УДК 004.934: 004.588

ВИЗУАЛЬНОЕ РАСПОЗНАНИЕ РЕЧИ

Д.А. МАКАР

Белорусский государственный университет информатики и радиоэлектроники (г. Минск, Беларусь)

Аннотация. С развитием технологий автоматического распознания речи, развивается интеграция визуальных компонентов коммуникации. В данной работе представлен принцип построения программного обеспечения, направленного на создание системы визуального распознания жестикуляции и артикуляции речевого аппарата. Исследование предполагает применение машинного обучения для анализа произносимых слов, основанное на методах компьютерного зрения, включая систему захвата данных, обработки и интеграции, а также воспроизведение речи с применением технологий синтеза.

Ключевые слова: визуальное распознание речи, автоматическое распознание речи, машинное обучение, нейронные сети, синтез речи.

VISUAL SPEECH RECOGNITION

D.A. MAKAR

Belarusian State University of Informatics and Radioelectronics (Minsk, Belarus)

Abstract. With the development of automatic speech recognition technologies, the integration of visual components in communication is also evolving. This paper presents the principles for developing software aimed at creating a system for visually recognizing gestures and the articulation of the speech apparatus. The study involves the application of machine learning for analyzing spoken worlds, based on computer vision methods, including data capture, processing, and integration systems, as well as speech reproduction using synthesis technologies.

Keywords: visual speech recognition, automatic speech recognition, machine learning, neural networks, speech synthesis.

Введение

С развитием технологий автоматического распознания речи (ASR), развивается интеграция визуальных компонентов коммуникации.

Цель программного обеспечения – создание системы, которая будет учитывать только визуальные компоненты.

Задачи:

1. Исследование артикуляции и жестикуляции речевого аппарата и интеграция в процесс распознавания;
2. Разработка модуля для захвата и анализа мимики и микромимики;
3. Реализация алгоритмов для восприятия и воспроизведения речи на основе визуальных данных.

Методика

Для системы захвата данных, необходимо использовать камеры для визуального распознания речи. Данная система сочетает в себе элементы компьютерного зрения и обработку языка.

Принцип архитектуры:

Камеры для захвата данных (лица и артикуляции), для получения 3D-данных о мимике и микромимики (например, Microsoft Kinect). Камеры устанавливаются таким образом чтобы обеспечить полноту фиксации изображения (крупный план лица) для максимального распознания мимики и движения губ. Система непрерывно или по запросу считывает при определенных триггерах (например, при обнаружении движения губ).

Для обработки визуальных данных потребуется компьютерное зрение, а для анализа движения лица такие библиотеки, как OpenCV и Mediapipe, что поможет извлечь ключевые точки на лице [1,2]. Пример принципа работы Mediapipe на рис 1. Благодаря алгоритмам

компьютерного зрения (например, Haar Cascades или HOG), система определяет и отслеживает лицо на видео или в режиме реального времени.

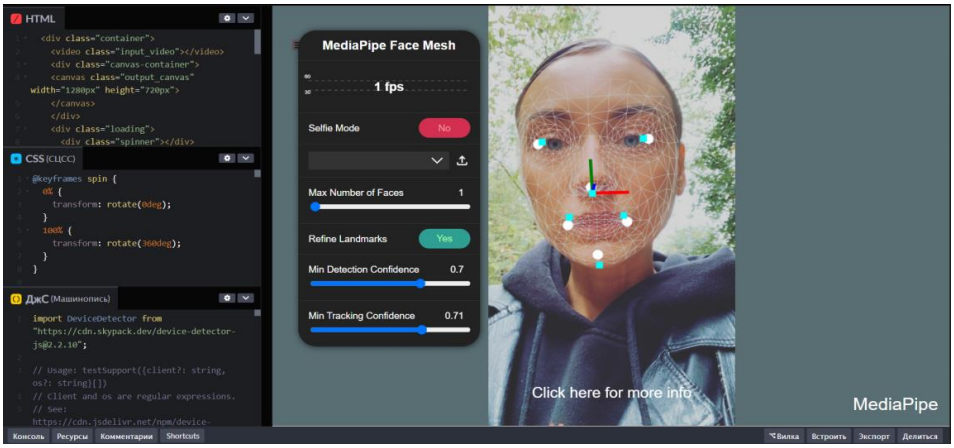

Рис. 1. Пример принципа работы Mediapipe

Используя методы обработки изображений, система сосредотачивается на области рта и анализирует движение губ. Для фильтрации применимы выделение контуров и использование машинного обучения для классификации звуков по движению губ.

Модели машинного обучения: обучение нейросети (например, CNN и RNN), для анализа временных рядов и выявлении взаимосвязи произносимых слов [3].

Моделирование артикуляции: создание модели для предоставления движения артикуляторов (языка, губ).

Реализация модуля воспроизведения речи: технологии синтеза речи (text-to-speech, TTS), для генерации звуковой дорожки, стогованной с полученными визуальными данными.

Языки программирования: Python – для прототипа и Deep Learning, C++ – для высокопроизводительных вычислений в реальном времени.

Библиотека и фреймворки: TensorFlow или PyTorch – для моделирования нейросети, OpenCV –для обработки изображений и Mediapipe – для захвата ключевых точек.

Интерфейс: Flask или Django – для взаимодействия пользователя с системой.

Визуальные паттерны звуковых сигналов сопоставляются с текстом. Для повышения уровня точности, можно использовать аудиозапись параллельно с видеопотоком, что позволит алгоритмам учитывать и визуальные, и звуковые данные.

Результат

Визуальное распознание речи, основанное на артикуляции и жестикуляции речевого аппарата, представляет собой многообещающую область – это сложная задача, требующая междисциплинарного подхода. Успешная реализация возможна при наличии качественных данных для обучения, а также подходящих алгоритмов обработки изображения и языка.

Заключение

Применение данного метода многогранно и не ограничивается областью людей с частичной или полноценной потерей речи, для более адаптивной коммуникации в социуме, но и в системе государственной безопасности, профессиональной деятельности и иной сфере.

Список литературы

1. Менуа Геворгян, OpenCV 4 с Python Blueprints: творческие проекты компьютерного зрения Bluid с последней версией OpenCV и Python 3, 2-е здание / Мю Геворгян, А. Мамиконян, М. Бейелер // Packt Publising Ltd, 2020.
2. Фреймвор MediaPipe [Электронный ресурс] // Google Al for Developers. Режим доступа: https://ai.google.dev/edge/mediapipe/framework. Дата обращения: 05.11.2024.
3. Мишель А. Нильсен, Нейронные сети и глубокое обучение / М.А. Нильсен // Determination Press, 2015.