РЕФЕРАТ

Отчет 16 с., 1 кн., 3 рис., 3 источн.

ЖЕСТОВЫЙ ЯЗЫК, СУРДОПЕРЕВОД, СИНТЕЗ ГОЛОСА, ПЕРЕВОДЧИК, СИСТЕМА ПОМОЩИ В ПОДДРЕЖКЕ ДИАЛОГА

Цель работы – разработка устройства синхронного перевода слов жестового языка в звуковой формат.

В процессе работы проводились экспериментальные и теоретические исследования нейронных сетей распознавания и перевода жестовых слов и синтезаторов речи.

В результате исследования была отобрана нейронная сеть MediaPipe Hands.

Основные показатели: низкие требования к камере, поддержка жестов двух рук и ориентирование на безмаркерную систему. В результате исследования был отобран синтезатор голоса Google Text-To-Speech API.

Основные показатели: естественное звучание синтезированного голоса, возможность изменения тембра и скорости голоса под требования потребителя, возможность создания голоса для синтеза.

Степень внедрения – тестовый образец работы нейронной сети и синтеза голоса.

Эффективность работы систем определяется точностью и естественностью работы итогового программного продукта. Обе системы могут быть интегрированы в один программный продукт для выполнения функции перевода жестовых слов в звуковой формат.

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящем отчете о НИР применяют следующие термины с соответствующими определениями:

Жестовый язык – самостоятельный язык, состоящий из жестов, каждый из которых производится руками в сочетании с мимикой, формой или движением рта и губ, а также в сочетании с положением корпуса тела

Сурдоперевод - процесс перевода устной речи на жестовый язык и в обратном порядке

Безмаркерная система – система, использующая для считывания жестов рук камеры или оптические датчики

Синтез речи - формирование речевого сигнала по печатному тексту с помощью программного продукта

Угол обзора — это та часть сцены перед камерой, которая попадает на матрицу и становится изображением

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

В настоящем отчете о НИР применяют следующие сокращения и обозначения:

МП – мегапиксель; 1 МП = 10^6 пикселей

Мб – мегабайт; 1 Мб = 2^{23} бит = 8388608 бит

лк – люкс, 1 лк = 1 люмен/$м^2$

ВВЕДЕНИЕ

В ходе первого этапа НИР проводилась экспериментальная и теоретическая исследовательская работа для определения оптимальной нейронной сети распознавания жестовых слов и программы синтеза голоса.

1. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ ОПТИМАЛЬНОГО ТИПА НЕЙРОННОЙ СЕТИ

В ходе исследовательской работы, целью которых было определение ключевых требований нейронной сети распознавания жестов кислей рук, было выявлено, что достаточными требованиями к нейронной сети можно считать следующие:

1. Нейронная сеть должна работать с безмаркерной системой, то есть считывать жест через камеру или оптический датчик.
2. Нейронная сеть должна показывать достаточный уровень распознавания жеста кисти руки как при усредненных, так и при отличных от усредненных условиях съемки, включая близкие к граничным. Под достаточным уровнем распознавания понимается такой процент считаных жестов, что их количество достаточно для сохранения смысла изначального сообщения;
3. Нейронная сеть должна работать с камерами или оптическими датчиками с низкими техническими показателями, тем самым, не имея критического влияния на итоговую себестоимость товара;
4. Нейронная сеть должна коррелировать свои выводы о значении жеста в соответствии с показаниями гироскопа, то есть распознавать не только жест кисти руки, но и положение кисти относительно камеры или оптического датчика;
5. Нейронная сеть должна считывать и распознавать жесты как минимум одной руки.

В ходе экспериментальных исследований существующих нейронных сетей распознавания жестов кистей рук было решено интегрировать в итоговый программный продукт нейронную сеть с открытым кодом MediaPipe Hands. Данная нейронная сеть соответствует заявленным в проекте требованиям, в последствии возможны изменения исходного кода в случае возникновения необходимости. Распознавание жестов происходит с помощью построения трехмерного скелета ладони, содержащего 21 точку наблюдения, через положение в пространстве которых определяется жест.

Ключевые особенности отобранной нейронной сети:

1. Для считывания жеста кисти руки данная нейронная сеть требует использования камеры;
2. Нейронная сеть имеет низкие требования к техническим характеристикам камеры, при этом диапазон условий, при которых считывание остается на достаточном уровне, соответствует требованиям, в которых устройство может быть использовано потребителем;

3. В нейронную сеть внедрено распознавание не только жеста кисти руки, но и положение кисти относительно камеры, что позволяет не использовать в итоговой конструкции гироскоп;
4. Нейронная сеть способна к распознаванию жестов как одной руки, так и двух одновременно. При этом механизм распознавания и требования к камере не изменяются.

2. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ ОПТИМАЛЬНОГО ТИПА ОПТИЧЕСКИХ ДАТЧИКОВ И КАМЕР

В ходе экспериментальных исследований заранее отобранные камеры с различным фокусным расстоянием и качеством съемки были поочередно помещены в условия, имитирующие те, в которых потребитель может использовать устройство: темное время суток/помещение, солнечная ппогода, туман, загрязнение камеры.

Использованные камеры:

1. 13 МП, фокусное расстояние 5 мм;
2. 5 МП, фокусное расстояние 20 мм;
3. 1 МП, фокусное расстояние 35 мм.

По итогу исследования было выявлено, что для работы достаточно использования камеры с разрешением не менее 1 МП и углом обзора не менее 100° (фокусное расстояние не менее 20 мм). Данные технические характеристики являются минимальными и обеспечивают достаточный уровень распознавания жестов кистей рук.

Результаты экспериментального исследования могут быть обобщены в следующих утверждениях, а также проиллюстрированы примерами работы нейронной сети с одной и той же камерой при комбинировании экспериментальных условий (рис.1):

1. Нейронная сеть считывает жесты с камеры в реальном времени, при этом отсутствует необходимость хранить в памяти устройства изображения, сделанные в процессе считывания;
2. Нейронная сеть показывает способность считывать жесты в условиях низкой (0,1 лк) и повышенной освещенности (10000 лк), которые приводят к снижению уровня контрастности между фоном и распознаваемой ладонью;
3. Нейронная сеть показывает способность считывать жесты в условиях нечеткости границ между фоном и ладонью;

Данные выводы свидетельствуют о соответствии выбранной нейронной сети заявленному требованию о достаточном уровне распознавания жестов кисти рук.

Каждое из используемых в экспериментальном исследовании условий, а также их комбинации, нацелены на симуляцию использования итогового устройства потребителем в повседневной жизни:

1. Условие низкой освещенности имитирует использование устройства в темное время суток;
2. Условие повышенной освещенности имитирует использование устройства при солнечной погоде;
3. Условие размытости границ имитирует ситуацию, в которой объектив камеры буде загрязнён;
4. Условие оптического дефекта имитирует попадание на объектив камеры в процессе работы нейронной сети луча солнца или частицы грязи.

Нормальные условия

Низкая контрастность (темнота), нечеткие границы

Низкая контрастность (свет), нечеткие границы

Нечеткие границы, световое пятно

Рис.1. Построение скелета в различных условиях

Описание условий, в которых проводилась съемка:

1. Под нормальными условиями понимается среднее значение освещенности жилой комнаты (1000 лк), при этом границы между фоном и ладонью четко различимы;
2. Под условиями низкой контрастности понимаются условия недостаточной (0,1 лк) или чрезмерной освещенности (10000 лк), при которых контраст между фоном и ладонью уменьшен;
3. Под нечеткими границами понимаются условие, при котором в связи с низкой контрастностью изображения или оптическими дефектами границы между фоном и ладонью теряют четкость или частично не видны;
4. Под оптическим дефектами понимается появление в кадре светового пятна, темного участка или иного подобного явления, уменьшающего четкость границ и/или частично их закрывающего.

3. РАЗРАБОТКА ЭКСПЕРИМЕНТАЛЬНОГО ОБРАЗЦА НЕЙРОННОЙ СЕТИ

В качестве экспериментального образца была использована интегрированная в приложение на ОС Android нейронная сеть MediaPipe Hands. Данное решение было принято в связи с легкостью интеграции нейронной сети в приложение и возможностью впоследствии использовать полученный экспериментальный образец как основу или составную часть итогового программного продукта.

Минимальные требования к смартфону для возможности запуска и корректной работы экспериментального образца нейронной сети:

1. Android
2. Поддержка Bluetooth
3. Процессор с поддержкой ARM64
4. 26МБ свободного места во внутренней памяти и выше

Экспериментальный образец нейронной сети имеет следующие атрибуты:

- Наименование исполняемого файла - handdetection.apk
- Размер исполняемого файла - 26 368 524 байт (25,1 Мб)
- Версия файла - 1.0
- Версия продукта - 1.0
- Внутреннее имя - handdetection
- Исходное имя файла - handdetection.apk
- Название продукта - Hand Reader
- Описание версии файла - 1.0
- Язык - English (United States)

Экспериментальный образец нейронной сети для распознавания жеста кисти руки использует фронтальную камеру смартфона, что приводит к необходимости распознавания зеркального изображения ладони.

Из-за наличия в нейронной сети возможности распознавать положение кисти руки относительно камеры устройства, на котором запущен экспериментальный образец нейронной сети, для полноценной работы достаточно наличие фронтальной камеры на смартфоне или камеры, подключенной к нему по сети Bluetooth.

Минимальные требования к камере, предъявляемые экспериментальному образцу нейронной сети, соответствуют минимальным требованиям к камере, заявленным как результат экспериментального исследования работы нейронной сети в различных условиях, симулирующих реальные.

4. РАЗРАБОТКА ЭКСПЕРИМЕНТАЛЬНОЙ БИБЛИОТЕКИ ЖЕСТОВ ДЛЯ НЕЙРОННОЙ СЕТИ

Для получения представления о работе экспериментального образца нейронной сети, было решено обучить ее распознавать правую и левую руку.

Определение происходит через построение трёхмерного скелета ладони, при этом более яркими белыми кругами обозначены места сочленения костей, смоделированного скелета, находящиеся ближе остальных к камере.

Поскольку экспериментальный образец нейронной сети считывает жесты кисти руки через фронтальную камеру смартфона, это вызвало необходимость учитывать зеркальность изображения. Для этого нейронная сеть, распознавая кисть руки, намерено дает ей противоположное название:

- В случае, если рука распознается как левая, экспериментальный образец нейронной сети выводит сообщение, что рука на экране – правая;
- В случае, если рука распознается как правая, экспериментальный образец нейронной сети выводит сообщение, что рука на экране – левая.

Данный метод позволяет скомпенсировать зеркальность изображения и вывести истинное наименование руки в кадре (рис.2). Успешность распознания – 93 %.

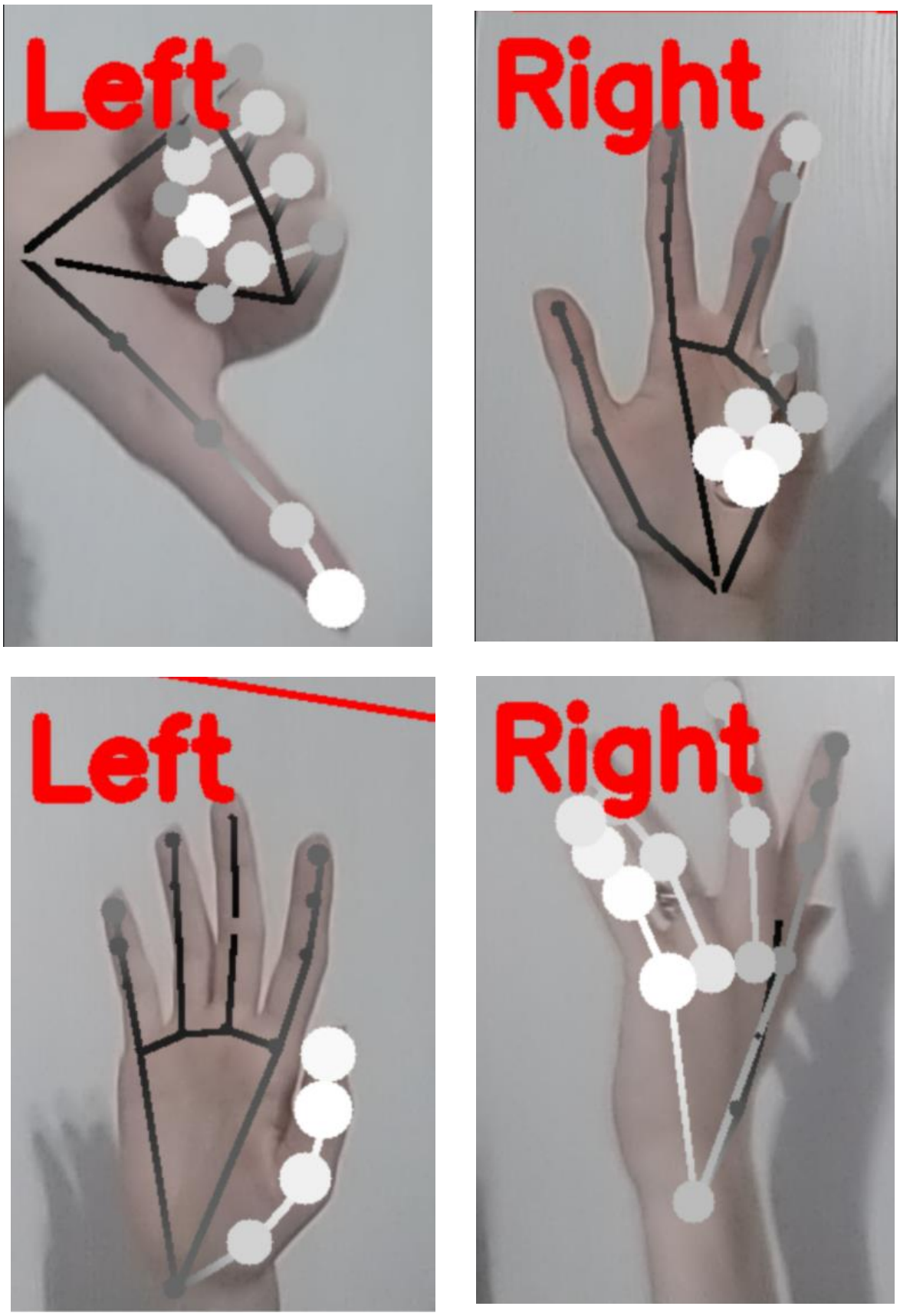

Рис.2. Определение левой и правой ладони в нормальных условиях

5. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ ОПТИМАЛЬНЫХ ТИПОВ ОБЛАЧНЫХ СЕРВИСОВ

В ходе теоретической и экспериментальной исследовательской работы были рассмотрены различные типы подходов к синтезу голоса, а также программные продукты, предназначенные для синтезирования голоса, и были определены основные требования к системе синтеза голоса:

1. Синтезированный голос должен иметь «естественное» звучание, поскольку от этого параметра зависит опыт использования готового устройства потребителем, а также уровень понимания синтезированной речи собеседником;
2. Стоимость использования системы синтеза голоса в случае работы с облачными сервисами не должна иметь сильное влияние на итоговую себестоимость устройства;
3. Необходимо наличие как минимум двух типов голосов – мужского и женского, желательно наличие возможности создания персонализированного голосового модуля для сохранения индивидуальности потребителя при использовании устройства.

Были рассмотрены следующие системы синтеза голоса:

1. Google Text-To-Speech API,
2. Vokalizer,
3. ESpeak,
4. RHVoice,
5. Speechpro.

По итогу исследовательской работы была выбрана система синтеза голоса Google Text-To-Speech API, разработанная Google’s AI technologies, алгоритм работы представлен на рис.3.

Основные критерии выбора:

1. Наличие возможности выбора потребителем соотношения цены и «естественности» звучания голоса:
    a. WaveNet синтезирует голос, имитирующий интонации живого человека, то есть данный вариант максимально приближен к реальной речи человека, но стоимость использования составляет 16$ за 1 млн символов;
    b. Basic синтезирует голос, отличающийся от реальной человеческой речи наличием лишь интонаций (к примеру, знак вопроса в конце предложения) и наличием пауз между словами, то есть данный вариант более роботизирован, но стоимость использования составляет 4 $ за 1 млн символов;

2. Возможность регулировать влияние системы на себестоимость устройства через выбор одного из двух профилей;
3. Возможность интеграции в приложение на смартфон на базе ОС Android при запросе на интеграцию со стороны пользователя.

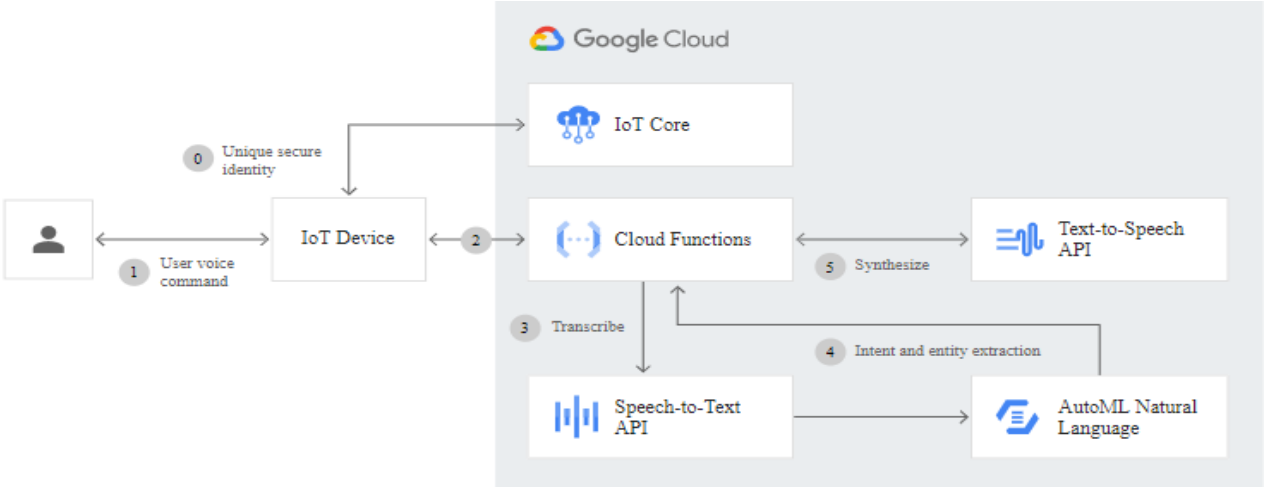

Рис.3 Алгоритм работы Google Text-To-Speech API

ЗАКЛЮЧЕНИЕ

По итогу этапа были сформированы основные требования к нейронной сети распознавания жестов кистей рук, проведено экспериментальное исследование, показавшее соответствие нейронной сети MediaPipe Hands этим требования.

В ходе экспериментальных исследований был составлен перечень искусственно созданных условий, которые использовались для имитации возможных условий использования готового устройства, а также проведены измерения минимально необходимых для корректной работы нейронной сети качества съемки и фокусного расстояния камеры.

Разработан экспериментальный образец нейронной сети, представляющий из себя приложения на ОС Android на основе нейронной сети MediPipe Hands. Планируется использование экспериментального образца как основы или составной части итогового программного продукта.

Создана экспериментальная библиотека, после обучения нейронной сети экспериментальный образец получил функционал определять левую и правую руки в кадре.

Были сформулированы основные требования к системе синтеза голоса, в ходе экспериментального исследования были протестированы 5 типов систем на соответствие этим требованиям, отобрана система синтеза голоса Google Text-To-Speech API.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. On-Device, Real-Time Hand Tracking with MediaPipe // Google AI Blog. – URL:https://ai.googleblog.com/2019/08/on-device-real-time-hand-tracking-with.html
2. MediaPipe Hands // MediaPipe. – URL:https://google.github.io/mediapipe/solutions/hands.html
3. Text-to-Speech // Google Cloud. – URL: https://cloud.google.com/text-to-speech?hl=ru