УДК 004.93'12

МЕТОД РАСПОЗНАВАНИЯ РУССКОЙ ЖЕСТОВОЙ РЕЧИ ДЛЯ СИСТЕМ АВТОМАТИЧЕСКОГО ПЕРЕВОДА В РЕАЛЬНОМ ВРЕМЕНИ

Костюк А. Д.
Научный руководитель – кандидат технических наук, доцент Штенников Д. Г.
Университет ИТМО
od@itmo.ru

Работа выполнена в рамках темы НИР «Методика для автоматизации распознавания и оцифровки русской жестовой речи в реальном времени».

Введение

Общее количество людей с частичной или полной потерей слуха превышает 430 млн человек [1], поэтому особую актуальность приобретает задача автоматического распознавания жестовой речи, позволяющая обеспечивать коммуникацию с людьми, имеющими нарушения слуха. В распознавании жестов обычно выделяют два наиболее распространённых подхода: анализ плотных видеоданных и использование скелетных представлений ключевых точек позы тела и рук [2]. Видео-ориентированные методы демонстрируют высокую выразительность, но являются требовательными к вычислительным ресурсам и плохо масштабируются [3]. Подходы, основанные на скелетных представлениях, позволяют снизить размерность данных и являются более устойчивыми к изменениям фона и освещения [4]. Существует множество решений, ориентированных на иностранные жестовые языки и ограниченное число классов. В то же время направление распознавания русской жестовой речи с большим количеством жестов в режиме реального времени остаётся недостаточно исследованным.

Основная часть

В работе предложен метод распознавания жестов русской жестовой речи, основанный на использовании скелетного представления и адаптации свёрточных нейронных сетей к обработке пространственно-временных данных. Входными данными выступают координаты ключевых точек рук и позы тела, извлекаемые из видеопоследовательности с применением предобученной модели MediaPipe Holistic. Каждый жест представляется в виде временной последовательности фиксированной длины, что позволяет обеспечить сопоставимость примеров при обучении модели. Ключевой особенностью данного подхода является преобразование последовательностей координат в двумерное представление, интерпретируемое как псевдоизображение с тремя каналами, соответствующими пространственным координатам ключевых точек. Такое представление позволяет использовать современные архитектуры компьютерного зрения для анализа динамических жестов без использования тяжеловесных рекуррентных или трансформерных моделей. В предлагаемом методе для классификации используется предобученная на ImageNet архитектура EfficientNet, адаптированная к задаче распознавания жестов, что позволяет эффективно извлекать информативные признаки из скелетного представления и сохранять вычислительную эффективность, необходимую для применения в системах реального времени.

Выводы

Предложенный подход обеспечивает распознавание 1001 жеста русской жестовой речи с точностью более 80 % на тестовой выборке, сформированной из объединённого корпуса датасетов SLOVO и BUKVA общим объёмом 23 944 видеозаписей. Результаты проведённого исследования подтверждают возможность масштабируемой

классификации большого числа жестов при сохранении вычислительной эффективности, достаточной для применения в режиме реального времени. Разработанный алгоритм может быть реализован в виде программного модуля распознавания и интегрирован в системы автоматического перевода РЖЯ, образовательные платформы и сервисы поддержки коммуникации.

Литература

1. World Health Organization et al. World report on hearing. – World Health Organization, 2021.
2. Wang C., Yan J. A comprehensive survey of rgb-based and skeleton-based human action recognition //IEEE Access. – 2023. – Т. 11. – С. 53880-53898.
3. Huang J., Chouvatut V. Video-based sign language recognition via resnet and lstm network //Journal of Imaging. – 2024. – Т. 10. – №. 6. – С. 149.
4. Li C. et al. Skeleton-based gesture recognition using several fully connected layers with path signature features and temporal transformer module //Proceedings of the AAAI conference on artificial intelligence. – 2019. – Т. 33. – №. 01. – С. 8585-8593.