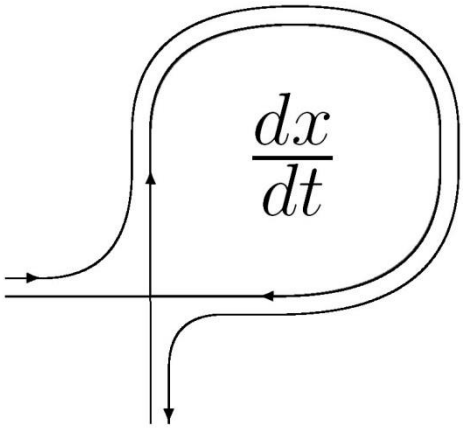

ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ И ПРОЦЕССЫ УПРАВЛЕНИЯ
N. 3, 2023
Электронный журнал,
рег. Эл № ФС77-39410 от 15.04.2010
ISSN 1817-2172

http://diffjournal.spbu.ru/
e-mail: jodiff@mail.ru

Машинное обучение

Построение эмоционального образа человека на основе анализа особых точек в последовательных кадрах видеоряда

Аверьянов Д.Д. [1,*], Желудев М.В.[2,**], Кияев В.И.[3,***]

[1]ООО "Роберт Бош
[2]ООО "Роберт Бош"
[3]Санкт-Петербургский государственный университет

[*]dmitryaverianov@gmail.com
[**]mikhail.zheludev@ru.bosch.com
[***]kiyaev@mail.ru

Аннотация. Работа посвящена разработке алгоритма классификации поведения человека в контексте детектирования правдивости или лживости высказываний, представленных в формате видеофайлов. Анализ видеофайла проводился в рамках временного окна, в котором анализировались как изменения в микромоторике лицевых мускулов, так и речевые признаки. В нашем случае мимика отражается математическим представлением в виде вектора, содержащего необходимую цифровую информацию о состоянии лица, которое характеризуется положениями особых точек (ключевых точек носа, бровей, глаз, век и т. д.). Вектор мимики формируется в результате обучения нелинейных моделей. Вектор, характеризующий речь, формируется на основе эвристических характеристик звукового сигнала. Темпоральную агрегацию векторов для финальной классификации поведения производит отдельная нейронная сеть. В работе приведены результаты точности и быстродействия алгоритма, которые показывают, что новый подход конкурентоспособен по отношению к существующим методам.

Ключевые слова: классификация видео; детектор лжи; трансформеры; лицевые ориентиры, речевой сигнал; анализ аудио; видеоаналитика; машинное и глубокое обучение.

1. Введение

Переход к цифровой экономике однозначно предполагает широкое использование цифровых технологий – в частности, широкое применение методов искусственного интеллекта (Artificial Intelligence), машинного и глубокого обучения (Machine and Deep Learning). Методы искусственного интеллекта (AI) предполагают применение любых компьютерных методов и технологий, с помощью которых можно моделировать и имитировать работу головного мозга человека, используя современные сложные алгоритмы распознавания текстов, звуков, образов. К ним можно также отнести интеллектуальные методы формирования и принятия решений, способы управления группами робототехнических устройств, выполняющих общую задачу в условиях неопределенности. Машинное обучение (ML) и глубокое обучение (DL) – это совокупность методов машинного обучения на многослойных нейросетях, основанных на обучении не по специализированным алгоритмам, разработанным под конкретные задачи, а по имеющимся общим признакам и представлениям. Эти признаки и представления конкретизируются и уточнятся в ходе обучения и в дальнейшем используются в качестве базовой основы для подготовки и в некоторых случаях реализации необходимого решения.

Нейронные сети успешно применяются для отождествления человека по системе характерных точек на его лице. Такие системы широко используются в охранных и мониторинговых системах. В Москве и Санкт-Петербурге в системах видеонаблюдения в постоянном режиме работает около трехсот тысяч видеоустройств с различной функциональностью и обладающих достаточно высоким разрешением. Такие устройства с точностью до 90% позволяют выделить из насыщенного человеческого потока лицо, которое объявлено в розыск. При этом процесс «Захват - Выделение - Анализ конфигурации особых точек - Отождествление» происходит на основе статического изображения, выхваченного из потока лиц. В этом случае решается задача типа «Да – Нет» с заданным уровнем вероятности.

Гораздо сложнее обстоит дело, если на основании серии наблюдений нужно построить образ, отражающий текущее эмоциональное состояние человека и сформировать модель развития этого состояния. В некоторых случаях такая оценочная модель, построенная в режиме реального времени, была бы чрезвычайно полезной в ситуациях, когда от человека требуется большое или максимальное напряжение сил при выполнении ответственной работы – например, выполнения задания в сложных условиях с трудно предсказуемым развитием событий. По нашему мнению, в таких ситуациях интеллектуальная система мониторинга видеонаблюдения с последующим покадровым анализом видеоряда на основе соответствующим образом натренированных нейронных сетей позволит с достаточной степенью точности оценить текущее эмоциональное состояние человека, смоделировать эмоциональный образ и прогнозировать развитие ситуации. Обучение сетей можно проводить на последовательном покадровом видеоряде, в котором соответствующим образом подготовленные актеры реализуют различные эмоциональные состояния (радость, безразличие, равнодушие, усталость, раздражение, страх и т. д.) и при этом с высокой степенью точности специальные видеоустройства отслеживают изменение микромоторики лицевых мускулов, движений и размера зрачков глаз и оттенков речи. Покажем возможности такого подхода на частном случае определения, говорит ли человек правду или ложь.

Использование физиологических параметров человека довольно давно применяется при определении его эмоционального состояния – в частности, при определении говорит ли человек правду или, мягко говоря, лжёт. Однако практика показывает, что специальным образом подготовленные люди, актеры или люди с низким IQ способны «обойти» такую систему. В отличие от такой, физиологической модели детектора лжи система, основанная на цифровом покадровом анализе микромоторики лицевых мускулов и оттенков речи, с высокой степени вероятности может подтвердить достоверность высказывания человека или объявить его высказывания недостоверными.

Цифровой детектор лжи – востребованная технология, отвечающая запросам современного общества, таким как совершенствование охранных систем, разработка цифрового переговорщика и прочих [1]. Распознавание правдивости или ложности высказываний является частным случаем распознавания текущего поведения человека [2]. Обычно такие задачи решаются путем векторизации изображений в текущий момент времени, а затем агрегированием этой векторизованной информации внутри временного окна [5]. В настоящей работе будем придерживаться этой стандартной парадигмы и предложим новый метод векторизации, основанный на особых точках лица.

1.1. Подходы к описанию речевого сигнала

Существуют подходы позволяющие описать речевой сигнал при помощи аудио признаков. Чаще всего для анализа используются простые просодические признаки – такие как изменение частоты, громкости, темпа голоса [1]. Значения признаков изменяются при стрессе, однако, не позволяют однозначно отделить ложь от неуверенности. Исходя из этого, для более точной классификации требуются иные признаки при описании речевого сигнала, такие как уровень частотной дрожи, контроль над вибрациями голосовых связок [2]. Данные параметры более полно характеризует частотные изменения речи.

В работе [3] упоминается, что в случае, когда человек врёт спонтанно, он будет сопровождать свой ответ длительными паузами, что будет приводить к увеличению времени ответа в то время, как заученные ложные фразы, наоборот, имеют наискорейший отклик. Таким образом, паузы в речи так же являются признаком, так как свидетельствуют о затруднениях при ответе на вопрос.

1.2. Получение вектора признаков изображения лица при анализе видео

Анализ мимики и жестов на видео полезен для оценки истинности высказываний человека, так как, когда человек лжет, он испытывает стресс, а его движения перестают быть естественными [1]. Для распознавания мимики используются лицевые ориентиры – ключевые точки, релевантные для определения формы лица, поворота головы, положения и движений глаз и губ [1]. Тем не менее, недостаточно просто отслеживать движения ориентиров лица. Внешние факторы, такие как яркий свет, ветер или холод, могут спровоцировать изменение мимики [1]. При распознавании лицевых ориентиров стоит ориентироваться не на факт их смещения, а на продолжительность и амплитуду движений точек лица. Это обусловлено тем, что микромоторика лицевых мускулов часто не осознается человеком, ее появление может быть непредсказуемым и кратким, поэтому системы распознавания мимической активности направлены не на распознавание движений ориентиров лица, а фиксацию быстрой смены мимики [1].

В ситуациях, когда информация о перемещении всех ориентиров лица избыточна, используется фокусное отслеживание зон лица [5]. Отслеживание направления взгляда позволяет определить, в каких зонах сосредоточено внимание человека, а в какие, наоборот, взгляд человека не направляется. Методы компьютерного зрения позволяют извлекать из видео ориентиры лица независимо от поворота головы. Увеличение репрезентативности обучающего набора данных снижает влияние факторов фона и освещения на выделение признаков лица [7].

Эффективным способом распознавания ориентиров лица человека является сверточная нейронная сеть BlazeFace [8], которая используется как для детекции лиц, так и для определения ключевых точек. В задаче обнаружения лица человека сеть BlazeFace извлекает признаки из изображения лица размером 128×128 пикселей [8]. Изменения особенностей лицевой поверхности интерпретируются как мимические движения.

Таким образом, перечислена информация, которая будет использоваться для дальнейшей векторизации видеоконтента, включающая как двумерный контент изображения, так и одномерный контент звуков аудиозаписи.

1.3. Методы анализа изменений признаков во времени

Рекуррентные нейронные сети (Recurrent neural network – RNN), такие как сеть долгой краткосрочной памяти (Long short-term memory – LSTM), долгое время были классическими методами анализа и классификации различных последовательностей [16]. В таких моделях каждый элемент рассматриваемой последовательности анализируется только в контексте предшествующих ему элементов, так что ближайшие к нему элементы по времени имеют большую значимость по отношению к более удаленным. Данный принцип работы модели может подходить для анализа последовательности слов в предложении или предсказании траектории движения объекта на видео. В задаче классификации видеопоследовательности такой подход может не позволить сделать корректное предсказание, так как имеет свойство быстро забывать начало последовательности и отдавать большее значение последним кадрам.

Имеются различные подходы для уменьшения влияния подобных проблем, например использование двунаправленных LSTM-моделей или добавление механизма внимания [16, 18], однако они не позволяют решить эти проблемы полностью. Существует подход, предложенный Facebook Research, позволяющий устранить вышеперечисленные проблемы и продолжить эффективно использовать рекуррентную нейронную сеть.[1] Данный метод заключается в использовании иерархии сверточных слоев в узлах RNN сети. Хотя данный подход позволяет производить параллельные вычисления, тем не менее согласно [10], вычислительные мощности экспоненциально зависят от точности получаемых результатов. Подход, позволяющий исправить все перечисленные противоречия, предложили Vaswani A. et. al. [18]. В основе предложенной архитектуры лежит механизм внимания, который лег в основу архитектуры трансформера [11]. Этот подход решает проблему забывания, так как каждый элемент последовательности получает информацию о любом другом. Кроме того, такой подход не требует последовательного выполнения и, соответственно, может выполняться параллельно, что делает трансформеры значительно более эффективными с точки зрения производительности.

[1] https://www.opennet.ru/opennews/art.shtml?num=49837

В таблице 1 приведено сравнение производительности трансформера, рекуррентной сети и сверточной сети.

Таб. 1. Сравнение сложности различных архитектур нейронных сетей для анализа последовательностей [12]

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(\log_k n)$
Self-Attention	$O(r \cdot n \cdot d)$	$O(1)$	$O\left(\frac{n}{r}\right)$

В классическом виде, предложенном авторами, трансформер используется для решения задачи на основе применения модели глубокого обучения Sequence-to-Sequence (seq2seq) [17]. Такой трансформер включает в себя энкодер, который получает на вход последовательность и преобразует ее в новую последовательность, и декодер, который выполняет обратную операцию применительно к задаче классификации.

Таким образом, выше перечислены подходы к анализу поведения на базе аудио- и видеоконтента, преимущества и недостатки существующих подходов. На основе этого предложен новый подход, устраняющий недостатки других подходов, перечисленных в нашем обзоре.

2. Материалы и методы

Разработанный подход предполагает сканирование видеопотока временным окном в 10 секунд, в результате чего у нас имеется 2 тензора информации. Первый из них относится к видеоконтенту и содержит признаки, характеризующие видимое состояние лица говорящего, а второй – аудиопризнаки, характеризующие интонацию (оттенки) голоса. Новизна этого подхода заключается в способе извлечения признаков видеоконтента. Тензоры, полученные с каждого кадра внутри временного окна, соединяются с аудиопризнаками, образующими вектор, далее они агрегируются вдоль шкалы времени нейросетью на базе новой архитектуры трансформера [11].

2.1. Извлечение вектора признаков лица

Предлагается подход, где вектор признаков лица состоит из двух компонент, первая из которых – вектор мимики лица, полученный обработкой особых точек, а вторая это углы поворота головы рисунок 1. Для получения векторов признаков лица, содержащего информацию как о мимике, так и о положении головы, была использована сверточная нейронная сеть BlazeFace [8], предназначенная для обнаружения 478 лицевых ориентиров человека (или особых точек) в трехмерном пространстве [8].

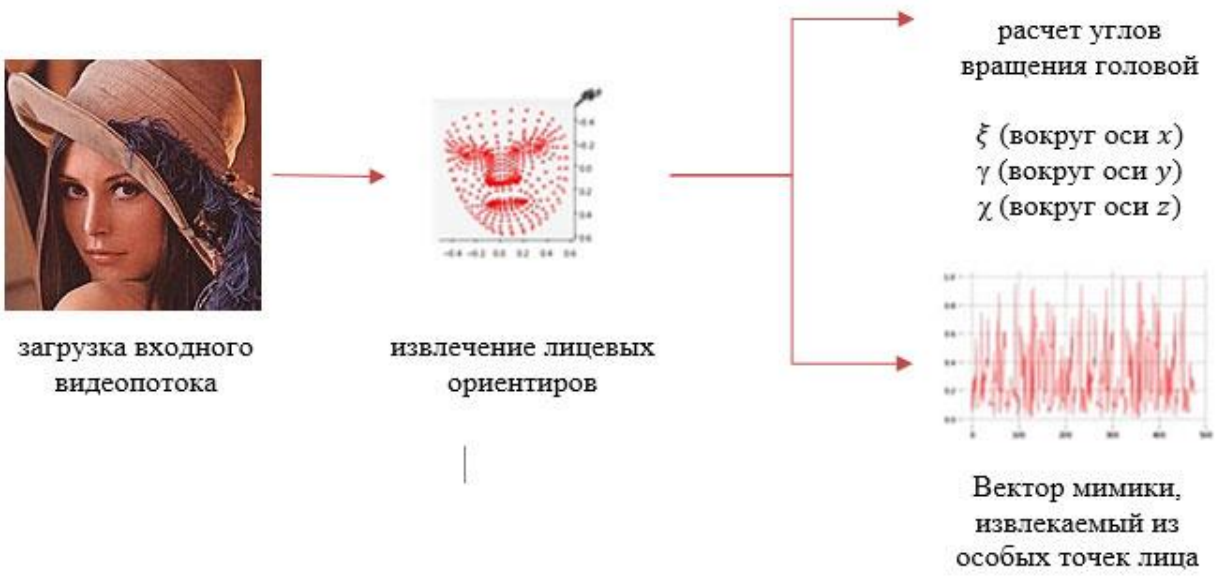

Рис. 1. Схема извлечения признаков лица, в которой на основе особых точек строится вектор мимики, состоящий из 478 признаков и 3 углов поворота головы

Выбор данного решения обусловлен трехмерностью пространства точек, в отличие от других решений, предлагающих двумерные точки. Определение пространственных координат лицевых ориентиров вокруг осей *X,Y,Z* позволяет строить карту глубины изображения для определения рельефа лица, поворота и наклона головы человека. Прогнозирование Z координат позволяет избежать ситуаций, когда лицо частично закрыто.

На основе топологии особых точек лица рассчитывается общий вектор признаков лица.

Пусть

$$\{x_1, x_2, \dots, x_n\} \in R^3 \tag{1}$$

подмножество трехмерного пространства, содержащее лицевые ориентиры.

Тогда

$$Z = \sum_{i=1}^{n} x_i \tag{2}$$

центр данного множества.

Определим упорядоченное множество

$$S = \{y_1, y, \dots, y_n\},$$
где $y_i = x_i - Z.$ (3)

Вектор

$$M = [|y_1|, |y_2|, \dots, |y_n|] \tag{4}$$

назовем вектором мимики, поскольку он содержит информацию о взаимном расположении особых точек, или лицевых ориентиров.
Определим далее вектор

$$\|M\| = \frac{M}{|M|} \tag{5}$$

который обладает следующими свойствами:

1. $\|M\|$ инвариантен относительно ортогональных вращений вокруг точки Z.
2. $\|M\|$ инвариантен относительно масштаба топологии расположения лицевых ориентиров в пространстве R^3.
3. $\|M\|$ чувствителен к изменению относительного расположения лицевых ориентиров.

В силу перечисленных выше свойств, вектор будет называться$\|M\|$вектором мимики.
Рассмотрим теперь еще одно множество

$$R = \left\{\frac{y_1}{|y_1|}, \frac{y_2}{|y_2|}, \dots, \frac{y_n}{|y_n|}\right\} \tag{6}$$

где R является проекцией S на единичную сферу с центром в Z.
Пусть T - оператор ортогонального вращения пространства вокруг точки Z. Очевидно, что:

$$T(\|M\|) = \|M\| \tag{7}$$

для любого T.
И если

$$T(R) = R, \tag{8}$$

Таким образом имеется взаимно-однозначное соответствие между ортогональными вращениями и множествами $\{T(R) \vee R \in O(3)\}$.

2.2. Получение углов поворота головы

В вопросе детекции правдивости высказываний человеком на основе видеоаналитики, движения головой также являются невербальным сигналом помимо мимики [1], поэтому 3 угла поворота головы должны быть добавлены к вектору мимики, чтобы дополнить общий вектор признаков лица.

Данные углы поворота характеризуют вращение трехмерного пространства вокруг точки Z, которое определяется по множеству T(R).

Для этой цели была обучена полносвязная нейронная сеть на парах взаимно-однозначного соответствия:

$$(\boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\chi}) \leftrightarrow \boldsymbol{T}(\boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\chi}) \leftrightarrow \boldsymbol{R}(\boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\chi})(\boldsymbol{R}), \tag{9}$$

где исходное множество R строилось из особых точек 120 лиц, образующих репрезентативную выборку разного типа лиц, смотрящих прямо.
Параметры $\{\xi, \gamma, \chi\}$ генерировались случайным образом на основе нормального распределения. Для проверки эффективности применения углов вращения головой в задаче определения истинности высказывания, была использована рекуррентная нейронная сеть LSTM. Точность модели составила 0.64, что доказывает значимость выделенных нами признаков.

2.3. Извлечение вектора признаков речевого сигнала

В результате применения разработанного алгоритма извлекаются следующие признаки:

1. доля повышений/понижений громкости относительно среднего значения, средняя громкость фрагмента;
2. доля повышений/понижений тона относительно среднего значения, среднее значение частоты речи (тона) выбранного фрагмента;
3. уровень частотного дрожания (jitter) [2];
4. значение темпа (количество произносимых слов в минуту);
5. доля сигнала, не относящаяся к речи;
6. категориальный признак – тишина (значение положительное, если среднее значение громкости на интервале в 1 секунду меньше заданного уровня);
7. класс эмоции – 7 классов: злость, скука, тревога, радость, печаль, отвращение, отсутствие эмоций (набор данных Emo-DB [13]).

При анализе параметров громкости используется амплитудная огибающая, которая получается за счет прохода по записи окном заданного размера (по умолчанию 512 элементов вектора признаков) и выделением максимального значения в рамках этого окна.

При подсчете доли увеличения/понижения громкости в рамках заданного временного фрагмента считается количество этих фрагментов, отклоняющихся от среднего более чем на стандартное значение.

Фрагмент принимает метку «тишина», если модуль громкости в децибелах меньше среднего значения всей записи на 6 дБ (значение получено эмпирически).

Значения доли повышения и понижения частоты вычисляются аналогичным способом (за счет подсчета отклонений от среднего уровня).
Для извлечения класса эмоций по спектрограмме модель сверточной нейронной сети (CNN) с полносносвязным классификатором была обучена (на наборе данных Emo-DB), которая принимает на вход аудио-сигнал в виде мел-кепстральных коэффициентов (MFCC), которые сосредоточены на частотах, имеющих большое значение для человеческой речи и слуха [3, 4]. Класс эмоций извлекается каждую секунду, фрагменты меньшей длительности дополняются нулями до необходимого размера.

2.3. Агрегирование изменений признаков во времени и классификация потока признаков

В предыдущих подразделах были описаны методы преобразования видеоряда в поток векторов признаков. Полученные векторы содержат как аудио, так и визуальные признаки. Каждый из векторов описывает соответствующий кадр в отдельности, однако не содержит

информацию о других кадрах, о всей последовательности, каких-либо изменениях или переходных процессах.

Ранее было описано преимущество использования энкодера трансформера по сравнению с LSTM в качестве модели для изучения временной составляющей последовательности векторов признаков, поэтому для темпорального агрегирования признаков будет использоваться трансформер.

Для окончательной классификации видео, преобразуем векторы признаков, полученные из каждого кадра и описанные при получении углов поворота головы, в двумерный тензор. В этом двумерном векторе по строкам индексируется время, а по столбцам – извлеченные признаки, который далее обрабатывается архитектурой нейронной сети, называемой трансформером. Эта сеть преобразует двумерный тензор в вектор, который подается на классификатор, представляющий собой последовательность полносвязных слоев рисунок 2.

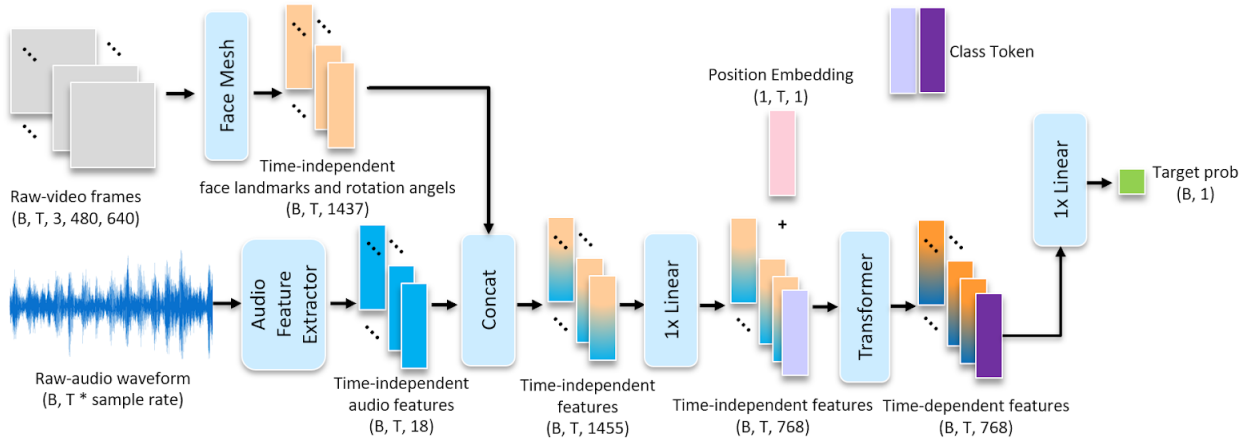

Рис. 2. Схема предлагаемого подхода

На вход подаются тензоры видеоряда, а также аудио сигнал. Из аудиосигнала каждую секунду извлекаются вектора признаков, которые конкатенируются с признаками лица. Далее суммарный вектор подается в трансформер, выход трансформера проводится через полносвязный модуль, после чего подается на классификатор.

3. Результаты

В обучающий набор включены записи группы людей в естественных условиях. Перед каждым из участников группы стояла задача высказать правдивые или ложные ответы на заранее предопределенные вопросы. База данных содержит серию размеченных видео, где записаны интервью. Всего проводилось 3 этапа интервью в среднем по 30 человек длительностью по 5 минут. Тестировали равное количество мужчин и женщин в возрасте между 21 и 33 для мужчин и 18 и 34 для женщин.

Для сравнения эффективности предлагаемого алгоритма классификации видеопоследовательности были проведены эксперименты с измерениями точности и объемом потребляемых ресурсов. Эти значения сравнивались с существующими решениями, представляющими некоторые эталоны (Top of the Art) в данной тематике [8].

В качестве альтернативы предложенному вектору признаков лица была рассмотрена предобученная на наборе данных ImageNet сверточная нейронная сеть ResNet50, которая извлекала признаки из обнаруженных на изображениях лиц. Для поиска и отслеживания

самих лиц использовалась предобученная модель TinaFace с основой ResNet50 с двумя Deformable блоками в конце, GroupNorm в качестве слоя нормализации и модулями Feature Pyramid Network и Inception, так как данная модель является одной из наиболее эффективных моделей для обнаружения лица с точки зрения как производительности, так и точности [14].

В качестве альтернативы классическому трансформеру были рассмотрены современные подходы к анализу и классификации видео, такие как использование трехмерных сверточных слоев и применение визуального трансформера. ResNet3d была выбрана как классический вариант модели с трехмерными сверточными слоями, а в качестве визуального трансформера был выбран TimeSformer, которая является передовым решением на данный момент [15]. Эти модели получали на вход изображения лиц, полученные с помощью модели TinaFace. ResNet3d использовался в конфигурации r3d с глубиной 18 блоков. TimeSformer использовался в конфигурации с 8 последовательными Divided Space-Time Attention блоками внимания, скрытым представлением размерностью 768 и полносвязными сетями размерностью 3072.

Итоговые значения точности, потребляемой памяти и времени работы, полученные в результате экспериментов, представлены в таблицах 2 и 3. По результатам можно увидеть, что наибольшую точность показывает TimeSformer, принимающий на вход изображения лиц и самостоятельно извлекающий из них признаки, а также использующий аудио признаки, итоговая точность модели составила 84,91%. Второй по точности является аналогичная модель без использования аудио признаков, ее точность составила 83,2%. Немного меньшую точность относительно данных моделей показывает предложенный метод использования точек лица, углов поворота лица и аудиальных признаков и трансформера – итоговая точность 83%.

Таб. 2. Сравнительная таблица точности классификации видеопоследовательности различными конфигурациями моделей в задаче бинарной классификации истина/ложь

<table>
<tr><th>Обнаружение и извлечение лица</th><th>Извлечение признаков из лица</th><th>Извлечение признаков относительно времени</th><th>Извлечение аудио признаков</th><th>Классификация</th><th>Точность (accuracy), %</th></tr>
<tr><td colspan="2">Точки лица и углы поворота</td><td>Transformer</td><td>—</td><td rowspan="6">Полносвязный слой</td><td>66,9</td></tr>
<tr><td colspan="2">Точки лица и углы поворота</td><td>Transformer</td><td rowspan="4">Громкость
Темп
Тон
Дрожь
Эмоции</td><td>83,0</td></tr>
<tr><td rowspan="3">TinaFace</td><td>ResNet50</td><td>Transformer</td><td>72,5</td></tr>
<tr><td colspan="2">ResNet3d</td><td>80,1</td></tr>
<tr><td colspan="2">TimeSformer</td><td>84,91</td></tr>
<tr><td>TinaFace</td><td colspan="2">TimeSformer</td><td>—</td><td>83,2</td></tr>
</table>

Таб. 3. Сравнительная таблица потребляемой памяти и времени выполнения при обучении и валидации на CPU

Конфигурация модели	Использование RAM при обучении, МБ	Использование RAM при валидации, МБ	Скорость при обучении, итер/сек	Скорость при валидации, итер/сек
Точки лица и углы поворота, Трансформер	5 500	2 300	1,1	1,1
Точки лица и углы поворота, Аудио, Трансформер	5 700	2 500	4,0	4,0
TinaFace, ResNet3d	5 800	4 000	224,95	174,8
TinaFace, TimeSformer	24 000	5 600	170,6	117,6
TinaFace,TimeSformer, Audio	24 200	5 800	173,5	119,9

Тестирование велось на оборудовании со следующими характеристиками:

- CPU: AMD EPYC (2nd Gen) 7702 Tetrahexaconta-core (64 Core, 3.35GHz);
- GPU: NVIDIA A100 SMX4 40Gb;
- RAM: 3600 Hz;
- SSD: M.2, 7 000 МБ/сек чтение, 5 300 МБ/сек запись;
- Batch size: 1;
- Временное окно: 100 видео-фреймов;
- Частота дискретизации аудио: 16 000 Hz.

В работе была предложена методика обработки видео с целью классификации видео- и звукового контента на предмет правдивости высказываний человека. В отличии от общепринятых подходов был предложен новый метод обработки особых точек лица, при которых информация разделялась на две составляющие: инвариантные относительно вращений и показатели самого вращения. Данный подход, как видно из приведенных выше таблиц, позволяет сократить вычислительные ресурсы при сохранении оптимального уровня точности. Отметим, что данное преимущество было показано на процессорах, но не на видеокартах, поскольку методика выявления особых точек ограничена решением, имеющим только процессорную реализацию.

Предлагаемый метод показывает сопоставимые с имеющимися (State-of-the-Art) решениями результаты. Однако наша методика требует значительно меньших вычислительных ресурсов как с точки зрения потребляемой памяти, так и с точки зрения времени выполнения на процессоре.

4. Заключение

В статье представлен новый подход к проблеме классификации поведения человека в контексте ситуации, когда во время интервью человек говорит правду или лжет. Подход основан на нескольких алгоритмах, в основе которых лежат модели глубокого обучения.

первая из них находит особые точки лица, вторая превращает их в вектор признаков, релевантный рассматриваемой задаче классификации, и последняя агрегирует эти признаки, получаемые каждую секунду, во времени. Было проведено тестирование и сравнение с State-of-the-Art подхода к этой задаче.

Показано, что точность на тестовой выборке составляет 84,91%, что приближается к State-of-the-Art методам анализа поведения, дающим точность 86%, однако предложенный метод намного меньше потребляет памяти, а также значительно превосходит в быстродействии – и это делает возможным успешно применить данный метод на мобильных платформах.

Дальнейшее развитие предлагаемой методики позволит, на наш взгляд, распознавать и использовать более тонкие эффекты анализа мимической микромоторики мускулов лица, поворотов головы, перемещения зрачков и изменения тональности звуковых сегментов речи для построения эффективной модели, отражающей эмоциональное состояние (эмоциональный образ) человека и прогнозировать развитие этого состояния. Такую методику можно использовать в различных прикладных задачах, в которых требуется профессиональное (производственное) тестирование и обязательный учет эмоциональных особенностей состояния человека.

5. Литература

[1] Goupil L. et al. Listeners’ perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature // Nature Communications. 2021. Vol. 12, № 1. P. 861.

[2] Teixeira J.P., Oliveira C., Lopes C. Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters // Procedia Technology. 2013. Vol. 9. P. 1112–1122.

[3] Burzo M. et al. Multimodal deception detection // The Handbook of Multimodal-Multisensor Interfaces, Volume 2. 2018.

[4] Chow A., Louie J. Detecting lies via speech patterns. 2017.

[5] Zhang, X., Sugano, Y., Fritz, M. & Bulling, A. 2017, "It's Written All over Your Face: Full-Face Appearance-Based Gaze Estimation", IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 2299.

[6] Kathi, M.G. & Shaik, J.H. 2021, "Estimating the smile by evaluating the spread of lips", Revue d'Intelligence Artificielle, vol. 35, no. 2, pp. 153-158.

[7] Zhang, X., Sugano, Y., Fritz, M. & Bulling, A. 2015, "Appearance-based gaze estimation in the wild", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4511.

[8] Bazarevsky, V. et. al., BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs, CoRR, abs/1907.05047. 2019.

[9] Kaiming H. et. al., Deep Residual Learning for Image Recognition, CVPR 2016, 2016.

[10] Bertatius G. et. al., Is Space-Time Attention All You Need for Video Understanding?, ICML 2021, 2021.

[11] Vaswani A. et. al., Attention Is All You Need, NIPS 2017, 2017.

[12] Gong Y., et. al., AST: Audio Spectrogram Transformer, Interspeech 2021, 2021.

[13] Burkhardt F. et al. A Database of German Emotional Speech // Interspeech. 2005. P. 1517–1520.

[14] Zhu Y., et. al., TinaFace: Strong but Simple Baseline for Face Detection, arXiv preprint arXiv:2011.13183, 2020.

[15] Tran D., et. al., A Closer Look at Spatiotemporal Convolutions for Action Recognition, CVPR 2018, 2018.

[16] Olah C., Understanding LSTM Networks // colah.github.io. 2015. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs (дата доступа 22.01.2023).

[17] Alammar J., Visualizing A Neural Machine Translation Model (Mechanics of Seq2Seq Models With Attention) // URL: https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/ (дата доступа 22.01.2023).

[18] Vaswani A., Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 2017

Construction of an emotional image of a person based on the analysis of key points in consecutive frames of a video sequence

Averianov D.D. [1,*], Zheludev M.V.[2,**], Kiyaev V.I.[3,***]

[1]OOO "Robert Bosch"
[2]OOO "Robert Bosch"
[3]Saint Petersburg State University

[*]dmitryaverianov@gmail.com
[**]mikhail.zheludev@ru.bosch.com
[***]kiyaev@mail.ru

Abstract. The work is devoted to the development of an algorithm for classifying human behavior in the context of detecting the truthfulness or falsity of statements presented in video file format. The analysis of the video file was carried out within the time window, in which both changes in the micromotility of the facial muscles and speech signs were analyzed. In our case, facial expressions are represented by a mathematical representation in the form of a vector containing the necessary digital information about the state of the face, which is characterized by the positions of special points (key points of the nose, eyebrows, eyes, eyelids, etc.). The mimic vector is formed as a result of training non-linear models. The speech characterizing vector is formed on the basis of the heuristic characteristics of the audio signal. The temporal aggregation of vectors for the final classification of behavior is performed by a separate neural network. The paper presents the results of the accuracy and speed of the algorithm, which show that the new approach is competitive with respect to existing methods.

Keywords: video classification; lie detector; transformers; facial landmarks, speech signal; audio analysis; video analytics; machine and deep learning.