ПРИМЕНЕНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ СОЗДАНИЯ АНИМАЦИИ ЛИЦА НА ОСНОВЕ РЕЧИ

В.Г. Спицын, д.т.н., проф. ОИТ ИШИТР,
В.А. Коровкин, аспирант гр. А7-39.
Томский Политехнический университет
E-mail: alcasar@tpu.com

Введение

Сегодня при процедурной генерации мимики лица у виртуальных агентов большое значение имеет движение губ (или технология липсинк).

Множество современных исследований на тему генерации анимации головы и лицевой мимики используют методвы машинного обучения. В качестве основой параметрической модели используют сети LSTM-RNN и общая схема работы таких алгоритмов представлена на рисунке ниже. [1]

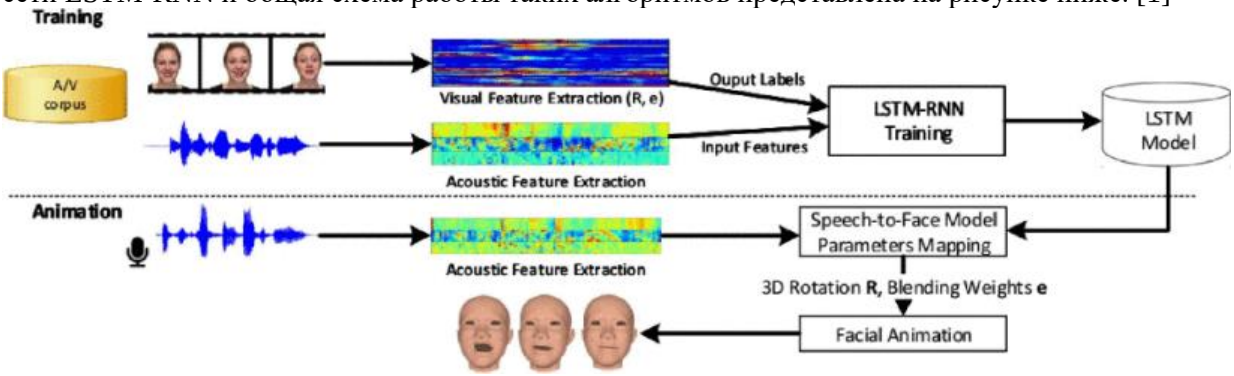

Рисунок 1. Общая схема работа алгоритмов, основанных на сетях LSTM-RNN

Концепция метода

Данный метод в качестве источнников обучения использует аудиовизуальный поток данных, на основе которого отдельно выделяются акустические призанки и им соответствующие визуальные. Полученные признаки и их соотвествие используются во LSTM моделе. После обучения модель использует только аудитопоток, на основе которого строится маска движения лицевых и мимических мышц и вращения головы. Главным недостатком такого метода является сложность его применения в динамических системах, переобучения и совмещения с другими методами лицовой анимации.

Поэтому для решения вышеобозначенных проблем была разработана группа сверточных нейронных сетей на овнове подхода Nvidea и Remedy Entertainment.

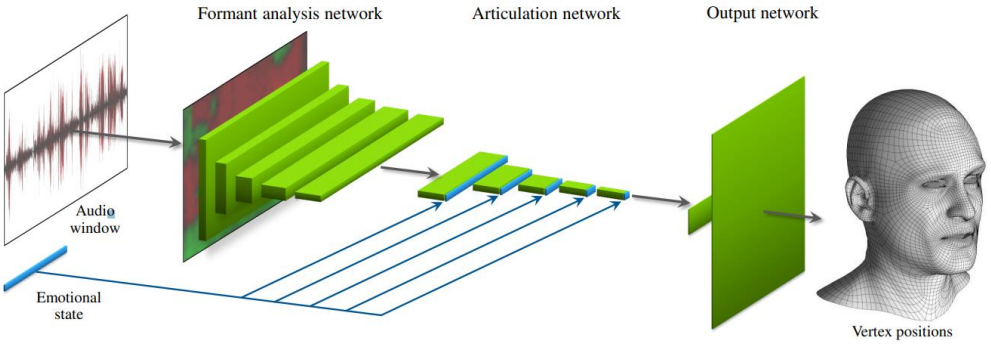

Рис. 1. Общая схема работа алгоритмов, основанных на сетях LSTM-RNN

В данном подходе можно выделить следующие преимущества:

1) классификация эмоций при распозновании речи;
2) выделени дваижения с помощью лицевых ключевых точек для изменения вертексной модели в процессе анимации.

Таким образом, получается параметрическая модель, которая может быть строеннв в мультимодальные решения реальноного времени.

В органильном исследовании использовались высококачесвтенные анимационные динные, полученнные с систем захвата движения (применение технологии Motion Capture), длительностью 3-5 минут. К сожалению, такой датасет является непубличным и слишком уникальным. Для разработки системы был использован датасет SAVEE (Surrey Audio-Visual Expressed Emotion), в котором записаны различные категории эмоций. К недостаткам данного датасета стоит отнести то, что он состоит всего из четырех британских актеров мужского пола с шестью основными эмоциями (отвращение, гнев, счастье, печаль, страх и удивление) и нейтральным состоянии. Всего из датасета было выбрано 480 фонетически сбалансированных предложений для каждого эмоционального состояния. Аудио дискретизируется с частотой 44,1 кГц, а видео – со скоростью 60 кадров в секунду. Фонетическая транскрипция также предоставляется вместе с набором данных. Всего доступно около 102 тыс. кадров для обучения и проверки моделей.

Сразу стоит отметить, что ключевые точки могут быть обнаружены тремя разными способами:

1) размечены вручную;
2) получены с помощью технологии MoCap;
3) выделены с помощью алгоритмов и методов машинного обучения.

В датасете SAVEE лица актеров уже отмечены синими маркерами для отслеживания движений лицецевых мышц во время записи. Они и будут главными визуальными признаками.

В качетсве аудиопризнака извлекается частотный спектральный коэффициент MEL [2] (функция MFSC). Для каждого видеокадра в общей сложности извлекается 36 ключевых точек.

В качестве функции активации использовался гиперболический тангес на каждом уровне с параметром Adam. Архитектурв сети представляется собой следюущие:
Архитектура используемой сети CNN состоит из 2х сверточных слоев.

- Первый слой имеет 64 фильтра размерностью 7
- Слой pooling с размером 4 и шагом 2
- Второй сверточный слой имеет 128 фильтров с размерностью 5
- Слой pooling с размером 2 и шагом 2
- Скрытый полностью связанным слой с 1024 нейронами
- Скрытый полностью связанным слой с 1024 нейронами

Заключение

В результате проведенного исследования была предложена модель сверточной неройнной сети, которая позволяет генерировать ключевые лицевые точки с учетом произносимой речи и испытываемой эмоции. Проведенные численные эксперменты показывают высокую точность классификации работы предложенной архитектуры. Точность состовляет ≈ 94%. Точность распознования сопоставима с современным уровнем.

В дальнейшем планируется отказаться от речевого анализа и воспользоваться в качестве поступающих данных непосредственно морфемы.

Список литературы

1. M. Brand. Voice puppetry // In Proceedings of the 26th annual conference on Computer graphics and interactive techniques. – ACM Press/Addison-Wesley Publishing Co. – 1999. – P. 21–28
2. P. Kakumanu and et al. Speech driven facial animation. In Proceedings of the 2001 workshop on Perceptive user interfaces // ACM. – 2001. – P 1–5.
3. S. Suwajanakorn, S.M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio // ACM Transactions on Graphics (TOG). – 2017. – V. 36. – №4. – P. 95 – 108.
4. Коровкин В.А. Применение методов машинного обучения для решения задачи классификации эмоции на изображении по ключевым точкам // Информационные технологии в науке, управлении, социальной сфере и медицине. Сборник научных трудов VI Международной научной конференции. Под редакцией О.Г. Берестневой, В.В. Спицына, А.И. Труфанов, Т.А. Гладковой. 2019. С. 100-104
5. Коровкин В. А. Распознавание и классификация лицевых эмоций на основе визуальной информации на видеопотоке // Молодежь и современные информационные технологии: сборник трудов XVII Международной научно-практической конференции студентов, аспирантов и молодых ученых (Томск, 17–20 февраля 2020 г.) / Томский политехнический университет. – Томск: Изд-во Томского политехнического университета, 2020. – С. 39-40