Нейросетевая модель классификации эмоций по видеоизображению лица для мобильных устройств

П.В. Демочкина[1], А.В. Савченко[1]

[1]*Национальный исследовательский университет Высшая школа экономики, Большая Печерская 25/12, Нижний Новгород, Россия, 603155*

Аннотация

Для классификации эмоций по видеоизображению лиц предложен вычислительно эффективный алгоритм, в котором для извлечения характерных признаков лица на каждом видеокадре используется нейросетевая модель MobileNet. Эта модель предварительно обучалась для задач идентификации лиц и распознавания возраста и пола, после чего было выполнено ее дообучение для классификации эмоций лиц на статических изображениях. Извлеченные векторы признаков покомпонентно агрегируются с помощью статистических функций (среднее значение, стандартное отклонение и т.п.) в единый дескриптор всего видео. Экспериментальное исследование для набора данных AFEW показало, что классификация таких дескрипторов приводит к достаточно точной и быстрой обработке видеоданных, приемлемой для реализации на мобильном устройстве в режиме реального времени.

Ключевые слова

Распознавание эмоций на видео, анализ изображений лиц, мобильные устройства

1. Введение

Задача классификации эмоций по видеоизображениям лиц заключается в следующем [1]: необходимо поступающей на вход последовательности из $T>1$ видео кадров $\{X(t)\}$, $t = 1, 2, ..., T$ поставить в соответствие одну из заранее определенных эмоций – меток класса $c \in \{1, ..., C\}$. Здесь t – номер кадра во входной последовательности, – общее количество кадров, C – число различных эмоций. Предполагается, что для обучения задан набор из $N>1$ эталонных видеоданных $\{X_n(t)\}$, метки классов c_n которых известны. Существующие наиболее точные методы решения задачи для повышения точности обычно включают в себя ансамбли классификаторов, используют не только видео, но и аудио модальность, а также основываются на вычислительно сложных сверточных нейронных сетях, что препятствует их реализации на мобильных и встроенных устройствах в режиме реального времени. В связи с тем, что основная информация об эмоции содержится в выражении лица наблюдаемого человека, для преодоления указанной проблемы вычислительной сложности в настоящей работе рассматривается возможность использования эффективных нейросетевых архитектур MobileNet, специальным образом предобученных для задач классификации атрибутов лиц.

2. Предложенный подход

Предложенный метод классификации эмоций по видео изображен на Рисунке 1. Вначале для каждого кадра применяется детектор лиц (например, MTCNN), после чего осуществляется извлечение характерных признаков выделенного лица на основе многозадачной (multi-task) нейросетевой модели MobileNet v1 [2] , которая была предобучена на наборе данных VGGFace2 для задач идентификации лиц и предсказания пола и возраста. В настоящей работе эта сеть была дообучена на изображениях из набора данных AffectNet для классификации 7 базовых эмоций (злость, отвращение, страх, радость, грусть, удивление и нейтральное выражение лица). Выход предпоследнего слоя полученной сети и использовался в качестве 1024-мерного вектора характерных признаков эмоций лица каждого кадра. Далее выполняется

агрегация векторов признаков всех кадров – для каждой компоненты (признака) вычисляются статистические функции (среднее, стандартное отклонение, минимум и максимум). В результате каждый видеофайл обучающего множества описывается с помощью дескриптора размерности 4096, который используется для обучения традиционных классификаторов. Наивысшая точность была достигнута с использованием машин опорных векторов (SVM) с линейным ядром. Экспериментальное исследование предложенного подхода проводились на наборе данных Acted Facial Expressions In The Wild (AFEW 8.0) из конкурса EmotiW2019. Результаты сравнительного анализа с методами, использующими одну нейросетевую модель, представлены в Таблице 1.

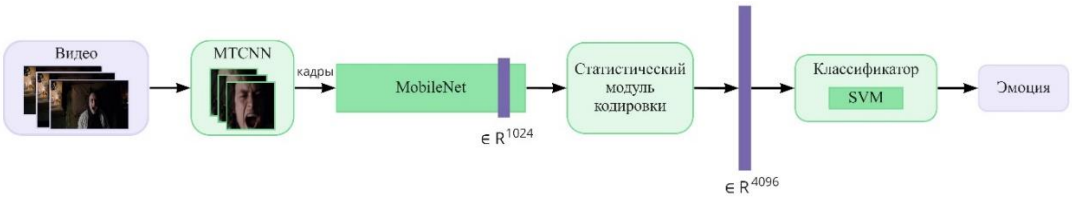

Рисунок 1: Предлагаемая нейросетевая модель классификации эмоций по видеоизображению

Таблица 1
Результаты классификации эмоций по видео, наборе данных AFEW 8.0

Метод	Точность (%)	Время извлечения признаков (мс)
Noisy student with iterative training	55,17	-
Noisy student w/o iterative training	52,49	-
DenseNet-161	51,44	-
FAN	51,18	-
LBP-TOP (baseline) [1]	38,90	-
Предложенный подход, VGGFace	43,86	103,6
Предложенный подход, VGGFace2	50,13	56,2
Предложенный подход, multi-task MobileNet [2]	**54,05**	**20,6**

3. Заключение

Предложенный подход с последовательным дообучением нейросетевой модели для задач идентификации лиц, классификации пола, возраста и эмоций, является не только одним из наиболее точных методов, но и характеризуется достаточно малым временем извлечения признаков лиц на каждом кадре с использованием CPU. В результате обученная модель была реализована в демонстрационном мобильном приложении (https://github.com/HSE-asavchenko/MADE-mobile-image-processing/tree/master/lesson6/src/FacialProcessing). В будущем планируется исследовать возможность повышения точности за счет использования механизма внимания вместо простой покомпонентной агрегации признаков всех кадров.

4. Благодарности

Исследование выполнено за счет гранта Российского научного фонда (проект № 20-71-10010).

5. Литература

[1] Dhall, A. Collecting large, richly annotated facial-expression databases from movies / A. Dhall, R. Goecke, S. Lucey, T. Gedeon // IEEE Multimedia. – 2012. – Vol. 3. – P. 34-41.

[2] Savchenko, A.V. Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet // Peer J. Computer Science. – 2019. – Vol. 5:e197. DOI 10.7717/peerj-cs.197.