Автоматическое распознавание слов из ограниченного словаря на основе последовательности изображений с видео

Ткаченко Г. С.
Научный руководитель Фильченков А. А.
Рецензент Карпов А. А.

Университет ИТМО

2017

Мотивация

Технология распознавания речи по визуальным признакам может быть использована для

- Взаимодействия с компьютером
- Идентефикации/верификации спикера (визуальные пароли)
- Камер безопасности
- Улучшения модели распознавания речи

Цель работы

- Предложить подход к распознаванию слов ограниченного словаря на основе только визуальной информации
- Система должна распознавать короткие последовательности слов

Требования

- Качественная аудио и видео дорожка
- Спикер в анфасе
- Ограниченный словарь

Источники

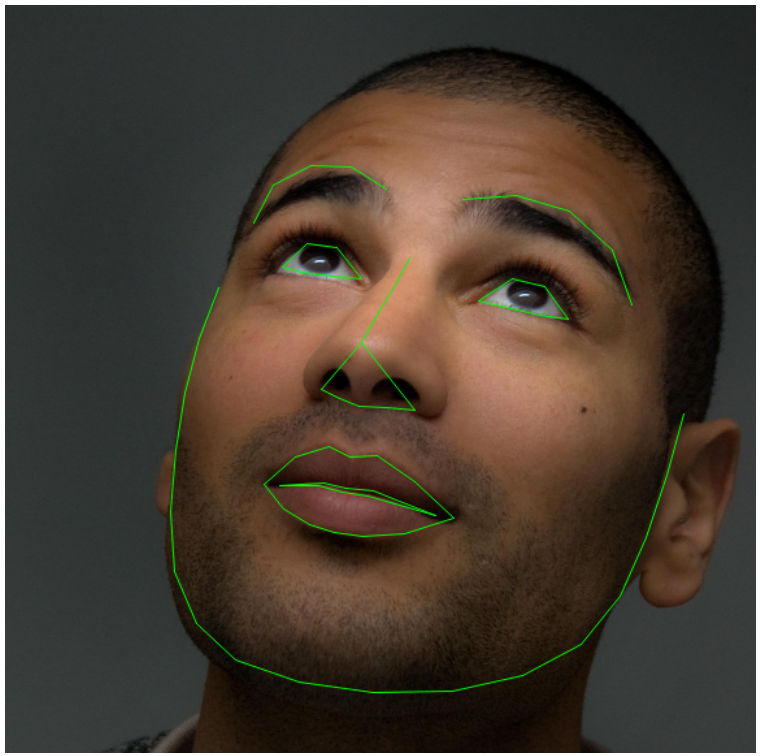
- Видео хостинги
  - Интервью
  - Новости
- Готовые датасеты
  - **GRID** (33 спикера, 1000 видео для каждого)
  - CUAVE (36 говорящих, словарь из 10 цифр)
  - LILiR TwoTalk corpus (4 диалога по 12 минут между двумя людьми)
  - AVLetters1 (10 говорящих, по 3 повторения каждой буквы английского алфавита)
- Записанный датасет подмножества команд Siri

Вспомогательные модели

- Модель предсказания распределения вероятности визем на фрейме
- Модель предсказания меток слов

Подходы

- Геометрические признаки
- Анализ контура губ
  - Active Appearance Models (AAM)
  - Active Shape Model (ASM)
  - **Градиентный бустинг на регрессионных деревьях**
- Анализ значений пикселей
  - Нейронные сети



Оптимизации

- "Сырые"координаты точек не подходят из-за изменения угла и масштаба
  - Нормализация - перенос координат уголков губ в точки с координатами (-1, 0) и (1, 0)
- "Выбросы"и случайный шум мешает обучению
  - Экспоненциальное сглаживание второго порядка
- Многие виземы короткие и 25 FPS недостаточно для распознавания
  - Линейная экстраполяция данных с 25 FPS до 100 FPS

Получение распределения по виземам

- Получение разметки по фонемам с помощью готовой модели, на основе датасета TIMIT
- С помощью таблицы соответствия, получение соответствующей фонемы

Визема	Фонема
uh	uh uw ux
O	ao ix ow oy
...	...

Модель предсказания визем

Нейронная сеть (отображение из вектора признаков видео в распределение по виземам)

Проблемы

- Данных с одного фрейма недостаточно для предсказания виземы
- Соседние фреймы сильно коррелируют между собой

Решения

- Конкатенация векторов с соседних фреймов
- Сжатие вектора
  - Метод главных компонент
  - Нейронные сети (autoencoder)

Постановка задачи

- Онлайн задача
  - На вход в сети подаются данные в режиме онлайн
  - На выходе ожидаются метки слов
- **Оффлайн задача**
  - Сеть имеет возможность прочитать весь вход
  - На выходе ожидаются метки слов

Модель

- Классические нейронные сети - фиксированный вход и выход
- Рекуррентные нейронные сети - фиксированный выход
- Encoder/Decoder LSTM - подходит для оффлайн задачи

Тестовая выборка

- Модель распознавания визем - датасет GRID. 24 спикера в обучающей выборке, 6 спикеров в тестовой выборке
- Модель распознавания меток - датасет Siri. 8 в обучающей, 2 в тестовой

Оптимизации модели распознавания визем

Описание модели	Точность
Window 1 (only current frame) 128(sigmoid)+64(sigmoid)+14(sigmoid)	12%
Window 10-5, autoencoder 72 128(sigmoid)+64(sigmoid)+14(sigmoid)	53%
Window 10-5, pca 72(opt) 128(sigmoid)+64(sigmoid)+14(sigmoid)	54%

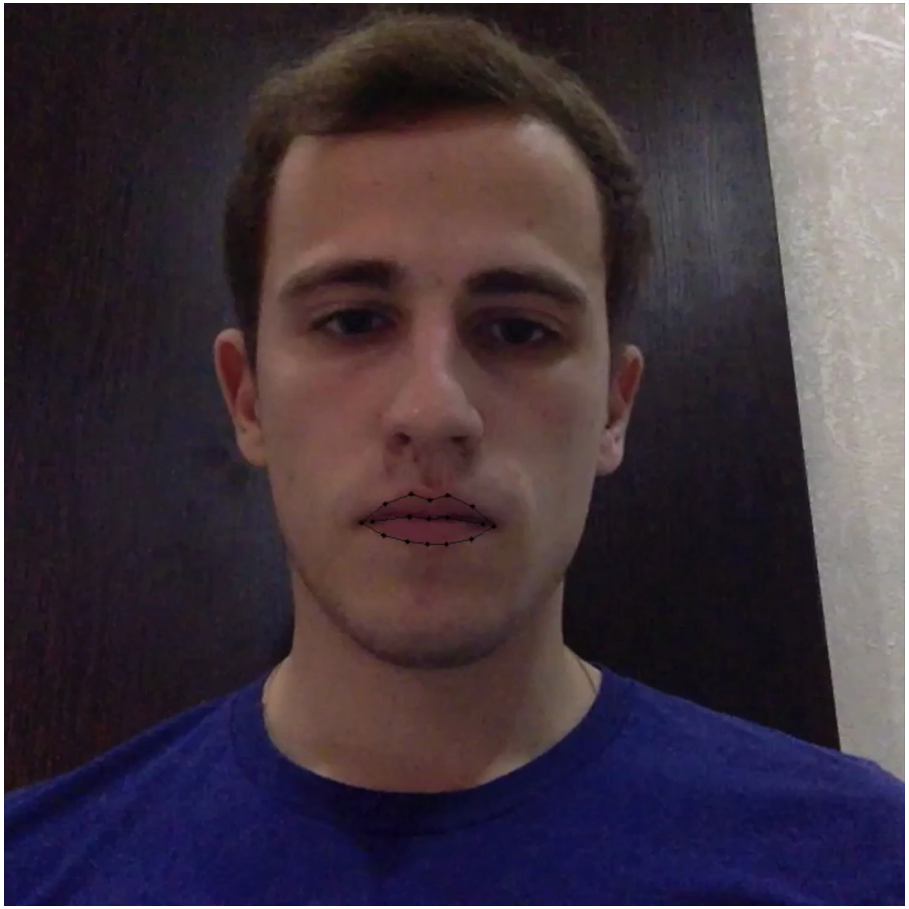
Распознавание визем

Описание модели	Точность
Window 10-5, pca 72(opt) ($\alpha = 0.97$) 128(sigmoid)+64(sigmoid)+14(sigmoid)	52%
Window 20-10, pca 80(opt) ($\alpha = 0.95, \beta = 0.1$) 140(sigmoid)+80(sigmoid)+14(sigmoid)	55%
Window 15-10, pca 75(opt) ($\alpha = 0.95, \beta = 0.1$) 110(relu)+80(relu)+14(relu)	56%
Window 15-10, pca 75(opt) ($\alpha = 0.95, \beta = 0.1$) 110(relu,dropout=0.1)+ 80(relu,dropout=0.15)+14(relu)	58%

Оффлайн задача

Метод	Точность	Метод	Точность
Manual algorithm (5 testers)	44%	KNN-based approach Geometric features	49%
SVM-based approach Eigenlips	68%	SVM-based approach HOG	70%
NN classifier	75%	**Our approach** params = (128, 40, 0.2)	**85%**

Демо



Заключение

- Предложен лучший из найденных методов для распознавания коротких последовательностей слов ограниченного словаря на основе визуальных признаков для speaker-independent задачи
- Предложена оптимизация в виде экспоненциального сглаживания для систем распознавания речи
- Спроектирован и реализован полный workflow для автоматического распознавания речи

Спасибо за внимание!

Вопросы?