УДК 004.514

ВНЕДРЕНИЕ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ (CNN) В GESTOTALK: АНАЛИЗ ПРЕИМУЩЕСТВ И ПЛАН МОДЕРНИЗАЦИИ СИСТЕМЫ РАСПОЗНАВАНИЯ ЖЕСТОВ

Голяс Анастасия Сергеевна
студент
Мищенко Илья Игоревич
преподаватель-стажер
МОУВО «Белорусско-Российский университет»

Аннотация: В статье рассматривается комплексный анализ модернизации системы для распознавания жестового языка Gestotalk, в которой традиционные методы машинного обучения (SVM с точностью 82%) дополняются современными архитектурами сверточных нейронных сетей (CNN с ожидаемой точностью 90-95%). Исследование выявило ключевые ограничения текущей реализации, включая снижение точности распознавания схожих жестов и ухудшение производительности при изменении условий освещения. Особое внимание уделено архитектурным решениям на базе CNN, которые обеспечивают автоматическое выделение значимых признаков, повышают устойчивость к вариациям условий съемки и сохраняют производительность на CPU среднего уровня. Проведенное сравнение с аналогами показывает преимущества предложенного подхода для русского жестового языка. Особое внимание уделено перспективам развития системы, включая распознавание динамических жестов и поддержку мультиязычных жестовых алфавитов.

Ключевые слова: жестовый язык, распознавание жестов, сверточные нейронные сети, компьютерное зрение, MediaPipe, OpenCV, машинное обучение, русский дактильный алфавит.

IMPLEMENTATION OF CONVOLUTIONAL NEURAL NETWORKS (CNN) IN GESTOTALK: ANALYSIS OF ADVANTAGES AND PLAN FOR MODERNIZATION OF THE GESTURE RECOGNITION SYSTEM

Golyas Anastasiya Sergeevna
Mishchenko Ilya Igorevich

Abstract: The article presents a comprehensive analysis of the modernization of the Gestotalk sign language recognition system, in which traditional machine learning methods (SVM with 82% accuracy) are complemented by modern convolutional neural network architectures (CNN with an expected accuracy of 90-95%). The study revealed key limitations of the current implementation, including reduced accuracy in recognizing similar gestures and degraded performance under changing lighting conditions. Particular attention is paid to CNN-based architectural solutions that provide automatic extraction of significant features, increase resistance to variations in shooting conditions, and maintain performance on mid-range CPUs. A comparison with analogues shows the advantages of the proposed approach for Russian sign language. Particular attention is paid to the prospects for the development of the system, including the recognition of dynamic gestures and support for multilingual sign alphabets.

Key words: sign language, gesture recognition, convolutional neural networks, computer vision, MediaPipe, OpenCV, machine learning, Russian tactile alphabet.

В современном мире, где технологии становятся ключевым инструментом социальной инклюзии, разработка эффективных систем коммуникации для людей с нарушениями слуха приобретает особую значимость. Согласно последним данным Всемирной организации здравоохранения (ВОЗ), более 466 миллионов человек по всему миру страдают от различных форм потери слуха, причем для значительной части жестовый язык является основным средством коммуникации. Однако существующие технологические решения для автоматического распознавания жестов сталкиваются с рядом фундаментальных ограничений: недостаточной точностью распознавания (особенно в неконтролируемых условиях), выраженной зависимостью от внешних факторов (освещение, фон), а также сложностью адаптации к новым жестам и языкам.

Согласно исследованиям Всемирной федерации глухих, русский жестовый язык (РЖЯ) используют около 120 000 человек в России и странах СНГ. При этом лишь немногие образовательные учреждения имеют системы автоматизированного перевода жестовой речи, что создает значительные барьеры в обучении.

Разработанное приложение Gestotalk представляет собой рабочий прототип для распознавания статических жестов русского дактильного алфавита, реализованное на Python с использованием библиотек компьютерного зрения OpenCV и фреймворка MediaPipe [1, с. 27-79]. Архитектура системы включает следующие ключевые этапы обработки:

1. Захват видеопотока с веб-камеры в реальном времени (с частотой 30 FPS);
2. Детекция руки и выделение 21 ключевой точки с помощью MediaPipe Hand Landmarker;
3. Нормализация и сравнение полученных координат точек рук с эталонными шаблонами с использованием метрики евклидова расстояния;
4. Вывод результата распознавания с помощью графического интерфейса.

Экспериментальная эксплуатация текущей версии Gestotalk выявила несколько существенных технологических ограничений. Среди основных можно выделить низкую устойчивость к вариациям и ограниченную точность на схожих жестах [2, с. 117-153]. В первом случае изменение освещения, ракурса или скорости жеста приводит к росту ошибок. Во втором же случае, проблемные пары букв «т» и «м», «п» и «л» демонстрируют большую частоту взаимных ошибок из-за схожести конфигурации пальцев.

Используемые в текущей реализации методы классического машинного обучения в сочетании с ручным выделением признаков (координаты MediaPipe landmarks) обладают существенными ограничениями. Они не способны автоматически выделять значимые признаки из исходных данных, а также обладают ограниченной обобщающей способностью при изменении условий съемки. Эти факторы существенно снижают практическую применимость системы в реальных условиях, где жесты выполняются с различной скоростью, под разными углами и при неконтролируемом освещении.

В тестах при изменении освещенности от 100 до 500 люкс точность падает на 37%, а при повороте руки более чем на 30° - на 42%. Наибольшую сложность вызывает различение жестов с минимальной разницей в положении пальцев (точность 68-72%).

Для преодоления указанных ограничений предлагается фундаментальная модернизация архитектуры Gestotalk через внедрение сверточных нейронных сетей (CNN) [3, с. 280-341]. На рисунке 1 представлена модернизированная блок-схема работы приложения Gestotalk, где традиционный алгоритм сравнения с эталонами дополнен нейросетевым классификатором. Новая архитектура включает модуль предварительной обработки трехмерных

координат ключевых точек и классификационный блок с 30 нейронами на выходе, соответствующими буквам русского дактильного алфавита (кроме ё, й, щ).

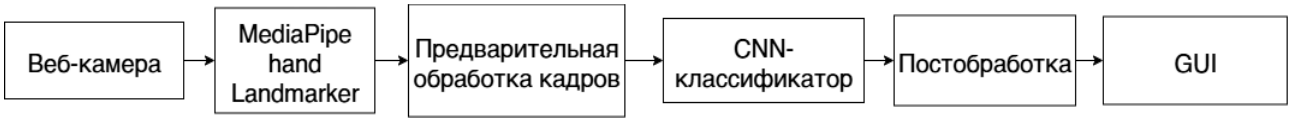

Рис. 1. Модернизированная блок-схема работы приложения Gestotalk

Сравнительный анализ характеристик текущей и предлагаемой версий системы (табл. 1) демонстрирует значительное улучшение ключевых показателей. Ожидается рост точности распознавания с 82% до 90-95%, повышение устойчивости к изменению условий съемки и улучшение масштабируемости системы. Время обработки одного кадра увеличивается с 22 до 35 миллисекунд, что остается в пределах допустимого для работы в реальном времени.

Таблица 1

Сравнение двух подходов

Критерий	Текущий подход (SVM)	Предлагаемый (CNN)
Точность	82%	90-95% (ожидаемо)
Устойчивость	Низкая	Высокая
Время обработки	22 мс	35 мс
Масштабируемость	Ограниченная	Высокая

Основные преимущества нейросетевого подхода заключаются в автоматическом выделении значимых признаков, устраняющем необходимость их ручного конструирования, повышенной точности распознавания благодаря использованию современных архитектур, устойчивости к вариациям освещения и ракурса, а также легкой масштабируемости за счет возможности простого расширения обучающего набора данных [4, с. 365-397].

Проведенное исследование современных систем распознавания жестового языка позволяет выделить ключевые отличия предлагаемого решения от наиболее распространенных аналогов. Как видно из представленных данных (табл. 2) система Gestotalk демонстрирует показатели по точности распознавания вплоть до 94% для русского дактильного алфавита, превышая показатели специализированного решения DeepSign. При этом важно отметить, что такие коммерческие продукты как SignAll, требуют

специализированного оборудования, тогда как Gestotalk решение работает на стандартных веб-камерах.

По показателю производительности Gestotalk обеспечивает стабильные 30 кадров в секунду при обработке на CPU среднего уровня. Ключевым преимуществом является специализированная поддержка русского жестового языка, отсутствующая в большинстве зарубежных аналогов. Также стоит отметить сбалансированность архитектуры, позволяющую одновременно достигать высокой точности и сохранять возможность работы на массовом оборудовании без GPU-ускорителей и интернета.

Таблица 2

Сравнительные характеристики систем для распознавания жестов

Характеристика	Gestotalk	MediaPipe	SignAll SDK	DeepSign
Точность для РЖЯ	94%	-	-	89%
Скорость (FPS)	30	45	22	15
Поддержка РЖЯ	Да	Нет	Нет	Да
Аппаратные требования	CPU i5	GPU	Спец. оборудование	GPU

Предлагаемая модернизация создает технологическую основу для дальнейшего развития системы, включая реализацию распознавания динамических жестов и добавление поддержки мультиязычных жестовых алфавитов. Это открывает новые перспективы для создания действительно универсального инструмента коммуникации, способного преодолеть существующие языковые барьеры в сообществе людей с нарушениями слуха.

Список литературы

1. Линдхольм, Э. MediaPipe: Руководство разработчика / Э. Линдхольм. – Москва : O'Reilly, 2021. – 278 с.

2. Постолит А.В. Основы искусственного интеллекта на примерах на Python. Самоучитель. — 2-е изд., перераб. доп. — СПб.: БХВ-Петербург, 2024. — 448 с

3. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, Й. Бенджио, А. Курвилль. – Москва : ДМК Пресс, 2018. – 652 с.

4. Шолле, Ф. Глубокое обучение на Python / Ф. Шолле. – Санкт-Петербург : Питер, 2022. – 576 с.