Распознавание выражений лиц на основе адаптации классификатора видеоданных пользователя

Е.Н. Чураев [1], А.В. Савченко [1,2]
[1] Национальный исследовательский университет Высшая школа экономики, Лаборатория алгоритмов и технологий анализа сетевых структур, 603093, Россия, г. Нижний Новгород, ул. Родионова, д. 136;
[2] Сбер, Лаборатория искусственного интеллекта, 121170, Россия, г. Москва, Кутузовский проспект д. 32, строение 2

Аннотация

В настоящей работе предложен метод распознавания выражений лиц по видео, позволяющий значительно увеличить точность при помощи адаптации модели к эмоциям конкретного пользователя, например, владельца мобильного устройства. На первом этапе нейросетевая модель, предварительно обученная распознавать выражения лиц на статических фото, применяется для извлечения визуальных признаков лиц на каждом видеокадре. Далее они агрегируются в единый дескриптор для короткого фрагмента видео, после чего обучается нейросетевой классификатор. На втором этапе предлагается выполнить адаптацию этого классификатора с использованием небольшого набора видеоданных с выражениями лиц конкретного пользователя. После принятия решения пользователь может корректировать предсказанные эмоции для дальнейшего повышения точности персональной модели. В рамках экспериментального исследования для набора данных RAVDESS показано, что подход с адаптацией модели под конкретного пользователя позволяет значительно (на 20 – 50 %) повысить точность распознавания выражений лиц по видео.

Ключевые слова: распознавание выражений лиц, адаптация нейросетевого классификатора, распознавание лиц.

Цитирование: **Чураев, Е.Н.** Распознавание выражений лиц на основе адаптации классификатора видеоданных пользователя / Е.Н. Чураев, А.В. Савченко // Компьютерная оптика. – 2023. – Т. 47, № 5. – С. 806-815. – DOI: 10.18287/2412-6179-CO-1269.

Citation: Churaev EN, Savchenko AV. Facial expression recognition based on adaptation of the classifier to videos of the user. Computer Optics 2023; 47(5): 806-815. DOI: 10.18287/2412-6179-CO-1269.

Введение

Распознавание выражений лиц (англ. Facial expression recognition) по видеоизображению является одной из наиболее сложных проблем компьютерного зрения. Анализ выражений лиц зачастую позволяет понять внутреннее состояние человека и эмоции, которые он испытывает в данный момент времени. В задачах распознавания эмоций человека по выражению лица наиболее часто используется категориальное представление эмоций, включающее нейтральное состояние и шесть базовых эмоций Экмана [1]: гнев, отвращение, страх, счастье, печаль, удивление, нейтральность. Иногда в существующих наборах данных добавляется еще одна эмоция, например, спокойствие [2] или презрение [3]. Другим дискретным способом представления эмоций является модель Facial Action Coding System (FACS) [1], которая описывает движения мышц лица, отвечающих за выражение эмоций и настроения. В FACS выделены более 40 мускулатурных действий на лице человека и дан их числовой код, что позволяет квантифицировать различия в выражении эмоциональных состояний между людьми. Также для описания эмоций может использоваться непрерывная пространственная модель, которая представляет собой двухмерное пространство валентности и возбуждения (VA-пространство) [4]: валентность показывает, насколько положительным или отрицательным является эмоциональное состояние, а возбуждение показывает, насколько оно пассивно или активно.

Понимание эмоционального состояния людей может быть важным во многих сферах, например, для предотвращения возможных конфликтных ситуаций по данным видеонаблюдения [5], определения внимания и стресса водителя во время движения [6], отслеживания изменений в эмоциях человека в зависимости от окружения и различных внешних событий [7], определения реакции покупателей на рекламную компанию [8], в рекомендательных системах [9] и т.п.

В настоящее время практически все современные методы распознавания эмоций по изображению лица используют глубокое обучение и свёрточные нейронные сети (СНС). Зачастую для многих прикладных задач одним из важных требований является офлайн-обработка видео в режиме реального времени на мобильных или встроенных устройствах без отправки данных на удалённый сервер. Такое пожела-

ние обусловлено как вопросами безопасности персональных данных [10], так и необходимостью работать без доступа к сети Интернет. Спектр различных устройств, на которых модель должна эффективно работать, довольно широк: от бюджетного смартфона до персонального компьютера, поэтому для извлечения эмоциональных признаков из видеокадров имеет смысл использовать эффективные по вычислительной сложности и затратам памяти нейросетевые модели, основанные на архитектурах MobileNet [11] и EfficientNet [12].

К сожалению, даже современные подходы к распознаванию эмоций по видеоизображению лица показывают довольно низкую точность на различных наборах данных. Например, для набора данных Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [2] точность распознавания известных методов оказывалась равной 70 – 80 % [13 – 15], что может быть недостаточно для промышленного применения. Поэтому в предыдущей работе авторов [16] представлен новый подход, основанный на идее дикторозависимого распознавания речи [17, 18], позволяющий увеличить точность распознавания эмоций по видеоизображению лиц путём адаптации модели под выражения лица конкретного пользователя.

В настоящей статье проведено расширенное исследование такого адаптивного подхода. В традиционный алгоритм распознавания выражения лица на видео с помощью предобученной нейросетевой модели предложено добавить обратную связь, позволяющую пользователю проверять результат классификации и при необходимости корректировать его. Эти данные накапливаются и используются для дообучения нейросетевого классификатора для конкретного пользователя. При применении подобного подхода для распознавания выражений лиц по видео используются термины «дикторозависимая» (персональная) и «дикторонезависимая» (универсальная) модель, широко применяющиеся в литературе по распознаванию эмоций в речи [17]. Проведено обширное экспериментальное исследование с использованием разных моделей EmotiEffNet [12] для извлечения визуальных признаков из кадров. Полученные результаты и сделанные по ним выводы представляют интерес для широкого круга специалистов в области распознавания образов и анализа изображений лиц.

1. Постановка задачи

Задача распознавания выражений лиц по видео может быть сформулирована следующим образом: необходимо определить метку класса эмоции $c \in \{1,\ldots,C\}$ пользователя системы на основе последовательности из n видеокадров $\{X_i\}$, $i = 1, 2,\ldots, n$, содержащих его лицо. Здесь n – число кадров во входном видео, а C – количество классов различных эмоций (счастье, злость и т.д.), обычно $C = 7$ или $C = 8$. При этом предполагается, что на всём видео эмоция пользователя оставалась неизменной. Традиционные методы решения задачи основаны на анализе ключевых точек лица и определении геометрических признаков [19]. В них визуальные признаки извлекаются только из определённых областей лица, таких как глаза, нос и рот. Подобные алгоритмы вычисляют различные функции, такие как локальные бинарные шаблоны (Local Binary Patterns) [20] или локальная фазовая квантизация (Local Phase Quantization) [21], затем используют их для распознавания выражения лица.

В настоящее время для решения задачи чаще применяются глубокие СНС, которые могут быть менее требовательными к вычислительным ресурсам и более энергоэффективными [22, 23], а также демонстрируют более высокую точность предсказаний в сравнении с традиционными подходами [24]. Использование легковесных СНС позволяет решать подобные задачи в режиме реального времени на широком спектре различных устройств. Одними из наиболее точных для многих свободно распространяемых наборов данных являются предложенные ранее одним из авторов нейросетевые модели MobileEmotiFace [11] и EmotiEffNet [12].

Стандартный подход, использующий СНС для распознавания эмоций по видеоизображению лица, включает в себя несколько этапов. На первом этапе производится разбиение видео на кадры и подготовка этих кадров к извлечению визуальных признаков [25]. На кадрах находятся области с лицом человека, например, с помощью детектора MTCNN (Multi-Task Convolution Neural Network) [26], далее эти области обрезаются. Выделенные лица могут выравниваться по ракурсу с помощью трансформации изображения, когда 5 ключевых точек лица, возвращаемые детектором, переносятся на заранее заданные фиксированные позиции.

На втором этапе осуществляется извлечение признаков лиц на каждом кадре. Стандартным подходом для данной задачи является использование СНС, предобученных на распознавание лиц, таких как ResNet-50 (VGGFace2) [27], VGG13 (VGGFace) [28] и т.д. Недостатком таких моделей является довольно большая вычислительная сложность, а следовательно, и время работы. Для ускорения данного шага использовались легковесные модели, основанные на MobileNet-v1 [11] и EfficientNet [12], которые были предобучены на наборе данных VGGFace2. В работе [11] утверждается, что невозможно использовать признаки атрибутов лица, таких как возраст, пол, этническая принадлежность, для распознавания выражений лиц. Это связано с тем, что лицевые признаки одного человека даже для различных эмоций будут практически идентичны, что не позволит добиться достаточной точности в задаче распознавания эмоций по изображению лица. Поэтому эти модели были дообучены для распознавания восьми эмоций (нейтральность, спокойствие, счастье, грусть, гнев, испуг,

удивление и презрение) на большом наборе данных AffectNet [3]. Точность, демонстрируемая этими моделями, оказалась сравнимой, а иногда даже выше точности известных моделей распознавания эмоций для большого числа свободно распространяемых наборов данных [12].

Подавая на вход СНС вырезанное из *i*-го кадра изображение лица, на выходе одного из последних слоёв извлекается вектор визуальных признаков f_i. На третьем этапе извлечённые признаки лиц на каждом кадре объединяются в дескриптор входного видеофрагмента. Такой дескриптор может подаваться на вход специально обученного классификатора для оценки выражения лица на видео.

Обычно применяют один из двух способов агрегации эмоциональных признаков лиц в видеодескриптор. Первый способ основан на объединении визуальных признаков с помощью статистической функции, в частности, вычислялись среднее или максимальное значение признаков для всех кадров в видео [11]. Данный тип дескриптора использовался для обучения классификатора логистической регрессии.

Второй способ основан на обработке видео как последовательности кадров с помощью рекуррентных нейронных сетей или моделей внимания, например, Frame Attention Network (FAN) [29]. Подобные сети предназначены для адаптивного объединения визуальных признаков. В экспериментах рассматривались два типа моделей внимания, описанных авторами оригинальной статьи [29]: single attention и self-attention.

Модель «Single attention» использует полносвязанный слой с сигмоидной функцией активации σ, применяемый к признакам кадра (f_i) для получения весов внимания (α_i):

$$\alpha_i = \sigma\left(f_i^T q^0\right). \tag{1}$$

Здесь q^0 – параметр полносвязанного слоя. Затем эти веса используются для получения дескриптора всего видео:

$$f_v^{'} = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i}. \tag{2}$$

Модель «Self-attention» [30] использует слой внимания из библиотеки Keras, который реализует следующую обработку признаков кадров с идентичными ключами и значениями. В оригинальном алгоритме «Self-attention» используется три компоненты на входе в модель: запрос (queue), ключ (key) и значение (value). Запрос — это вектор, представляющий информацию, которую мы хотим найти в исходных данных. Ключ — это вектор, описывающий информацию, которую мы ищем в исходных данных. А значение — это вектор, содержащий информацию, которую мы хотим получить из исходных данных.

$$q_i = f_i^T w^q; k_i = v_i = f_i^T w^k, \tag{3}$$

$$\alpha_i = \frac{q_i k_i^T}{\sqrt{d_k}}, \tag{4}$$

$$\beta_i = \frac{\alpha_i v_i}{\sum_{j=1}^{n} \alpha_j}, \tag{5}$$

$$\gamma_i = \sigma\left(\beta_i q^0\right), \tag{6}$$

$$f_v^{'} = \frac{\sum_{i=1}^{n} \gamma_i f_i}{\sum_{i=1}^{n} \gamma_i}. \tag{7}$$

Полносвязанные слои используются для вычисления векторов q_i, k_i и v_i с помощью матриц весов w^q и w^k. Затем для каждого вычисляются значения внимания α_i как скалярное произведение q_i на k_i, нормированное на корень квадратный из размерности ключей (d_k). Далее α_i используется для вычисления весов β_i, которые подставляются в формулы (6), (7), аналогичные вычислениям в модели single attention (1), (2). В конце моделей добавляется полносвязанный слой с активацией softmax для классификации выражений лиц, и вся модель классификации видео обучается в режиме end-to-end.

2. Предложенный подход

К сожалению, точность распознавания выражений лиц для описанного в предыдущем параграфе подхода оказывается достаточно низкой – 40 – 70 % для типовых наборов данных [11], что не всегда приемлемо для практических приложений. Основная причина низкой точности состоит в существенных отличиях используемого для обучения наборов видеоданных и эмоциональных особенностей конкретного пользователя, а также различия в условиях съемки. Поэтому в настоящей работе был расширен подход, позволяющий значительно увеличить точность за счёт адаптации модели под конкретного пользователя [16]. В алгоритм был добавлен шаг, позволяющий пользователю проверять и корректировать точность распознавания модели. При исправлении или подтверждении предсказанной эмоции пользователю даётся выбор из C выражений лиц, где C – число классов различных эмоций. Эти ответы накапливаются и затем используются для обучения новой персональной модели (рис. 1).

Первым этапом нейросетевые классификаторы видеодескрипторов обучались на доступном наборе данных, который не содержал лица пользователей, этот этап представлен в верхней части рис. 1. В результате была получена универсальная (дикторонезависимая) модель предсказания эмоций по видеоизображению лица. После этого на втором шаге для каждого пользователя предлагается проводить адаптацию – дообучение (fine-tuning) дикторонезависимой модели с использованием только видеоданных выбранного пользователя в предположении о том, что истинные метки эмоций известны. При этом стоит отметить,

что СНС, применяемые для извлечения эмоциональных признаков, не дообучаются, поэтому число требуемых примеров видео каждого пользователя может быть относительно мало.

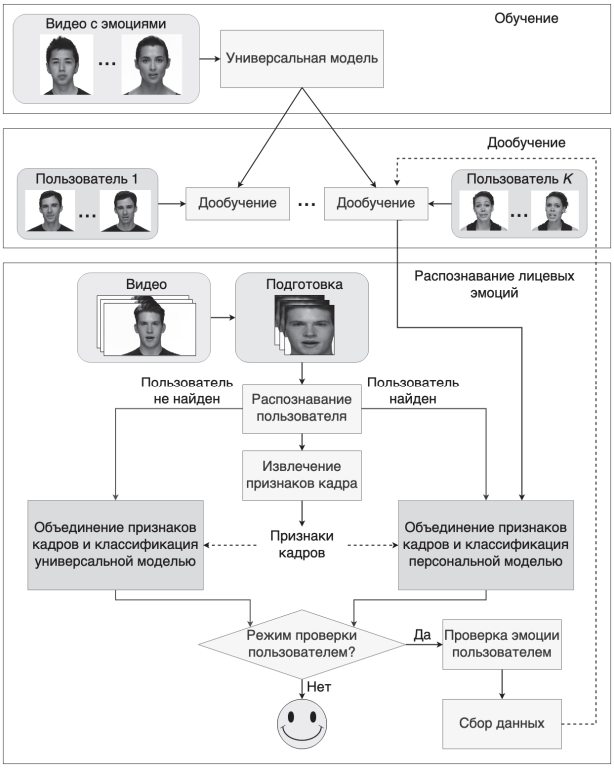

Рис. 1. Предлагаемый адаптивный подход для распознавания выражений лиц на видео

На третьем этапе, в нижней части рис. 1, представлен алгоритм распознавания эмоций. Входное видео разбивается на кадры, на каждом из них с помощью детектора выделяется область лица, которая подготавливается для подачи на вход в СНС. Нейронные сети для классификации атрибутов лиц обычно принимают на вход изображения лиц в размере 224×224 или 112×112 [11, 31], поэтому в начале картинка масштабируется и выравнивается для получения лица, смотрящего прямо в камеру. В процессе такого преобразования пропорции лица деформируются, но это не влияет на качество итогового решения модели, обученной с использованием большого числа изображений лиц с идентичными размерами ширины и высоты.

В предложенном подходе одновременно с извлечением эмоциональных признаков лица применяется модель для идентификации пользователя. Если лицо найдено в базе зарегистрированных пользователей и для него существует дообученная (персональная) модель [16], то она будет использоваться для распознавания эмоций. В противном случае используется предобученный классификатор (универсальная модель).

Существует довольно много методов, которые позволяют с высокой точностью распознать пользователя по фото и видео его лица. Один из наиболее современных алгоритмов использует специальную функцию потерь ArcFace (Additive Angular Margin Loss) для обучения СНС извлекать характерные признаки лица [31]. В настоящей статье особое внимание уделялось возможности работы на мобильных устройствах, поэтому при реализации алгоритма использовались легковесные модели [11], предобученные на наборе данных VGGFace2.

Для предоставления возможности пользователю улучшать точность распознавания эмоций в алгоритм (рис. 1) добавлена возможность адаптации классификатора. Если активирован режим валидации результата пользователем, то после распознавания выражения лица пользователь может подтвердить или скорректировать результат работы модели. В случае, если классификатор дал неверное предсказание, пользователь выбирает другую эмоцию, соответствующую тому, что было показано на видео. Эти данные накапливаются, передаются на этап «Дообучения» и используются для улучшения и тренировки персональных моделей. При наличии небольшого числа размеченных пользовательских видео такой подход позволяет значительно увеличить точность распознавания эмоций (в идеале 99 – 100 %), если видео, подаваемое на вход модели, было записано при тех же условиях съемки (ракурс, освещение и т.п.), которые использовались во время сбора данных для дообучения модели. Конечная точность предсказания зависит от качества входного видео. Если из-за плохого освещения или сильного поворота головы не удаётся провести идентификацию пользователя, то будет использоваться универсальная модель, точность предсказания которой ниже.

3. Экспериментальный набор данных

При обучении модели для распознавания эмоций по видео в настоящей работе использовался набор данных RAVDESS [2]. Он содержит видео и аудиозаписи с 24 актёрами (12 мужчин и 12 женщин). Актёры поют и произносят две одинаковые фразы: «Kids are talking by the door» и «Dogs are sitting by the door». Разметка набора данных включает в себя 8 стандартных эмоций. Каждый видеофрагмент длительностью 3 секунды и частотой 25 кадров в секунду содержит наблюдение за одним актёром с одной эмоцией.

Этот набор данных считается довольно простым из-за того, что во многих статьях [32, 33] получил точность более 90 %. Для примера, в табл. 1 приведены результаты, полученные в работах, где использовалось рандомизированное разбиение на тренировочный и тестовый наборы данных. При таком разбиении один и тот же актёр мог попасть и в тренировочную, и в тестовую выборки, но с разными эмоциями. Таким образом, модель уже на этапе обучения знала некоторую информацию об актёрах, которые встретятся в тестовом наборе. Соответственно, точность предсказания была значительно выше. Во всех рабо-

тах, где применялось подобное разбиение, была получена довольно высокая точность (англ. accuracy) распознавания эмоций, которая вычислялась как отношение количества видео с правильно предсказанными эмоциями к общему числу видео из тестового набора данных.

Табл. 1. Точность распознавания выражений лиц на видео для набора RAVDESS при случайном разбиении на тренировочное и тестовое множества

Метод классификации	Модальность	Точность (%)
VGG16 (video) + MFCC/pitch/energy (audio) [8]	Video, Audio	98,14
VGG16 [8]	Video	97,77
Deep Belief Network, Gabor filter for facial landmarks (video) + MFCC/pitch/energy (audio) [32]	Video, Audio	97,92
Deep Belief Network, Gabor filter for facial landmarks [32]	Video	96,53
SVM, landmarks + COG [33]	Video, Audio	97,26
Modified VGG16 [34]	Video	92
MobileNet-v1 + Single attention [13]	Video	97,34

К сожалению, в этих работах использовалось случайное разбиение набора данных на тренировочную и тестовую выборки. В таком случае видеофайлы с одним актёром попадают в обе выборки, поэтому при разбиении случайным образом классификатор уже имеет некоторую информацию об актёрах из тестового набора и из-за этого показывает высокую точность классификации эмоций. Несмотря на высокую точность на тестовом множестве, если применить эту модель для классификации эмоции нового пользователя, которого нет ни в одном из наборов, то точность предсказания будет значительно ниже.

Поэтому в предыдущей работе авторов настоящего исследования [13] и в ряде других статей [14, 15] набор данных RAVDESS разделяется по актёрам, когда все видео части актёров попадают в обучающее множество, а видео оставшихся людей – в тестовую выборку. При таком подходе к разделению на тестовое и тренировочное множества, каждое из множеств содержит только уникальных актёров. В этом случае точность классификации ощутимо снижается (табл. 2).

Например, в статье [14] набор данных был разделён по актёрам в следующем соотношении: 75 % актёров в тренировочном наборе и 25 % в тестовом. Авторы использовали модель активности лицевых мышц для распознавания типа эмоции в режиме реального времени. OpenFace [36] использовался для извлечения значений активации лицевых мышц. Stacked Auto Encoder (SAE) использовал эти значения в качестве вектора признаков. Результатом работы SAE было наилучшее сочетание мышц, описывающее конкретную эмоцию. Авторам этой работы удалось добиться точности в 84,91 % для распознавания выражений лиц на видео.

Табл. 2. Точность распознавания выражений лиц на видео для набора RAVDESS при разбиении на тренировочное и тестовое множества по актёрам

Метод классификации	Модальность	Точность (%)
SAE [14]	Video	84,91
VGG16 [35]	Video	79,74
MobileNet-v1 + Single attention [13]	Video	81,01
Sequential (bi-LSTM) [15]	Video	57,08
Sequential (bi-LSTM) (video) + CNN-14 (audio) [15]	Video, Audio	80,08
bi-LSTM + MLP (video) and xlsr-Wav2Vec2.0 + MLP (audio) [15]	Video, Audio	86,70

В работе [35] для извлечения признаков из видео использовалась модель VGG16 с предобученными весами на наборе данных ImageNet. Затем рекуррентная нейронная сеть применялась для обработки последовательности признаков кадров. Набор данных RAVDESS разделялся в той же пропорции, что и в данной работе (80 % видео в обучающую выборку, остальные 20 % – в тренировочную). Авторы этой работы достигли точности в 79 % для распознавания эмоций на видео.

В настоящей статье воспользуемся следующим разбиением набора RAVDESS: актёры 1, 2, 3, 4 и 5 были выделены в качестве пользователей системы и использовались в режиме персональной классификации эмоций, а универсальная модель обучалась на 3864 видео остальных 19 актёров. Пример разбиения на тренировочное и тестовое множества приведён на рис. 2. Для обучения персональной модели видеофайлы для каждого из 5 актёров тестового множества были перемешаны случайным образом и разбиты на тренировочный (80 % всех видео) и тестовый наборы. Точность персонализированного классификатора вычислялась для каждого из пяти актёров на оставшихся 20 % его видео. Кроме того, дополнительно проводились эксперименты по определению необходимого количества видео для достижения высокой точности распознавания эмоций универсальной моделью.

4. Результаты экспериментальных исследований

Предложенный подход (рис. 1) был реализован с помощью TensorFlow 2.5. Для извлечения эмоциональных признаков использовались СНС одного из авторов, показавшие наилучшую точность для различных свободно распространяемых наборов фото и

видео эмоциональных лиц, а именно, модель MobileEmotiFace на основе архитектуры MobileNet v1 [11] и две модели EmotiEffNet на основе архитектур EfficientNet-B0 и EfficientNet-B2 [12]. Извлеченные признаки классифицировались с помощью многоклассовой логистической регрессии (однослойной нейронной сети прямого распространения) и двух моделей внимания. Дескриптор видео вычислялся с помощью функции покомпонентного максимума, вычисленной по векторам признаков всех видеокадров [16].

Рис. 2. Пример разбиения набора RAVDESS на тренировочное и тестовое множества

На первом шаге универсальная модель обучалась на всех актёрах из набора RAVDESS, кроме актёров 1, 2, 3, 4 и 5. После дообучения этой модели были получены пять новых моделей, адаптированных для актёров из тестового набора. Для всех моделей внимания использовался Adam оптимизатор со скоростью обучения 0,0001 и категориальной функцией кросс-энтропийных потерь. Все нейросетевые классификаторы обучались 20 эпох. Персонализированная модель логистической регрессии дообучалась в течение 70 эпох, в то время как 50 эпох было достаточно для получения хорошего результата у моделей внимания.

Средняя точность классификаторов представлена в табл. 3.

Классификаторы, где для извлечения признаков в качестве основы использовалась сеть EmotiEffNet, показали более высокую точность в сравнении с моделью, где использовался MobileEmotiFace. Лучшая точность в 77,62 % для универсальной модели была получена при использовании EmotiEffNet-B0 для извлечения признаков и self-attention модели для классификации эмоций. После дообучения моделей под конкретных пользователей точность значительно возросла и в некоторых случаях достигла 100 %.

На рис. 3, 4 представлены матрицы перепутывания (англ. confusion matrix) для наиболее точных результатов, полученных при использовании EmotiEffNet-B0 для извлечения визуальных признаков и self-attention модели для классификации эмоций. На матрице для универсальной модели видно, что модель иногда делает ложные предсказания, в то время как персональная модель демонстрирует отличную точность распознавания выражения лица.

Табл. 3. Точность (%) распознавания эмоций по видеоизображению лица

СНС для извлечения признаков лиц	Классификатор	Универсальная модель	Персональная модель
MobileEmotiFace	Logistic regression	71,9	90,95
	Single attention	**74,76**	**100**
	Self-attention	73,81	**100**
EmotiEffNet-B0	Logistic regression	73,65	91,43
	Single attention	72,59	99,05
	Self-attention	**77,62**	**100**
EmotiEffNet-B2	Logistic regression	68,75	93,34
	Single attention	**69,99**	99,52
	Self-attention	68,09	**100**

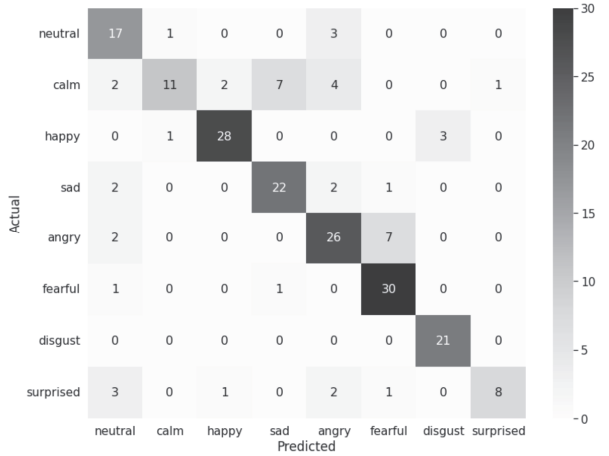

Рис. 3. Матрица перепутывания для универсальной модели, основанной на EmotiEffNet-B0 и self-attention

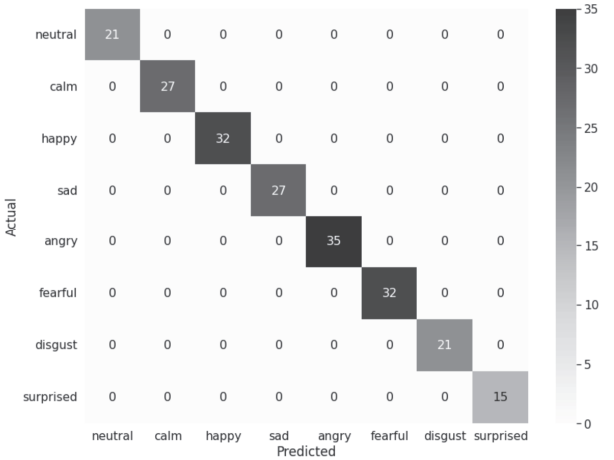

Рис. 4. Матрица перепутывания для персональной модели, основанной на EmotiEffNet-B0 и self-attention

Чтобы проверить эффективность режима проверки эмоций пользователем и влияние числа видео в тренировочном наборе на точность универсальной модели, был проведён ряд экспериментов. В этих экспериментах проводилось дообучение универсальной модели на различном числе видео с эмоциями актёра. На рис. 5 представлены модели для признаков эмоций, извлечённых с помощью EmotiEffNet-B0. По вертикальной оси отображается среднее арифметическое оценок точности, полученных во время распознавания эмоций, а по горизонтальной оси – число видео с эмоциями пользователя. Значению 0 здесь соответствует универсальная модель. Значение 1 означает, что тренировочный набор данных содержал только один видеофайл для каждой из эмоций, значение 2 – два видеофайла для каждой из эмоций и т.д. Таким образом, общее число видеофайлов в тренировочном наборе может быть рассчитано по формуле: $M \times C$, где M – количество видео каждой эмоции, C – общее число эмоций.

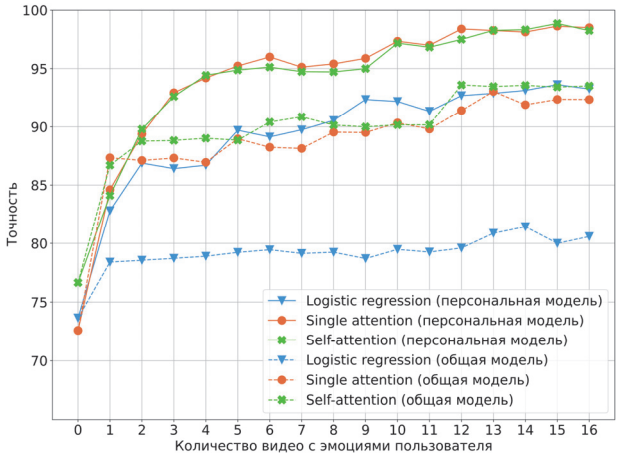

Рис. 5. Точность персональной и общей моделей, основанных на EmotiEffNet-B0, в зависимости от числа видео, использованных для дообучения классификатора

Кроме того, для сравнения представлены результаты экспериментов по дообучению универсальной модели на всём наборе видео пользователей из тестового набора, разбитых случайным образом. Вместо обучения персональных моделей для каждого из пользователей, обучалась одна общая модель, используемая для распознавания выражения лиц актёров из тестовой выборки, при этом размеры тренировочного и тестового наборов полностью совпадали с размерами этих наборов для экспериментов с персональными моделями. Пунктирные графики на рис. 5 отображают данные, полученные для общей модели.

На рисунке пунктирными линиями обозначены графики для общей модели и сплошными – для персональной модели. Здесь практически сразу при адаптации модели под пользователя происходит скачок точности. В среднем, после обучения на одном видеофайле для каждой эмоции точность распознавания эмоций возрастает на 11 % в сравнении с универсальной моделью. Для классификатора логистической регрессии общей модели точность увеличивается не так значительно, всего на 6 %. Точность классификатора логистической регрессии при одинаковом числе видеофайлов, используемых в дообучении модели, в среднем хуже точности моделей внимания на 5 % для персональной модели и на 12 % для общей модели.

Для персональной модели с использованием логистической регрессии значительное увеличение точности прекращается после того, как тестовый набор содержит более 10 видео для каждой из эмоций. Добавление новых видео в тестовый набор улучшает точность модели, но увеличение точности после каждого добавленного набора видео составляет менее 1 %. Общая модель логистической регрессии не демонстрирует значительного увеличения точности, в зависимости от добавления новых видео. В среднем точность изменяется менее чем на 1 %. Для моделей внимания последний рост более чем на 1 % фиксируется на 12 шаге у персональной модели, когда тестовый набор содержит 12 видео для каждой из эмоций, и на 13 шаге у общей модели. Затем рост точности также продолжается, но уже медленнее, менее 1 %.

Из графиков также видно, что персональные модели демонстрируют более высокую точность в распознавании выражений лиц пользователя, в сравнении с общей моделью, а также им необходимо меньше данных для получения хорошего результата. Кроме того, создание персональных моделей для каждого пользователя и использование модели для идентификации пользователя выглядит предпочтительнее и с точки зрения безопасности данных каждого пользователя, так как в предложенном подходе не требуется пересылать персональные видео на внешний сервер.

Заключение

В рамках данной работы реализован и верифицирован подход по высокоточному распознаванию эмоций лиц на видео (рис. 1), который может быть полезен не только для одного пользователя, но и в многопользовательских системах. При наличии небольшого количества размеченных видео с эмоциями конкретного пользователя может быть достигнуто значительное увеличение точности модели. Режим проверки результатов распознавания эмоций моделью будет полезен для новых пользователей, у которых нет размеченных данных. Тогда пользователь, используя этот режим, может оперативно выбрать необходимое число видео, которые будут использоваться для создания персональной модели. Результаты экспериментов (табл. 3) демонстрируют, что использование предложенных адаптивных моделей для распознавания пользовательских эмоций позволяет добиться улучшения точности более, чем на 20 % в сравнении с универсальной моделью. Наибольшую точность позволяют получить современные модели внимания.

Предложенная адаптация классификатора для распознавания эмоций целевого пользователя по видео не требует большого числа вычислительных ресурсов и может быть выполнена практически на любом пользовательском устройстве. Кроме того, нет необходимости в большом наборе размеченных данных для пользователя, под которого должна быть дообучена универсальная модель. Экспериментально продемонстрировано (рис. 5), что значительный рост точности распознавания эмоций наблюдается до того момента, пока тренировочный набор, используемый при адаптации модели, содержит до 5 видеофайлов для каждой из эмоции включительно, а затем рост точности замедляется.

Известно, что дополнительные сложности в распознавании выражения лица появляются в тех случаях, когда лицо может быть закрыто каким-то объектом, например, солнцезащитными очками или маской. Например, в статье [37] группа экспертов и наблюдателей распознавала выражения лиц людей с открытыми лицами, лицами, закрытыми маской, и лицами, закрытыми очками. Точность распознавания для частично закрытых лиц значительно снижалась. Кроме того, такие параметры, как варьирующееся освещение, возраст и положение лица, оказывают значительное влияние на точность работы алгоритмов распознавания [38]. Например, если набор данных, на котором обучалась модель, содержал только лица людей в возрасте от 20 до 40 лет, то нейросетевая модель будет хуже работать для более старшего поколения и детей. Предлагаемый подход позволяет в значительной степени преодолеть указанные проблемы, если условия съёмки и характеристики лица (форма, возраст, пол, причёска, наличие или отсутствие бороды, очки и т.п.) в собранном для дообучения наборе видео каждого пользователя совпадают с условиями, в которых разработанная система будет использоваться этим пользователем. Тем не менее одной из задач будущих исследований является проведение дополнительных экспериментов для варьирующихся освещенности и наличия помех.

В будущем планируется добавление новых модальностей к этому алгоритму, а именно дополнительно распознавать эмоции по речи пользователя, что должно увеличить точность распознавания и помочь в тех случаях, когда лица человека нет в кадре [39].

Благодарности

Работа выполнена при поддержке Российского научного фонда (проект № 20-71-10010).

References

[1] Ekman P. Basic emotions. In Book: Dalgleish T, Power MJ, eds. Handbook of cognition and emotion. New York: John Wiley & Sons; 1991: 45-60. DOI: 10.1002/0470013494.ch3.

[2] Livingstone SR, Russo FA. 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; 13(5): e0196391. DOI: 10.1371/journal.pone.0196391.

[3] Mollahosseini A, Hasani B, Mahoor MH. AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 2017; 10(1): 18-31. DOI: 10.1109/TAFFC.2017.2740923.

[4] Chang WY, Hsu SH, Chien JH. FATAUVA-Net: An integrated deeplearning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. in: 2017 IEEE Conf on Computer Vision and Pattern Recognition Workshops (CVPRW) 2017: 17-25. DOI: 10.1109/CVPRW.2017.246.

[5] Arunnehru J, Kalaiselvi Geetha M. Automatic human emotion recognition in surveillance video. In Book: Dey N, Santhi V, eds. Intelligent techniques in signal processing for multimedia security. Cham: Springer International Publishing Switzerland; 2017: 321-342. DOI: 10.1007/978-3-319-44790-2_15.

[6] Lee KW, Yoon HS, Song JM, Park KR. Convolutional neural network-based classification of driver's emotion during aggressive and smooth driving using multi-modal camera sensors. Sensors 2018; 18(4): 957. DOI: 10.3390/s18040957.

[7] Scherr SA, Elberzhager F, Holl K. Acceptance testing of mobile applications-automated emotion tracking for large user groups. 2018 IEEE/ACM 5th Int Conf on Mobile Software Engineering and Systems (MOBILESoft) 2018: 247-251.

[8] Naas SA, Sigg S. Real-time emotion recognition for sales. 2020 16th Int Conf on Mobility, Sensing and Networking (MSN) 2020: 584-591. DOI: 10.1109/MSN50589.2020.00096.

[9] Tkalcic M, Kosir A, Tasic J. Affective recommender systems: The role of emotions in recommender systems. The RecSys 2011 Workshop on Human Decision Making in Recommender Systems 2011: 9-13.

[10] Liu X, Xie L, Wang Y, Zou J, Xiong J, Ying Z, Vasilakos AV. Privacy and security issues in deep learning: A survey. IEEE Access 2020; 9: 4566-4593. DOI: 10.1109/ACCESS.2020.3045078.

[11] Savchenko AV. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. 2021 IEEE 19th Int Symposium on Intelligent Systems and Informatics (SISY) 2021: 119-124. DOI: 10.1109/SISY52375.2021.9582508.

[12] Savchenko AV, Savchenko LV, Makarov I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Trans Affect Comput 2022; 13(4): 2132-2143. DOI: 10.1109/TAFFC.2022.3188390.

[13] Churaev E, Savchenko AV. Touching the limits of a dataset in video-based facial expression recognition. 2021 Int Russian Automation Conf (RusAutoCon) 2021: 633-638. DOI: 10.1109/RusAutoCon52004.2021.9537388.

[14] Bagheri E, Bagheri A, Esteban PG, Vanderborgth B. A novel model for emotion detection from facial muscles activity. In Book: Silva MF, Lima JL, Reis LP, Sanfeliu A, Tardioli D, eds. Robot 2019: Fourth Iberian robotics conference. Cham: Springer; 2019: 237-249. DOI: 10.1007/978-3-030-36150-1_20.

[15] Luna-Jiménez C, Grio D, Callejas Z, Kleinlein R, Montero JM, Fernández-Martínez F. Multimodal emotion recogni-

tion on RAVDESS dataset using transfer learning. Sensors 2021; 21(22): 7665. DOI: 10.3390/s21227665.
[16] Churaev E, Savchenko AV. Multi-user facial emotion recognition in video based on user-dependent neural network adaptation. 2022 VIII Int Conf on Information Technology and Nanotechnology (ITNT) 2022: 1-5. DOI: 10.1109/ITNT55410.2022.9848645.
[17] Savchenko L, Savchenko AV. Speaker-aware training of speech emotion classifier with speaker recognition. In Book: Karpov A, Potapova R, eds. Speech and Computer. Cham: Springer Nature Switzerland AG; 2021: 614-625. DOI: 10.1007/978-3-030-87802-3_55.
[18] Li CJ, Spigner M. Partially speaker-dependent automatic speech recognition using deep neural networks. Journal of the South Carolina Academy of Science 2021; 19(2): 93-99.
[19] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Trans Pattern Anal Mach Intell 2001; 23(6): 681-685. DOI: 10.1109/34.927467.
[20] Shan C, Gong S, McOwan PW. Facial expression recognition based on local binary patterns: A comprehensive study. Image Vis Comput 2009; 27(6): 803-816. DOI: 10.1016/j.imavis.2008.08.005.
[21] Wang Z, Ying Z. Facial expression recognition based on local phase quantization and sparse representation. 2012 8th Int Conf on Natural Computation 2012: 222-225. DOI: 10.1109/ICNC.2012.6234551.
[22] Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. Int Conf on Machine Learning 2019: 6105-6114.
[23] Capotondi A, Rusci M, Fariselli M, Benini L. CMix-NN: Mixed low-precision CNN library for memory-constrained edge devices. IEEE Trans Circuits Syst II Express Briefs 2020; 67(5): 871-875. DOI: 10.1109/TCSII.2020.2983648.
[24] Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recogn Lett 2021: 141; 61-67. DOI: 10.1016/j.patrec.2020.07.042.
[25] Lomotin K, Makarov I. Automated image and video quality assessment for computational video editing. In Book: van der Aalst WMP, Batagelj V, Ignatov DI, Khachay M, Koltsova O, Kutuzov A, Kuznetsov SO, Lomazova IA, Loukachevitch N, Napoli A, Panchenko A, Pardalos PM, Pelillo M, Savchenko AV, Tutubalina E, eds. Analysis of images, social networks and texts. Cham: Springer Nature Switzerland AG; 2021: 243-256. DOI: 10.1007/978-3-030-72610-2_18.
[26] Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 2016; 23(10): 1499-1503. DOI: 10.1109/LSP.2016.2603342.
[27] Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A. VGGFace2: A dataset for recognising faces across pose and age. 2018 13th IEEE Int Conf on Automatic Face & Gesture Recognition (FG 2018) 2018: 67-74. DOI: 10.1109/FG.2018.00020.
[28] Barsoum E, Zhang C, Ferrer CC, Zhang Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. ICMI '16: Proc 18th ACM Int Conf on Multimodal Interaction 2016: 279-283. DOI: 10.1145/2993148.2993165.
[29] Meng D, Peng X, Wang K, Qiao Y. Frame attention networks for facial expression recognition in videos. 2019 IEEE Int Conf on Image Processing (ICIP) 2019: 3866-3870. DOI: 10.1109/ICIP.2019.8803603.
[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems 2017: 6000-6010.
[31] Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: Additive angular margin loss for deep face recognition. 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2019: 4690-4699. DOI: 10.1109/CVPR.2019.00482.
[32] Jaratrotkamjorn A, Choksuriwong A. Bimodal emotion recognition using deep belief network. 2019 23rd Int Computer Science and Engineering Conf (ICSEC) 2019: 103-109. DOI: 10.1109/ICSEC47112.2019.8974707.
[33] Alshamsi H, Kepuska V, Alshamsi H, Meng H. Automated facial expression and speech emotion recognition app development on smart phones using cloud computing. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conf (IEMCON) 2018: 730-738. DOI: 10.1109/IEMCON.2018.8614831.
[34] Rzayeva Z, Alasgarov E. Facial emotion recognition using convolutional neural networks. 2019 IEEE 13th Int Conf on Application of Information and Communication Technologies (AICT) 2019: 1-5. DOI: 10.1109/AICT47866.2019.8981757.
[35] He Z, Jin T, Basu A, Soraghan J, Di Caterina G, Petropoulakis L. Human emotion recognition in video using subtraction pre-processing. ICMLC '19: Proc 2019 11th Int Conf on Machine Learning and Computing 2019: 374-379. DOI: 10.1145/3318299.3318321.
[36] Baltrušaitis T, Robinson P, Morency LP. OpenFace: An open source facial behavior analysis toolkit. 2016 IEEE Winter Conf on Applications of Computer Vision (WACV) 2016: 1-10. DOI: 10.1109/WACV.2016.7477553.
[37] Noyes E, Davis JP, Petrov N, Gray KL, Ritchie KL. The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. Royal Soc Open Sci 2021; 8(3): 201169. DOI: 10.1098/rsos.201169.
[38] Bhattacharya S, Gupta M. A survey on: Facial emotion recognition invariant to pose, illumination and age. 2019 Second Int Conf on Advanced Computational and Communication Paradigms (ICACCP) 2019: 1-6. DOI: 10.1109/ICACCP.2019.8883015.
[39] Savchenko AV. Personalized frame-level facial expression recognition in video. In Book: Yacoubi ME, Granger E, Yuen PC, Pal U, Vincent N, eds. Pattern recognition and artificial intelligence. Cham: Springer Nature Switzerland AG; 2022: 447-458. DOI: 10.1007/978-3-031-09037-0_37.

Сведения об авторах

Чураев Егор Николаевич, 1993 года рождения, в 2016 году окончил Нижегородский государственный технически университет им. Р.Е. Алексеева по специальности 01.04.02 «Прикладная математика и информатика», в 2020 году поступил в аспирантскую школу по компьютерным наукам НИУ ВШЭ. Область научных интересов: компьютерное зрение, распознавание эмоций, машинное обучение, компиляторы, оптимизации алгоритмов.
E-mail: *echuraev@hse.ru* .

Савченко Андрей Владимирович, 1985 года рождения, в 2008 году окончил Нижегородский государственный технический университет им. Р.Е. Алексеева по специальности «Прикладная математика и информатика». В 2010 году защитил диссертацию на соискание ученой степени кандидата технических наук по специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ». В 2015 г. присвоено ученое звание доцента по специальности 05.13.18. В 2016 году присуждена учёная степень доктора технических наук по специальности 05.13.01 «Системный анализ, управление и обработка информации». В настоящее время работает профессором кафедры информационных систем и технологий и ведущим научным сотрудником лаборатории алгоритмов и технологий анализа сетевых структур в Национальном исследовательском университете Высшая школа экономики – Нижний Новгород. В настоящее время является научным директором лаборатории искусственного интеллекта Сбера. Автор более 100 научных работ. Область научных интересов: обработка мультимедийной информации, распознавание образов. E-mail: *avsavchenko@hse.ru*.

ГРНТИ: 28.23.15
Поступила в редакцию 30 декабря 2022 г. Окончательный вариант – 15 апреля 2023 г.

Facial expression recognition based on adaptation of the classifier to videos of the user

E.N. Churaev [1], A.V. Savchenko [1,2]
[1] HSE University, Laboratory of Algorithms and Technologies for Networks Analysis, 603093, Nizhny Novgorod, Russia, Rodionova 136;
[2] Sber AI, 121170, Moscow, Russia, Kutuzovsky prospekt 32, building 2

Abstract

In this paper, an approach that can significantly increase the accuracy of facial emotion recognition by adapting the model to the emotions of a particular user (e.g., smartphone owner) is considered. At the first stage, a neural network model, which was previously trained to recognize facial expressions in static photos, is used to extract visual features of faces in each frame. Next, the face features of video frames are aggregated into a single descriptor for a short video fragment. After that a neural network classifier is trained. At the second stage, it is proposed that adaptation (fine-tuning) to this classifier should be performed using a small set of video data with the facial expressions of a particular user. After emotion classification, the user can adjust the predicted emotions to further improve the accuracy of a personal model. As part of an experimental study for the RAVDESS dataset, it has been shown that the approach with model adaptation to a specific user can significantly (up to 20 – 50 %) improve the accuracy of facial expression recognition in the video.

Keywords: facial expression classification, neural network classifier adaptation, speaker-dependent emotion recognition.

Citation: Churaev EN, Savchenko AV. Facial expression recognition based on adaptation of the classifier to videos of the user. Computer Optics 2023; 47(5): 806-815. DOI: 10.18287/2412-6179-CO-1269.

Acknowledgements: This work was supported by the Russian Science Foundation under RSF grant No. 20-71-10010.

Authors' information

Egor Nikolaevich Churaev (b. 1993) graduated from N. Novgorod State Technical University in 2016, majoring in Applied Mathematics and Informatics. In 2020, he entered the graduate school in Computer Science at the National Research University Higher School of Economics, Nizhny Novgorod. Research interests include computer vision, emotion recognition, machine learning, compilers, algorithm optimizations E-mail: *echuraev@hse.ru*.

Andrey Vladimirovich Savchenko, (b. 1985), graduated from N. Novgorod State Technical University in 2002, majoring in Applied Mathematics and Informatics. He defended his PhD in Mathematical Modeling, Numeric Methods and Software Complexes in 2010. He received the Doctor of Science degree in System Analysis, Control and Information Processing in 2016. Currently he works as the professor of Information Systems and Technologies department and leading researcher of the laboratory of Algorithms and Technologies in Network Analysis in National Research University Higher School of Economics, Nizhny Novgorod. Currently he holds the position of scientific director of Sber AI Lab. He is the co-author of more than 100 scientific papers. Research interests include multimedia processing and pattern recognition. E-mail: *avsavchenko@hse.ru*.

Received December 30, 2022. The final version – April 15, 2023.