ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2022, № 2, с. 80–85

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 004.8

ПОКАДРОВОЕ ОПРЕДЕЛЕНИЕ ЭМОЦИЙ НА ВИДЕОЗАПИСИ ПРИ ПОМОЩИ МНОГОСЛОЙНЫХ НЕЙРОННЫХ СЕТЕЙ

© 2022 г. Ф. Р. Ахияров[a], Л. А. Деревягин[a], В. В. Макаров[a,*], В. И. Цурков[b], А. Н. Яковлев[a]

[a] *МФТИ (НИУ), Москва, Россия*

[b] *ФИЦ ИУ РАН, Москва, Россия*

**e-mail: viktor.makarov@phystech.edu*

Поступила в редакцию 27.09.2021 г.
После доработки 18.10.2021 г.
Принята к публикации 29.11.2021 г.

Предложена архитектура многослойной нейронной сети для решения задачи определения эмоции человека на видеозаписи. Сформированы соответствующие признаки. Под эмоциями понимаются страх, радость, грусть и др. Для обучения используются известные и широко применяемые наборы данных. С помощью комбинирования в архитектуре удается доучить нейросеть и повысить точность ее окончательной версии.

DOI: 10.31857/S0002338822020020

Введение. Распознавание эмоций человека по мимике лица является важной научно-исследовательской проблемой, которая затрагивает множество дисциплин и областей [1–3]. Эта тематика актуальна в таких сферах, как медицина, психология [4] и безопасность. В данной работе рассматриваются подходы к распознаванию эмоций человека по визуальным признакам лица. Применяется глубокое обучение многослойных нейронных сетей.

Существует огромное разнообразие алгоритмов, способных распознавать эмоции человека по мимике лица [5]. Однако качество этих систем уменьшается из-за следующих обстоятельств:

маленькая выборка для обучения,

расхождение в пропорциях лица,

наигранность эмоций,

освещенность во время съемки,

окклюзия,

различный угол поворота головы,

внутриклассовое различие и межклассовое сходство,

этническая принадлежность, пол, возраст.

Использование многослойных нейронных сетей направлено на повышение точности определения эмоций на изображениях.

1. Краткий обзор существующих подходов. Наиболее точно на сегодняшний день эмоции человека были описаны П. Экманом в работе [6], где каждая эмоция была представлена при помощи кодирования лицевых движений, но данный подход сложно автоматизировать [7, 8]. Это связано с тем, что он содержит 46 основных категорий и более 50 дополнительных, а эмоции являются комбинациями таких групп. Поэтому для подготовки такие датасеты (пронумерованный набор изображений, фонограмм или видеозаписей с указанием исследуемых признаков каждого элемента) обрабатываются психологами вручную с учетом указанных обстоятельств.

Особый интерес представляет датасет Aff-Wild [9–11], который состоит из фрагментов видеороликов платформы YouTube, они не являются предзаписанными в видеостудии. Это значит, что условия записи материалов максимально приближены к действительности.

Существует два принципиально разных подхода в распознавании эмоций:

с предварительным алгоритмическим извлечением визуальных признаков и последующей машинной классификацией [12],

с использованием глубоких нейронных сетей без предварительного извлечения признаков.

Визуальные признаки могут быть извлечены при помощи выявления:

геометрических объектов лица (брови, нос, рот, глаза и др.) [13, 14] методами дескриптора line edge map, сравнения направленности градиентов [15, 16], активной модели формы ASM [17], курвлет-преобразования [18], использования структурных моделей [19] и др.;

текстурных особенностей методами фильтра Габора, дискретным вейвлет-преобразованием [20] и др.;

глобальных и локальных объектов методами главных компонент [21], оптического потока [22], морфологическими преобразованиями [23] и др.

Но при условии наличия достаточного датасета наиболее высокой точности классификации удается достичь именно при помощи автоматического выявления признаков и классификацией глубокими нейронными сетями [24–26].

2. Обучение нейронной сети. Сначала осуществляется предварительная обработка. Этот этап предподготовки информации позволяет минимизировать перечисленные ранее причины ошибок. В первую очередь выполняется обнаружение области лица, обрезка изображения и масштабирование для подведения под нужный размер. Затем производится изменение контрастности для уменьшения ошибок, возникающих из-за большой разницы освещения, и выделение общих визуальных особенностей лица. Под этими особенностями понимаются компоненты лица, такие, как губы, нос, рот, брови и т. д., а также немаловажный признак – текстура кожи. Этот этап нужен для более точной постановки задачи перед алгоритмом классификации.

Среди известных подходов обнаружения лиц на изображении существуют две категории.

1. Методы, построенные на конкретном наборе составленных правил, основанных на выделении независимых свойств изображений лиц. Здесь имеет место два этапа построения:

установка явных признаков, характерных для изображений лица,

обработка найденных признаков.

2. Методы, в которых задействован вычислительный вектор признаков, разделяющий изображение на два типа: лицо и не лицо. Выбор метода зависит от установленных ограничений и условий в процессе выполнения задачи. Выделяются следующие возможные ограничения:

пространственные характеристики положения лиц;

наличие или отсутствие ограничений на возможные искусственные помехи на лице;

количество лиц на изображении;

условия освещенности объектов;

цветность изображения;

приоритет в минимизации ложных обнаружений или в количестве обнаруженных лиц;

масштаб лиц и разрешение изображения.

Для анализа видеопотока была выбрана библиотека для проектирования нейросетей – Theano. Это библиотека, которая используется для разработки систем машинного обучения как сама по себе, так и в качестве вычислительного бекэнда для более высокоуровневых библиотек, в данном случае – Lasagne. Также применяется Nolearn как дополнительная и вспомогательная библиотека машинного обучения.

Вычисления проводились для различных датасетов. Для датасета RAVDESS [27] каждый видеоматериал разбивался на части по 0.5 с, и при обработке они перемешивались. Для датасета CK+ [28] из каждой директории брали 3–5 последних кадров, они отбирались многократно и случайным образом, а затем была проведена аугментация в виде поворота кадра в случайный угол на 10°, для того чтобы достичь баланса в классификации. Для датасета Aff-Wild из видео выбирались 4–16 кадров случайного отрезка, а сами видеоматериалы определялись также многократно и случайным образом, после проводилась аугментация для достижения баланса. Дальнейшая работа заключалась в фиксации лица и его переформировании в конкретное изображение лица в формат 256 × 256, чтобы пропустить его через сверточную нейронную сеть.

После прохождения нейронной сетью обучения требуется тестирование ее полученной архитектуры. Происходит это посредством обработки нейронной сетью новых данных. Для этого была сделана выборка из датасета RAVDESS. Для работы над видеоматериалами были также взяты выборки из CK+ и Aff-Wild. На основе этой выборки была проведена оптимизация гиперпараметров для получения более точного результата. Целевая функция использовала кортеж гиперпараметров и возвращала связанные с ними потери. Использовался случайный перебор всех

комбинаций на выборку их случайным образом при помощи библиотеки Keras Tuner, а именно с применением алгоритмов случайного поиска и HyperBand.

Случайный поиск заменяет полный перебор всех комбинаций на выборку их случайным образом. Метод может быть обобщен к непрерывным и смешанным пространствам. Случайный поиск может превзойти поиск по решетке особенно в случае, если только малое число гиперпараметров оказывает влияние на производительность алгоритма обучения машины. Следовательно, задача оптимизации имеет низкую внутреннюю размерность. Случайный поиск также легко параллелизуем и, кроме того, позволяет использовать предварительные данные путем указания распределения для выборки случайных параметров.

3. Комбинирование архитектурой нейронных сетей и датасетов. В рамках проведения экспериментов были получены следующие результаты тестирования нейронной сети для видеофиксации.

Датасет RAVDESS, точность – 69.35%. Классификация по эмоциям:

грусть,
злость,
нейтральное состояние,
отвращение,
радость,
спокойствие,
страх,
удивление.

Датасет CK+: точность – 82.3%. Классификация по эмоциям:

грусть,
злость,
отвращение,
презрение,
радость,
страх,
удивление.

Датасет Aff-Wild: точность – 60.7%. Классификация по состояниям:

нейтральное состояние,
возбужденное позитивное,
возбужденное негативное,
расслабленное позитивное,
расслабленное негативное.

Здесь в каждой из трех групп вводятся неотрицательные коэффициенты (вероятности), сумма которых равна единице. Цель распознавания – нахождение максимального коэффициента.

Остановимся на результатах датасета Aff-Wild. Он является единственным из представленных, в котором материал для анализа содержит изображения в различных ракурсах. Из этого делается следующий вывод: в датасетах CK+ и RAVDESS наличие схожих кадров лиц неизбежно привело бы к ухудшению показателей результата.

На предыдущем этапе разработки была создана и обучена нейросеть, которая показывает хорошую точность в условиях, близких к идеальным: при правильном освещении, фоне, расстоянии от камеры до лица. Но при ухудшении условий точность результатов падает. Поэтому следующей задачей является создание нейросети, которая обучена уже на данных, приближенных к реальным.

Для обучения был выбран датасет Aff-Wild, который участвовал в предыдущем этапе разработки, но в качестве датасета результативной выборки. Сама организация процесса обучения выглядит следующим образом.

Обучаем первичную сеть на простой задаче, удаляя первый класс. При этом датасет сбалансируется, так как в исходной версии было много нейтральных кадров с нулевой результативностью. На этом этапе кадры обрабатываются по отдельности без последовательности. Точность на этом этапе составляет 72%.

Из обученной в предыдущем этапе нейросети убираем последние слои, дойдя до слоя укрупнения (maxpool) с предыдущего блока слоев. Полученную нейросеть делаем первичной. После чего пропускаем через эту нейросеть датасет RAVDESS и сохраняем найденный промежуточный результат.

Таким образом получаем промежуточный датасет для обучения нейросетей по определению эмоций и их силы. В результате такой комбинации архитектур нейросети будут достаточно точны при пропускании через них материалов, приближенных к реальным условиям, потому что первичная нейросеть уже умеет с ними работать, а студийные условия выступают лишь как частный случай.

Нейросети должны объединять в себе точность, скорость работы и простоту реализации. Работа с изображением все так же предполагает работу со сверточными нейронными сетями в силу их спецификации. Точность сети можно предварительно оценить, исходя из результатов теста cifar-10. В тесте убирается первый слой, а после через нейросеть пропускаются картинки размером 32×32, разбитые на 10 классов.

Подходящие под все три критерия сети обладают большим количеством простых слоев и относятся к одному из двух видов:

полносвязные (dense), в которых результат свертки объединяется с исходными данными;

остаточные (residual), в которых результат свертки (или нескольких) суммируется с исходными данными.

Особенность обеих архитектур состоит в том, что градиент ошибки, являющийся фактором обучения, не угасает от слоя к слою, а равномерно обучает все слои сети. Кроме того, обе архитектуры используют после каждого слоя (либо перед каждым слоем) нормализацию внутри партии. Это значит, что из исходных данных вычитается среднее, а отклонение делается равным единице. Этот процесс заметно стабилизирует и ускоряет обучение и заодно повышает точность, однако замедляет работу примерно на 30%.

Выбран вариант нейросети с нарастанием. Используются тонкие свертки с нелинейностью elu и нормализацией результата, потому что исходные данные уже нормализованы, при добавлении новых данных остаются нормализованными, и нет смысла повторять вычисления.

Также для удобства реализации создан слой плотной свертки, в котором последовательно объединены:

тонкая свертка с ядром 1×1,

основная свертка с ядром 3×3 и нелинейностью,

нормализация по партии,

объединения с исходными данными.

Две первые операции эквивалентны обычной свертке, но требуют на порядок меньше вычислений, а значит и времени. Для нормализации (простой вариант) мы обучаем параметры beta и gamma, и итоговое значение будет

$$result = \frac{(conv - conv_{mean})}{conv_{std}} gamma + beta,$$

где $conv_{mean}$ – среднее по осям партии, длине и ширине свертки, $conv_{std}$ – стандартное отклонение по осям партии, длине и ширине свертки.

В отличие от стандартной архитектуры, из модели удалены полносвязные слои, образующие итоговый набор классов. Вместо них при достижении слоем размеров менее 5×5 происходит усреднение внутри слоев.

Видео читается покадрово при помощи библиотеки Scikit-video. Для удобства работы написан итератор, который перебирает элементы контейнерного класса без необходимости пользователю знать реализацию определенного контейнерного класса. Он определяет характеристики видео из заголовка и затем читает видео посекундно. Лишние кадры удаляются, а на оставшихся итератор находит лицо при помощи библиотеки Dlib, и изменяет его размер до 256×256 при помощи библиотек Scikit-image или CV2. Полученные изображения объединяем в партию и обрабатываем первичной нейросетью. Полноценно архитектура сети выглядит следующим образом:

первые слои сети – нормализация внутри партии и обнуление (dropout) с вероятностью 0.1,

далее следует свертка с ядром 5×5 и шагом 2,

4–8 плотных сверток с ядром 3×3 и 15 каналами,

укрупнение размера промежуточного изображения,

слой нормализации,

слой обнуления,

свертки 1 × 1 с числом каналов, равным половине входящих,

последний блок, который впоследствии будет отбрасываться,

обнуление с вероятностью 0.3,

свертка размером 90 × 1 × 1,

6 плотных сверток с ядром 1 × 1,

усреднение слоев,

нормализация по партии,

усреднение каналов до 5 созданных классов.

Контрольная выборка была взята из расширенного датасета RAVDESS, в которой три актера произносили одну фразу с разными эмоциональными оттенками. Итоговая нейросеть правильно классифицировала 91 файл из 100, тем самым можно сделать вывод об увеличении точности на 21% по сравнению с предварительными итогами.

Заключение. Последующее исследование будет связано с анализом и разработкой нейронных сетей для классификации эмоциональных состояний в фонограммах. Совместное применение аудио- и видеосистем увеличит точность распознавания.

СПИСОК ЛИТЕРАТУРЫ

1. *Александров А.А., Кирпичников А.П., Ляшева С.А., Шлеймович М.П.* Анализ эмоционального состояния человека на изображении // Вестн. технологического ун-та. 2019. Т. 22. № 8. С. 120–123.
2. *Заболеева-Зотова А.В.* Развитие системы автоматизированного определения эмоций и возможные сферы применения // Открытое образование. 2011. № 2. С. 59–62.
3. *Люсин Д.В.* Современные представления об эмоциональном интеллекте // Социальный интеллект: теория, измерение, исследования / Под ред. Д.В. Люсина, Д.В. Ушакова. М.: Изд-во Ин-та психологии РАН, 2004. С. 29–36.
4. *Гранская Ю.В.* Распознавание эмоций по выражению лица: Автореф. дис. ... канд. психологических наук по специальности 09.00.01. СПб., 1998.
5. *Бобе А.С., Конышев Д.В., Воротников С.А.* Система распознавания базовых эмоций на основе анализа двигательных единиц лица // Инженерный журнал: наука и инновации. 2016. № 9. С. 7.
6. *Ekman P.* Facial Action Coding System. Palo Alto. USA: Consulting Psychologist Press, 1978.
7. *Kollias D., Zafeiriou S.* Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace // arXiv preprint arXiv: 1910.04855, 2019.
8. *Kollias D.* Face Behavior a la carte: Expressions, Affect and Action Units in a Single Network // arXiv preprint arXiv: 1910.11111, 2019.
9. *Kollias D.* Analysing Affective Behavior in the First ABAW 2020 Competition // arXiv preprint arXiv:2001.11409, 2020.
10. *Kollias D.* Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond // Intern. J. Computer Vision (IJCV). 2019. № 127. P. 907–929.
11. *Kollias D.* Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study // arXiv preprint arXiv:2105.03790, 2020.
12. *Abdulrahman M., Eleyan A.* Facial Expression Recognition Using Support Vector Machines // Proc. 23nd Signal Processing and Communications Applications Conf. (SIU 2015). Malatya, Turkey, 2015. P. 276–279.
13. *Tripathi A., Pandey S.* Efficient Facial Expression Recognition System Based on Geometric Features Using Neural Network // Lecture Notes in Networks and Systems. 2018. V. 10. P. 181–190.
14. *Hernandez-Matamoros A., Bonarini A., Escamilla-Hernandez E., Nakano-Miyatake M., Perez-Meana H.* A Facial Expression Recognition with Automatic Segmentation of Face Regions // Communications in Computer and Information Science. 2015. V. 532. P. 529–540.
15. *Jumani S.Z., Ali F., Guriro S., Kandhro I.A., Khan A., Zaidi A.* Facial Expression Recognition with Histogram of Oriented Gradients Using CNN // Indian J. Science and Technology. 2019. V. 12. № 24. P. 1–8.
16. *Greche L., Es-Sbai N., Lavendelis E.* Histogram of Oriented Gradient and Multi Layer Feed Forward Neural Network for Facial Expression Identification // Proc. Intern. Conf. on Control, Automation and Diagnosis (ICCAD 2017). Hammamet, Tunisia, 2017. P. 333–337.

17. *Iqtait M., Mohamad F.S., Mamat M.* Feature Extraction for Face Recognition Via Active Shape Model (ASM) and Active Appearance Model (AAM) // IOP Conf. Series: Materials Science and Engineering. Tangerang Selatan, Indonesia, 2018. V. 332. P. 1–8.
18. *Candès E., Demanet L., Donoho D., Ying L.* Fast Discrete Curvelet Transforms // Multiscale Modeling & Simulation. 2006. V. 5. № 3. P. 861–899.
19. *Себряков Г.Г., Визильтер Ю.В.* Разработка методики построения специализированных экспертных систем для анализа цифровых изображений в задачах обнаружения и идентификации сложных структурных объектов // Вестн. компьютерных и информационных технологий. 1997. № 3. С. 31.
20. *Nigam S., Singh R., Misra A.K.* Efficient Facial Expression Recognition Using Histogram of Oriented Gradients in Wavelet Domain // Multimedia Tools and Applications. 2018. V. 77. № 21. P. 28725–28747.
21. *Varma S., Shinde M., Chavan S.S.* Analysis of PCA and LDA Features for Facial Expression Recognition Using SVM and HMM Classifiers // Techno-Societal 2018: Proc. 2nd Intern. Conf. on Advanced Technologies for Societal Applications. Berlin, Germany, 2019. V. 1. P. 109–119.
22. *Zhao J., Mao X., Zhang J.* Learning Deep Facial Expression Features from Image and Optical Flow Sequences Using 3D CNN // Visual Computer. 2018. V. 34. № 10. P. 1461–1475.
23. *Визильтер Ю.В., Выголов О.В., Желтов С.Ю., Князь В.В.* Метрический подход к семантико-морфологическому сравнению изображений // Вестн. компьютерных и информационных технологий. 2020. Т. 17. № 5 (191). С. 3–12.
24. *Рюмина Е.В., Карпов А.А.* Аналитический обзор методов распознавания эмоций по выражениям лица человека // Научно-технический вестник информационных технологий, механики и оптики. 2020. № 2. С. 163–176.
25. *Talegaonkar I., Joshi K., Valunj S., Kohok R., Kulkarni A.* Real Time Facial Expression Recognition Using Deep Learning // Proc. of Intern. Conf. on Communication and Information Processing (ICCIP). 2019 [Электронный ресурс]. URL: https://ssrn.com/abstract=3421486.
26. *Визильтер Ю.В., Горбацевич В.С., Желтов С.Ю.* Структурно-функциональный анализ и синтез глубоких конволюционных нейронных сетей // Компьютерная оптика. 2019. Т. 43. № 5. С. 886–900.
27. *Livingstone S.R., Russo F.A.* The Ryerson Audio-visual Database of Emotional Speech and Song (RAVDESS): A dynamic, Multimodal Set of Facial and Vocal Expressions in North American English // PLoS ONE. 2018. V. 13. № 5. С. 1–35.
28. *Lucey P.* The Extended Cohn-Kanade Dataset (CK+): A Complete Bataset for Action Unit and Emotion-specified Expression // Proc. IEEE CVPR Workshop on Biometrics. San Francisco: IEEE Computer Society, 2010. P. 94–101.