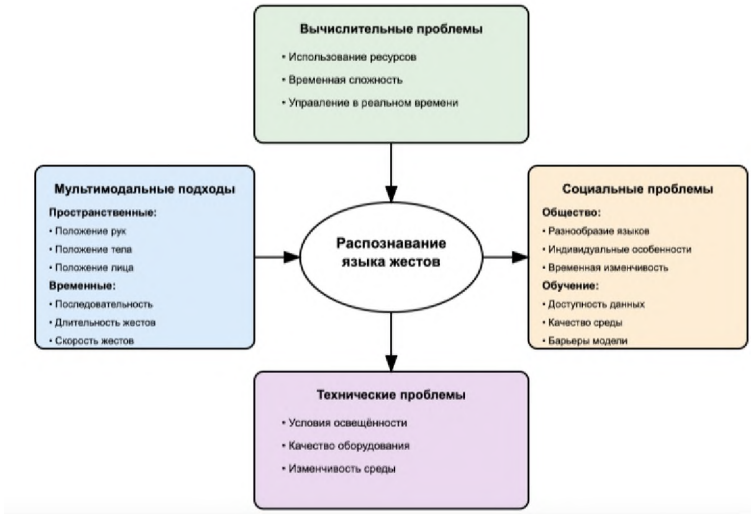

Рис. 2. Зависимость распознавания языка жестов от влияющих факторов окружающей среды

Особенности систем распознавания ЯЖ, обусловленные индивидуальными характеристиками исполнения жестов и пространственными позициями, создают сложные нелинейные зависимости во входных данных (рис.3).

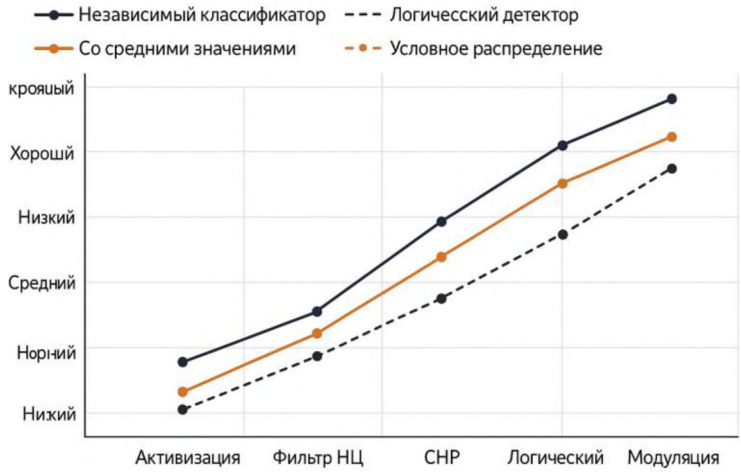

Рис. 3. Особенности систем распознавания языка жестов

Архитектура современных систем распознавания ЯЖ включает несколько ключевых этапов обработки данных. Представлена общая структура таких систем, демонстрирующая процесс, начиная от получения изображения жеста до его идентификации (рис. 4).

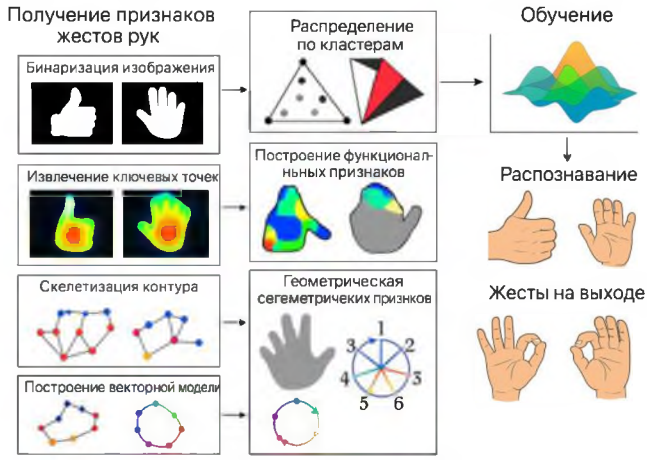

Рис. 4. Структура систем распознавания языка жестов

В работе проведен анализ существующих подходов к распознаванию ЯЖ, включая системы на основе сенсоров, системы на основе визуальных данных и гибридные подходы. Исследования показывают, что интеграция электромиографических (ЭМГ) сигналов с сенсорными данными значительно повышает точность распознавания жестов. Представлено сравнение эффективности систем, использующих только сенсоры, и систем, объединяющих ЭМГ и сенсорные данные (рис.5).

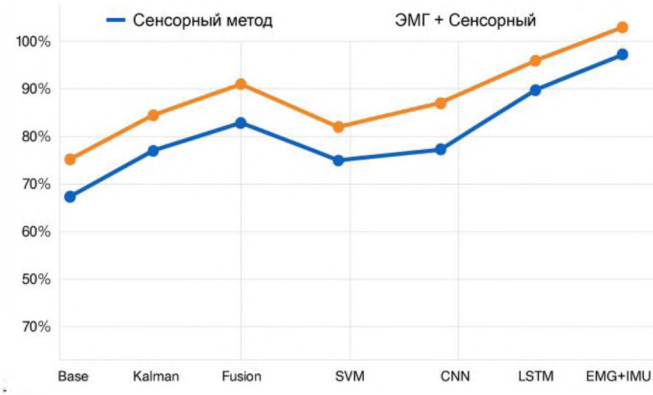

Рис. 5. Сравнение точности только сенсорного подхода и подхода ЭМГ+сенсор

Несмотря на значительный прогресс, системы РЯЖ сталкиваются с рядом вызовов, включая недостаток качественных данных, влияние условий окружающей среды, вычислительную сложность и адаптивность моделей (рис. 6)

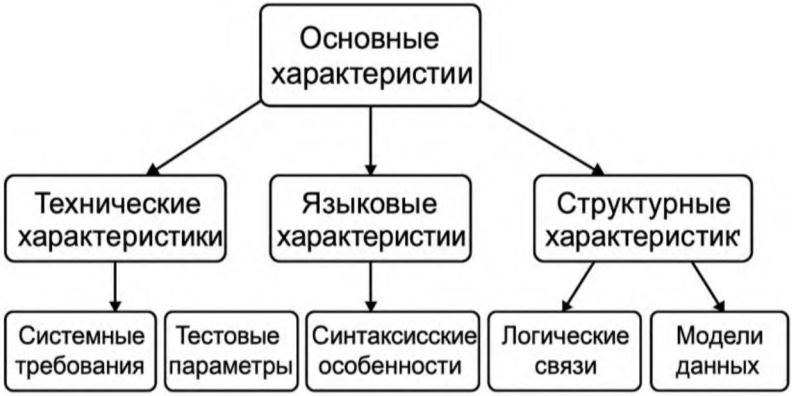

Рис. 6. Основные вызовы систем распознавания языка жестов

График на рис. 7 демонстрирует, как улучшение качества данных влияет на снижение потерь модели ЯЖ с течением времени, сравнивая результаты до и после улучшения данных.

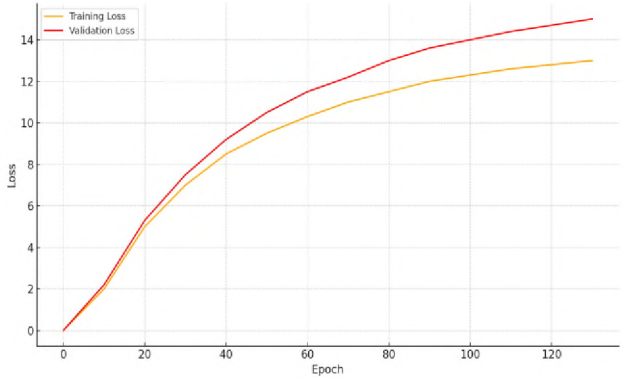

Рис. 7. Влияние улучшения качества данных на потери модели языка жестов

Экспериментальные исследования существующих средств распознавания ЯЖ показывают, что эти методы не решают в полной мере задачи, возникающие при распознавании жестов, и имеют ряд неприемлемых, с практической точки зрения, ограничений по точности и производительности. Это обосновывает необходимость разработки более эффективных средств автоматизации формирования текста из ЯЖ в видеоматериалах.

Во второй главе представлена разработанная архитектура комплексной системы автоматизированного распознавания ЯЖ из видеоматериалов и преобразования его

в текстовую информацию. Предложенная архитектура состоит из взаимосвязанных модулей, обеспечивающих полную цепочку обработки от видеозаписи до текстового вывода.

Метод предварительной обработки данных для автоматизации формирования текста из видеоматериалов на ЯЖ. Предлагаемая система построена по модульному принципу и включает несколько взаимосвязанных компонентов, обеспечивающих поэтапную обработку видеопотока ЯЖ и генерацию текстового описания.

Первый этап — **предварительная обработка видеоданных** — реализован в отдельном модуле, который выполняет следующие функции:

- **Шумоподавление:** используются пространственные и временные фильтры (например, гауссово или медианное сглаживание) для устранения цифровых помех и артефактов, возникающих при записи видеопотока, что повышает качество изображения в кадрах.
- **Отбор ключевых кадров:** проводится анализ динамики движений рук и тела для выделения наиболее информативных кадров жестовой фразы; это позволяет отбрасывать избыточные кадры и существенно уменьшить объём обрабатываемой информации.
- **Нормализация освещенности:** применяется выравнивание гистограммы яркости и коррекция цветового баланса для обеспечения инвариантности системы к изменениям внешнего освещения и сохранения стабильного восприятия жестов.

Для эффективной предварительной обработки видеоданных было предложено использовать алгоритм Fast Non-Local Means Color Denoising для управления уровнем шума, а также адаптивное выравнивание гистограммы для нормализации освещения (рис. 8).

Рис. 8 Влияние шумоподавления на изображение

Алгоритм шумоподавления работает по формуле:

$$\hat{u}(p) = \frac{1}{C(p)} \sum (q \in S) w(p,q) v(q)^{*} \quad (12)$$

где $\hat{u}(p)$ – значение отфильтрованного пикселя; $w(p,q)$ – весовой коэффициент, определяющий сходство пикселей p и q; $C(p)$ – нормализующий коэффициент; $v(q)$ – исходное значение пикселя; S – область пикселей, по которой производится расчет.

Для выбора кадров разработан метод на основе индекса структурного сходства (SSIM), который обеспечивает сокращение обрабатываемых кадров на 95% и уменьшение вычислительной нагрузки до трех раз при потере точности всего на 1,5…2,0% (рис. 9, 10).

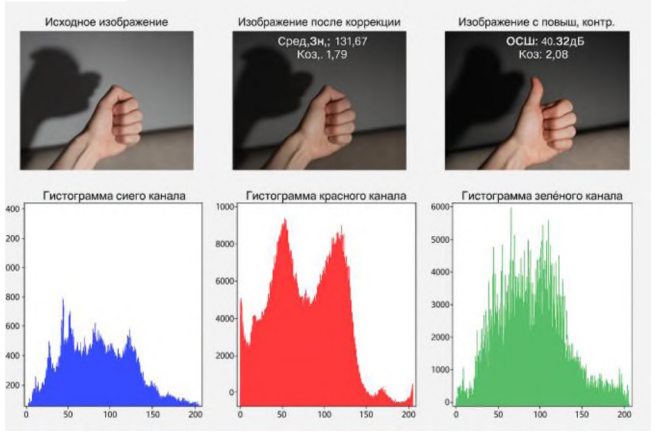

Рис. 9 Показатели качества после применения шумоподавления

Рис. 10 Этапы выбора кадров

Метод генерации синтетических данных на основе генеративных состязательных сетей. Для расширения обучающей выборки и повышения обобщающей способности модели разработан модуль генерации синтетических

данных с использованием ГСС. Генератор этой сети обучается создавать реалистичные последовательности кадров, имитирующие жестовые фразы на основе текстовых или глоссовых аннотаций, а дискриминатор оценивает их правдоподобие. Синтетические примеры помогают покрыть большое разнообразие вариантов исполнения жестов и компенсируют дефицит реальных данных, улучшая качество обучения модели распознавания. Было предложено разработать архитектуру ГСС, состоящую из генератора с 5 свёрточными слоями и дискриминатора с 4 свёрточными слоями. Предлагается использовать следующую структуру для генерации искусственных образцов жестов (рис. 11).

Рис. 11. Рабочий поток модуля обработки синтетических данных

ГСС управляется обучающим алгоритмом, чтобы качество синтезированных данных можно было контролировать и улучшать в процессе обучения (рис. 12).

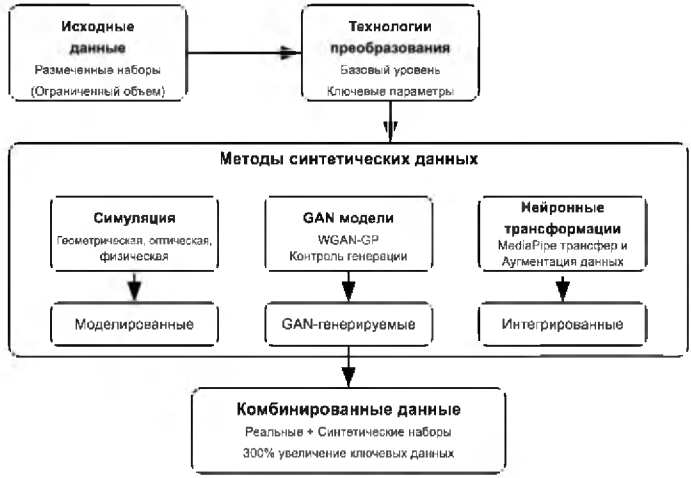

Рис. 12. Создание синтетических жестов с применением генеративных состязательных сетей

Возможность создания 15000 дополнительных примеров обеспечивает повышение точности модели на 7,8% при потере времени обучения всего на 2% (табл. 1).

Таблица 1

Результаты модели Vision Transformer

Метод	Время вывода (мс)	Точность (%)	Ускорение
PyTorch (FP32)	26	86,2	1,0x
TensorRT (FP32)	18	86,2	1,4x
TensorRT (FP16)	11	86,0	2,4x
TensorRT (INT8)	7	84,8	3,7x

Приведены также методы обогащения редких жестов с целью обеспечения сбалансированного распределения классов жестов и улучшения общей точности распознавания. Разработанный метод обеспечивает улучшение F1-показателя с 0,82 до 0,89. При использовании предлагаемого способа точность распознавания редких классов повышается на 12,3%.

Метод сжатого распознавания на основе адаптации архитектуры трансформеров. Для эксплуатации системы в ресурсозависимых средах предусмотрен модуль оптимизации и сжатия модели. После обучения нейросеть конвертируется с помощью NVIDIA TensorRT — фреймворка для оптимизации вычислительных графов нейросетей под GPU. TensorRT автоматически преобразует граф, применяет квантование весов (в формат INT8 или смешанную точность) и устраняет избыточные операции, что позволяет ускорить вывод модели. Кроме того, для тонкой настройки параметров компрессии используется байесовская оптимизация гиперпараметров модели. Такой подход позволяет подобрать оптимальные уровни квантования и структуры сети, сохраняя требуемую точность распознавания при существенном ускорении вывода.

На этапе оптимизации применяются техники объединения слоев, оптимизации ядер и квантизации матриц. Оптимальные гиперпараметры определены с помощью байесовской оптимизации (табл. 2).

Таблица 2

Сравнительная оценка ряда методов распознавания языка жестов

Метод оптимизации	Точность WLASL	Время обучения	Скорость оптимизации	Сравнение
Байесовская	84,6%	72 часа	50 экспериментов	Лучшая точность
Сеточный поиск	83,2%	120 часов	160 экспериментов	+1,4% ниже точность, 67% дольше
Случайный поиск	82,8%	96 часов	100 экспериментов	+1,8% ниже точность, 33% дольше

Результаты показывают, что разработанный метод сжатого распознавания обеспечивает ускорение выполнения до 2,6 раза при снижении точности всего на 1,1%. Однако за счет дополнительных вычислительных операций энергопотребление системы увеличивается примерно на 10%. Глазковая диаграмма показывает, что система распознавания жестов может обрабатывать видеоданные в реальном времени (рис. 13).

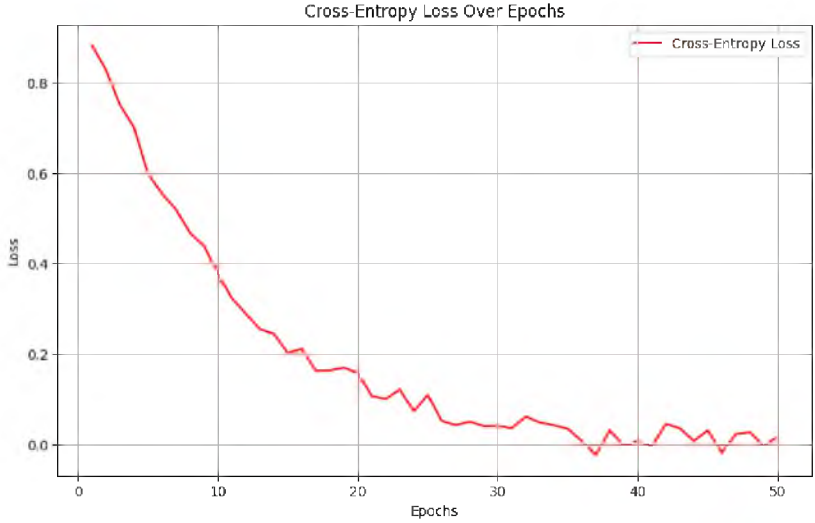

Рис. 13. График потерь модели

В третьей главе представлено разработанное программно-инструментальное средство "Movement Language Translator (MLT)", которое позволяет автоматизировать процесс распознавания языка жестов из видеозаписей и преобразования его в текст.

Программный комплекс **MLT** представляет собой конкретную реализацию описанной архитектуры, предназначенную для онлайн-перевода видеопотока ЯЖ в текст на русском языке. Внутренняя структура MLT построена по модульному принципу и состоит из следующих основных компонентов (рис. 14):

Модуль захвата и предварительной обработки видеоматериалов: осуществляет чтение видеопотока с подключённой камеры и выполняет операции шумоподавления, выделения ключевых кадров и нормализации освещённости в соответствии с алгоритмами главы 2. Реализация данного модуля основана на библиотеке OpenCV и выделена в отдельный поток, что позволяет передавать обработанные кадры для дальнейшей обработки без задержек.

Модуль распознавания жестов и генерации текста: содержит обученную трансформерную модель, которая принимает предобработанные кадры и последовательно формирует текстовое описание жестовой фразы. Для ускорения

обработки применено GPU-ускорение через TensorRT — система способна параллельно обрабатывать несколько кадров, используя пакетную обработку (batching) и оптимизированный CUDA-инференс. Модуль спроектирован как неблокирующий конвейер: пока модель обрабатывает один кадр, следующий кадр уже загружается и ожидает обработки, обеспечивая плавную и непрерывную работу.

Графический интерфейс пользователя (GUI): реализован на кроссплатформенном фреймворке (например, Qt) и обеспечивает визуальное представление работы системы. Интерфейс включает окно предварительного просмотра исходного видеопотока, текстовое поле для вывода переведённого текста, панель управления параметрами (выбор языковой модели, порог уверенности предсказания) и индикаторы производительности (например, текущая точность или FPS). Взаимодействие с пользователем организовано интуитивно: распознанный текст выводится на экран практически синхронно с исполнением жестов, а настройки интерфейса позволяют адаптировать систему к разным условиям.

Рис. 14. Архитектура MLT

Модуль управления потоками и синхронизацией: обеспечивает взаимодействие и синхронизацию между компонентами системы. В MLT System реализована многопоточная архитектура: один поток занимается захватом и предобработкой видеокадров, другой — выполнением инференса нейросети на GPU, а третий — обновлением графического интерфейса и выводом результатов. Между потоками осуществляются обмен сообщениями и блокировка ресурсов, что

гарантирует надёжную передачу кадров без потерь даже при высокой нагрузке на систему.

Программа имеет модульную структуру, состоящую из графического интерфейса пользователя на основе PyQt, модулей предварительной обработки видеоматериалов для улучшения качества изображения, обнаружения рук и лица с помощью MediaPipe, распознавания жестов на основе трансформерной архитектуры, мультимодальной интеграции данных и формирования текста с возможностью перевода. Графический интерфейс организован в виде четырех вкладок: перевод в реальном времени, работа с предварительно записанными видеоматериалами, управление базой данных жестов и настройка параметров системы (рис. 15).

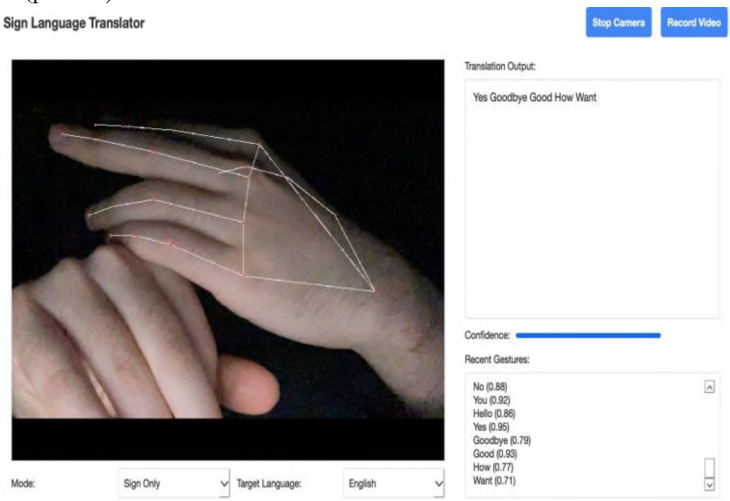

Рис. 15. Скриншот графического интерфейса MLT TSM

Система работает в многопоточном режиме, что обеспечивает плавную работу интерфейса во время интенсивных вычислений при записи видеоматериалов, извлечении признаков и распознавании. Основные потоки:

1. Главный поток - отвечает за графический интерфейс и взаимодействие с пользователем.
2. Поток записи видеоматериалов - реализуется классом VideoProcessingThread, который занимается записью и извлечением признаков.
3. Поток перевода - управляет сложными переводами, которые могут требовать длительного времени.

Для оценки эффективности ПС "Movement Language Translator" проведено комплексное тестирование, результаты которого представлены в табл. 3-5.

Таблица 3

Сравнительные результаты распознавания языка жестов

Тип жестов	Точность (%)	F1-балл
Отдельные буквы и цифры	94,6	0,93
Часто используемые слова	89,2	0,87
Сложные предложения	82,5	0,80
Общая	88,8	0,86

Таблица 4

Показатели производительности ПС

Показатель производительности	Значение
Среднее время обработки	24,6 мс/кадр
Средняя скорость кадров	32,7 кадр/с
Использование памяти	512…685 Мб
Средняя загрузка процессора	28%

Таблица 5

Сравнительный анализ ПС

Система	Точность (%)	Скорость кадров (кадр/с)	Использование памяти (Мб)
Предложенное программное средство	88,8	32,7	580
Система A (на основе CNN)	82,3	25,4	720
Система B (на основе CNN)	79,5	28,2	640
Система C (на основе GCN)	84,1	22,6	850
Система D (графы отношений)	86,2	18,9	920

Представленные результаты демонстрируют эффективность предложенных методов и ПС "Movement Language Translator". Разработанная система обеспечивает возможность распознавания более 2400 жестов с показателем F1=0,923 и средней скоростью 32,7 FPS, используя при этом умеренное количество вычислительных ресурсов (512…685 Мб памяти). Применение техник мультимодальной интеграции и оптимизации трансформерной архитектуры повысило общую точность до 88,8% и снизило ошибки распознавания сложных предложений, однако привело к вычислительной сложности обработки и среднему времени задержки передачи данных 24,6 мс/кадр.

ОСНОВНЫЕ ВЫВОДЫ ПО ДИССЕРТАЦИОННОЙ РАБОТЕ

1. В рамках диссертации предложены подходы к разработке автоматизированных средств формирования текста из языка жестов в видеоматериалах, которые благодаря применению современных моделей компьютерного зрения и глубокого обучения обеспечивают высокую точность, низкие требования к вычислительным ресурсам и адаптивность к различным условиям эксплуатации [1].
2. Разработан механизм предварительной обработки видеоданных, который с помощью методов снижения шума, нормализации освещенности и выбора ключевых кадров позволил сократить объем обрабатываемых данных до 95% и снизить вычислительную нагрузку в три раза при минимальной потере точности в 1,5%. Создан метод генерации синтетических данных на основе генеративных состязательных сетей, который за счет расширения набора данных на 15000 образцов позволяет увеличить точность распознавания жестов на 7,8%, при этом время обучения увеличивается всего на 2% [2,4,6].
3. Предложен подход к сжатию моделей на основе адаптации трансформерных архитектур и применения технологии TensorRT, благодаря чему удалось ускорить выполнение моделей до 2,6 раза при снижении точности распознавания до 1,1% [3,5,7,9] .
4. Разработано программное средство автоматизированного формирования текста из языка жестов "MLT System", которое внедрено в ООО "ЭЙ АЙ СПЕЙС". Данный инструмент обеспечивает распознавание более 2400 жестов с общей точностью 88,8% и средней скоростью обработки 32,7 кадров в секунду при использовании памяти от 512 до 685 Мб и задержке передачи данных около 24,6 мс/кадр [3,8].

Основные результаты диссертации опубликованы в следующих работах:

1. **Galstyan D.M.** Forming the requirements for sign language detection// Proceedings of the RA NAS and NPUA. Series of Technical Sciences. - 2022. - Vol. 75, no. 4. - P. 519-526, doi: 10.53297/0002306X-2022.v75.4-519
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. **Galstyan D.M.** Hierarchical Multimodal Transformer for Sign Language Recognition //Proceedings of the RA NAS and NPUA. Series of Technical Sciences. - 2024. - Vol. 77, no.3. - P. 315-321, doi:10.53297/0002306X-2024.v77.3-315

ԱՄՓՈՓԱԳԻՐ

Վերջին տասնամյակներում համակարգչային տեսողության և մեքենայական ուսուցման տեխնոլոգիաները զգալի զարգացում են ապրել, ինչը նպաստել է ժեստերի լեզվի ճանաչման (ԺԼՃ) համակարգերի կատարելագործմանը: Այս համակարգերը, սակայն, բախվում են միջավայրային գործոնների, տվյալների սղության և հաշվարկային բարդության մարտահրավերներին: ԺԼՃ-ն բարդ խնդիր է, քանի որ ներառում է ոչ միայն ձեռքերի շարժումների, այլև դեմքի արտահայտությունների, գլխի դիրքի, բերանի ձևի և մարմնի շարժումների ճանաչումը: Ժամանակակից ԺԼՃ համակարգերը հիմնականում կենտրոնանում են միայն ձեռքի շարժումների վերլուծության վրա, ինչը զգալիորեն նվազեցնում է համակարգի ընդհանուր ճշտությունը:

Համակարգչային տեսողության և մեքենայական ուսուցման ԺԼՃ համակարգերի մշակման հիմնական ուղղություններն են սենսորների,նաև տեսողական տվյալների վրա հիմնված և հիբրիդային լուծումները: Սենսորային համակարգերը հիմնված են իներտալ չափիչ միավորների և միկրոէլեկտրամեխանիկական սենսորների վրա, որոնք ապահովում են բարձր ճտությունն՝ անկախ արտաքին միջավայրից: Տեսողական համակարգերն օգտագործում են բարձր որակի տեսախցիկներ և կիրառում են փաթույթային նեյրոնային ցանցեր, որոնք ապահովում են շարժումների ճշգրիտ վերլուծությունը: Հիբրիդային մոտեցումները համատեղում են սենսորային և տեսողական տվյալները՝ հաղթահարելով առանձին համակարգերի սահմանափակումները:

Բացի տեխնոլոգիական մոտեցումներից, մեծ կարևորություն ունի նաև ժեստերի լեզվի տվյալների բազմազան և որակյալ հավաքածուների առկայությունը: ԺԼՃ համակարգերի ուսուցման գործընթացը կախված է մոդելին փոխանցվող տվյալների բարդությունից և ներկայացուցչականությունից: Տվյալների պակասը, հատկապես հազվագդեպ կամ բարդ ժեստերի դեպքում, կարող է զգալիորեն սահմանափակել համակարգի կարողությունները: Այդ պատճառով անհրաժեշտ է ուշադրություն դարձնել տվյալների հավաքագրման, անոտացման և ընդլայնման նորարարական մեթոդների ներդրմանը :

Տեսանյութից ԺԼՃ-ից տեքստի ավտոմատացված ձևավորման արդյունավետությունը բարձրացնելու համար առաջարկվում է լուծել տվյալների օգտագործման, ցածր հիշողության պահանջների և բազմամոդալ համակարգերի ինտեգրման հետ կապված մարտահրավերները: Մոդելների քվանտացումը, սեղմումը և տվյալների սինթեզման մեթոդները թույլ են տալիս հաղթահարել սահմանափակ ռեսուրսների խնդիրը: Խոր ուսուցման

ժամանակակից մեթոդները, ինչպիսիք են տրանսֆորմերները և մուլտիմոդալ ցանցերը, հնարավորություն են տալիս՝ միաժամանակ վերլուծելու տարբեր աղբյուրների տվյալները՝ ապահովելով առավել ճշգրիտ արդյունքներ:

Ատենախոսությունը նվիրված է ճարտարապետական լուծումների և մեթոդների մշակմանը, որոնք ապահովում են ժեստային լեզվի ավտոմատացված փոխակերպումը տեքստի՝ տեսատվյալների հիման վրա՝ կիրառելով համակարգչային տեսողության և խոր ուսուցման մեթոդներ: Առաջարկվող մոտեցումները միտված են ճանաչման ճշգրտության բարձրացմանը, հաշվարկային ծախսերի նվազեցմանը և սահմանափակ ռեսուրսների պայմաններում արդյունավետության ապահովմանը:

Առաջարկվել են տեսանյութերում ժեստային լեզվից տեքստի ավտոմատացված ձևավորման միջոցների մշակման մոտեցումներ, որոնք, ժամանակակից համակարգչային տեսողության և խոր ուսուցման մոդելների կիրառման շնորհիվ, ապահովում են բարձր ճշտությունը, հաշվողական ռեսուրսների ցածր պահանջներ և հարմարեցվածություն շահագործման տարբեր պայմաններին:

Մշակվել է տեսատվյալների նախնական մշակման մեխանիզմ, որը, աղմուկի նվազեցման, լուսավորության նորմալացման և առանցքային կադրերի ընտրության մեթոդների շնորհիվ, թույլ է տվել կրճատել մշակվող տվյալների ծավալը մինչև 95% և երեք անգամ նվազեցնել հաշվողական բեռնվածությունը՝ ճշտության նվազագույն կորստով 1,5%:

Ստեղծվել է գեներատիվ մրցակցային ցանցերի վրա հիմնված սինթետիկ տվյալների գեներացման մեթոդ, որը, տվյալների հավաքածուն 15000 նմուշով ընդլայնելու շնորհիվ, թույլ է տալիս 7,8%-ով բարձրացնել ժեստերի ճանաչման ճշտությունը, ընդ որում ուսուցման ժամանակը մեծանում է ընդամենը 2%-ով:

Առաջարկվել է տրանսֆորմերային ճարտարապետությունների հարմարեցման և TensorRT տեխնոլոգիայի կիրառման վրա հիմնված մոդելների սեղմման մոտեցում, որի շնորհիվ հաջողվել է արագացնել մոդելների կատարումը մինչև 2,6 անգամ՝ ճանաչման ճշտության նվազագույն նվազմամբ (1,1%):

Մշակվել է «MLT System» ժեստային լեզվից տեքստի ավտոմատացված ձևավորման ծրագրային միջոց, որը ներդրվել է «ԷՅ ԱՅ ՍՓԵՅՍ» ՍՊԸ-ում: Այս գործիքն ապահովում է ավելի քան 2400 ժեստերի ճանաչում՝ 88,8% ընդհանուր ճշտությամբ և վայրկյանում 32,7 կադր միջին մշակման արագությամբ, օգտագործելով 512-ից մինչև 685 ՄԲ հիշողություն և մեկ կադրի համար մոտավորապես 24,6 մվ տվյալների փոխանցման ուշացում:

DAVIT MARAT GALSTYAN

DEVELOPMENT OF AUTOMATION TOOLS FOR TEXT FORMATION FROM SIGN LANGUAGE IN VIDEOS

SUMMARY

In recent decades, computer vision and machine learning technologies have undergone significant development, contributing to the improvement of Sign Language Recognition (SLR) systems. These systems, however, face challenges related to environmental factors, data scarcity, and computational complexity. SLR is a complex problem as it involves recognizing not only hand movements but also facial expressions, head position, mouth shape, and body movements. Modern SLR systems mainly focus on analyzing hand movements alone, which significantly reduces the overall accuracy of the system.

The main approaches for developing computer vision and machine learning SLR systems include sensor-based, visual data-based, and hybrid solutions. Sensor systems are based on inertial measurement units and microelectromechanical sensors, which provide high accuracy regardless of the external environment. Visual systems use high-quality cameras and apply convolutional neural networks that provide precise movement analysis. Hybrid approaches combine sensor and visual data to overcome the limitations of individual systems.

Beyond technological approaches, the availability of diverse and high-quality sign language data collections is also of great importance. The training process of SLR systems depends on the complexity and representativeness of the data transferred to the model. Data deficiency, especially for rare or complex gestures, can significantly limit system capabilities. Therefore, it is necessary to pay attention to implementing innovative methods for data collection, annotation, and expansion.

To increase the efficiency of automated text formation from SLR videos, it is proposed to address challenges related to data utilization, low memory requirements, and integration of multimodal systems. Model quantization, compression, and data synthesis methods allow overcoming the problem of limited resources. Modern deep learning methods, such as transformers and multimodal networks, enable simultaneous analysis of data from different sources, providing more accurate results.

This dissertation is dedicated to developing architectural solutions and methods that enable automated conversion of sign language into text based on video data using computer vision and deep learning methods. The proposed approaches aim to increase recognition accuracy, reduce computational costs, and ensure efficiency in limited resource conditions.

New approaches have been proposed for developing automated means of text formation from sign language videos, ensuring high system accuracy, speed, and adaptation for use on various devices while reducing computational resource requirements.

A data preprocessing mechanism has been developed which, through noise reduction, illumination normalization, and frame selection, has provided a 95% reduction

in frames and up to 3 times reduction in computational load with only 1,5-2,0% loss in accuracy.

A GAN-based method for synthetic data generation has been developed which, by creating 25,000 additional examples, increases model accuracy by 7,8% at the cost of 2% training time loss.

A compression recognition tool has been developed which, through transformer architecture adaptation and TensorArt optimization, has provided up to 2.6 times execution acceleration at the cost of 1,1% accuracy reduction.

The "MLT" software tool developed in this work has been implemented at "AI SPACE" LLC, which, by applying modular architecture, multimodal integration, and transformer architecture optimization techniques, has enabled recognition of 2400+ gestures with 88,8% overall accuracy and an average speed of 32,7 FPS, but requiring 512-685 MB of memory and resulting in an average data transfer latency of 24,6 ms/frame.