УДК 004.032.26 DOI: 10.26102/2310-6018/2021.32.1.004

Метод распознавания эмоций человека по двигательной активности тела в видеопотоке на основе нейронных сетей

М.Ю. Уздяев, Д.М. Дударенко, В.Н. Миронов

Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Санкт-Петербург, Российская Федерация

Резюме: В данной статье рассматривается применение различных нейросетевых моделей для решения задачи распознавания эмоций человека по двигательной активности его тела на кадрах видеопотока без сложной предварительной обработки этих кадров. В работе представлены трехмерные сверточные нейронные сети: Inception 3D (I3D), Residual 3D (R3D), а также сверточно-рекуррентные нейросетевые архитектуры, использующие сверточную нейронную сеть архитектуры ResNet и рекуррентные нейросети архитектур LSTM и GRU (ResNet+LSTM, ResNet+GRU), которые не требуют предварительной обработки изображений или видеопотока и при этом потенциально позволяют достичь высокой точности распознавания эмоций. На основе рассмотренных архитектур предложен метод распознавания эмоций человека по двигательной активности тела в видеопотоке. Обсуждаются архитектурные особенности используемых моделей, способы обработки моделями кадров видеопотока, а также результаты распознавания эмоций по следующим метрикам качества: доля верно распознанных экземпляров (accuracy), точность (precision), полнота (recall). Результаты апробации предложенных в работе нейросетевых моделей I3D, R3D, ResNet+LSTM, ResNet+GRU на наборе данных FABO показали высокое качество распознавания эмоций по двигательной активности тела человека. Так, модель R3D показала лучшую долю верно распознанных экземпляров, равную 91 %. Другие предложенные модели: I3D, ResNet+LSTM, ResNet+GRU – показали точность распознавания 88 %, 80 % и 80 % соответственно. Таким образом, согласно полученным результатам экспериментальной оценки предложенных нейросетевых моделей, наиболее предпочтительными для использования при решении задачи распознавания эмоционального состояния человека по двигательной активности, с точки зрения совокупности показателей точности классификации эмоций, являются трехмерные сверточные модели I3D и R3D. При этом, предложенные модели, в отличие от большинства существующих решений, позволяют реализовывать распознавание эмоций на основе анализа RGB кадров видеопотока без выполнения их предварительной ресурсозатратной обработки, а также с высокой точностью выполнять распознавание эмоций в реальном масштабе времени.

Ключевые слова: нейросетевая модель, распознавание эмоций, сверточные нейронные сети, машинное обучение, обработка изображений, видеопоток

***Для цитирования*:** Уздяев М.Ю., Дударенко Д.М., Миронов В.Н. Метод распознавания эмоций человека по двигательной активности тела в видеопотоке на основе нейронных сетей. *Моделирование, оптимизация и информационные технологии.* 2021;9(1). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=929 DOI: 10.26102/2310-6018/2021.32.1.004

Method of human emotion recognition through analysis of body motor activity in a video stream using neural networks

M.Y. Uzdiaev, D.M. Dudarenko, V.N. Mironov

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russian Federation

Abstract: This paper presents the use of various neural network models to solve the problem of human emotion recognition by the motor activity of his body on frames of a video stream without complex preprocessing of these frames. The paper presents three-dimensional convolutional neural networks: Inception 3D (I3D), Residual 3D (R3D), as well as convolutional-recurrent neural network architectures using the convolutional neural network of the ResNet architecture and recurrent neural networks of the LSTM and GRU architectures (ResNet + LSTM, ResNet + GRU) which do not require preliminary processing of images or video stream and at the same time potentially allow achieving high accuracy of emotion recognition. Based on the considered architectures, a method for human emotion recognition from the motor activity of the body in a video stream is proposed. Architectural features of the used models, methods of processing video stream frames by models, as well as the results of emotion recognition according to the following quality metrics: the proportion of correctly recognized instances (accuracy), precision, recall are discussed. Approbation results of the proposed neural network models I3D, R3D, ResNet + LSTM, ResNet + GRU on the FABO data set showed a high quality of emotion recognition based on the motor activity of the human body. Thus, the R3D model showed the best share of correctly recognized copies, equal to 91%. Other proposed models: I3D, ResNet + LSTM, ResNet + GRU showed 88%, 80% and 80% recognition accuracy, respectively. Therefore, according to the obtained results of the experimental evaluation of the proposed neural network models, the most preferable for use in solving the problem of a person's emotional state recognition by motor activity, from the point of view of a set of indicators of the accuracy of emotion classification, are three-dimensional convolutional models I3D and R3D. At the same time, the proposed models, in contrast to most existing solutions, make it possible to implement emotion recognition based on the analysis of RGB frames of a video stream without performing their preliminary resource-consuming processing, as well as to perform emotion recognition in real-time with high accuracy.

Ключевые слова: neural network model, emotion recognition, convolutional neural networks, machine learning, image processing, video stream

For citation: Uzdiaev M.Y., Dudarenko D.M., Mironov V.N. Method of human emotion recognition through analysis of body motor activity in video stream using neural networks*Modeling, optimization and information technology*. 2021;9(1). Available from: https://moitvivt.ru/ru/journal/pdf?id=929 DOI: 10.26102/2310-6018/2021.32.1.004 (In Russ).

Введение

Эмоции являются одним из важнейших предикторов поведения человека. В современных моделях человеко-машинных интерфейсов широкого класса киберфизических систем [1-3], для распознавания эмоций, как правило, применяются две структурные компоненты эмоций – физиологическая и поведенческая [4]. Несмотря на то, что физиологическая компонента является наиболее валидной для определения эмоций, ее применение в реальных системах автоматического распознавания эмоций весьма затруднено необходимостью наличия дорогостоящего оборудования, высокими временными и трудовыми затратами на изменение, а также необходимостью непосредственного участия испытуемых в измерениях. С другой стороны, анализ поведенческой составляющей является более перспективным ввиду того, что поведенческие измерения не имеют таких высоких требований, их можно выполнять опосредованно на расстоянии, и они обеспечивают приемлемую валидность. Поведенческая компонента эмоций может быть представлена в различных модальностях: вербальное речевое поведение, невербальное речевое поведение, поза и двигательная активность тела субъекта, мимические выражения и т.д. Подробнее стоит остановиться на позе и двигательной активности тела человека. Использование данной

модальности для распознавания эмоций становится актуальным в тех ситуациях, когда наблюдается отсутствие или недостаток данных других модальностей. В данной работе предложен метод решения задачи распознавания эмоционального состояния по двигательной активности, основанный на использовании сверточно-рекурретных и трехмерных нейросетевых архитектур, которые не требуют предварительной обработки изображений или видеопотока и при этом потенциально позволяют достичь высокой точности распознавания эмоций.

Анализ известных подходов

В настоящее время существует множество подходов к решению задачи определения эмоционального состояния на видео или отдельных изображениях [5, 6]. С целью обеспечения большей универсальности, а также повышения точности существующих решений в области распознавания эмоций, различными исследователями были предложены усовершенствованные методы решения данной задачи, отличительной чертой которых является большее число оцениваемых факторов [7-9]. В связи с тем, что методы, ориентированные на оценку мимики лица или речи, являются крайне неустойчивыми к фоновым шумам и иным помехам, были разработаны методы, которые помимо вышеуказанных параметров оценивают также и позу человека [10-12]. Например, авторами [11] была разработана система, состоящая из нейронной сети, позволяющей извлекать необходимые визуальные признаки, и нейронной сети, которая реализует сопоставление данных признаков. Сеть для извлечения признаков включает в себя три подсети: первая отвечает за извлечение признаков лица, вторая за извлечение признаков тела и третья за извлечение признаков всего изображения. Сеть сопоставления объектов анализирует полученные данные из трех подсетей и предсказывает 26 дискретных категорий и 3 непрерывных измерения (лицо, тело, изображение). По результатам тестирования данная система продемонстрировала показатель точности, равный 73%. Следует отметить, что данная система имеет недостаток, связанный с излишним количеством обрабатываемой визуальной информации, а именно, отдельная обработка разными нейросетями физической активности (позы человека), лица и всего изображения делает соответствующие решения крайне ресурсоемкими. На сегодняшний день такого рода методы получили большое распространение и считаются более устойчивыми к качеству входных данных, поскольку способны функционировать в условиях с большим количеством людей в кадре, а также не зависят от наличия фоновых шумов в исследуемой среде. Ключевым недостатком таких подходов является недостаточно высокая точность распознавания эмоций, не превосходящая, как правило, 75% [13, 14], и обусловленная, в том числе недостаточно высокими показателями качества работы моделей нейронных сетей, лежащих в основе соответствующих решений. Использование на этапе извлечения признаков и этапе классификации современных нейросетевых моделей [15-17], демонстрирующих лучшую точность в задачах классификации действий человека на видеоданных, потенциально позволит сформировать решение, обладающее более высокими показателями точности определения эмоций по видеопотоку. Также следует отметить, что в современной научной литературе не достаточно внимания уделено моделям, методам и системам распознавания эмоций на основе анализа его двигательной активности в видеопотоке без сложной предварительной обработки кадров, такой как выделение ключевых точек, оптического потока и т.д. В рамках данной работы был разработан метод определения эмоционального состояния человека путем анализа двигательной активности тела человека в видеопотоке с использованием комбинированных нейросетевых моделей ResNet, I3D и R3D. Данный метод распознавания эмоционального состояния человека по

двигательной активности потенциально позволит с высокой точностью выполнять распознавание эмоций в реальном масштабе времени.

Описание разработанного метода

В соответствии с результатами проведенного анализа связанных методов и подходов, для определения эмоции по двигательной активности тела человека в рамках настоящего исследования предложен авторский метод решения данной задачи. В данной работе рассматриваются два основных подхода, которые обычно применяются в задачах обработки и анализа видеопоследовательностей. Первый подход к анализу последовательностей кадров состоит из двух этапов. На первом этапе происходит выделение с помощью различных методов пространственной информации о визуальных объектах на кадрах видео, а на втором этапе выполняется временной анализ последовательности представлений выделенной пространственной информации. Второй подход заключается в одновременном анализе пространственно-временной информации без разбиения на отдельные этапы обработки.

Рассмотрим сначала подробнее первый подход к анализу видеопоследовательностей и нейросетевые модели, соответствующие данному подходу. Для решения задачи по выделению пространственных признаков визуальных объектов целесообразно использовать глубокие сверточные нейронные сети (Convolutional Neural Network – CNN), предварительно обученные задаче классификации большого количества визуальных объектов на наборе данных Imagenet [18]. Такой прием является распространенным в широком спектре различных задач компьютерного зрения, таких как генерация текстового описания изображений [19, 20], видео [21, 22], распознавание действий [23, 24], локализация визуальных объектов [25-28] и соответствует парадигме переноса обучения (Transfer Learning) [29, 30]. В рамках первого подхода к анализу видеопоследовательностей при решении задачи определения эмоционального состояния человека путем анализа двигательной активности тела субъекта была сформирована обобщенная архитектура соответствующих нейросетевых моделей, представленная на Рисунке 1.

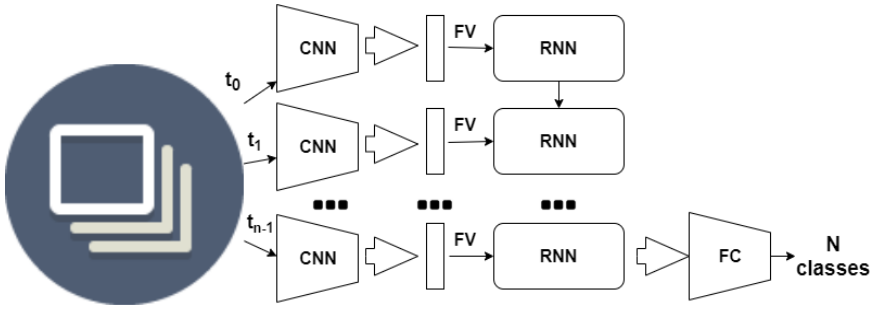


Рисунок 1 – Обобщенная архитектура сверточно-рекуррентной нейронной сети
Figure1 – High-level architecture of a convolutional-recurrent neural network

Согласно представленной выше архитектуре, каждый кадр исследуемого видеоряда t_i поступает на вход экстрактору признаков CNN. Выделяемая с помощью CNN последовательность признаковых представлений или векторов признаков FV_i содержит в себе наиболее значимую информацию о пространственных составляющих визуальных объектов. Данная последовательность далее обрабатывается при помощи рекуррентной нейронной сети RNN (Recurrent Neural Network), где по достижении последнего элемента последовательности выполняется классификация всей

последовательности при помощи полносвязного слоя нейронной сети FC (Fully Connected).

В рамках настоящего исследования было сформировано несколько реализаций описанной архитектуры, где в качестве экстрактора использовалась предварительно обученная CNN архитектуры ResNet, основной особенностью которой является использование т.н. Residual блоков, обеспечивающих учет входной информации на выходе блока. Это препятствует размытию градиентов на этапе обучения с увеличением количества слоев нейросети, что в итоге приводит к более детальной обработке информации и повышению качества классификации [31]. В данной работе использовалась нейронная сеть ResNet, состоящая из 18 слоев.

В качестве рекуррентных нейронных сетей, выполняющих временной анализ динамики векторов признаков, были выбраны однонаправленные RNN архитектуры долгой краткосрочной памяти (Long Term Short Memory – LSTM) [32] и вентильных управляемых элементов (Gated Recurrent Units – GRU) [33]. Основными особенностями архитектуры LSTM являются, во-первых, наличие т.н. памяти или внутреннего состояния ячейки, что позволяет сохранять контекстную информацию при обработке последовательностей большой длины, а также применение т.н. вентильных элементов (gate units), которые обеспечивают контроль записи, сохранения и стирания информации в ячейке с учетом контекста обрабатываемой временной последовательности. Данные особенности позволяют эффективно обрабатывать длительные временные последовательности. Архитектура GRU является упрощенным по сравнению с LSTM вариантом архитектуры ячейки рекуррентной нейронной сети. Простота реализации достигается за счет меньшего количества вентильных элементов и отсутствия ячейки состояния, что обеспечивает меньшее количество матрично-векторных операций. Несмотря на упрощенное устройство, данная архитектура также способна эффективно обрабатывать последовательные данные.

Таким образом, за счет комбинирования LSTM и GRU архитектур с моделью ResNet, в рамках исследования были получены две специфические реализации представленной выше обобщенной архитектуры сверточно-рекуррентной нейронной сети: ResNet+LSTM и ResNet+GRU. В отличие от подхода, в котором для выделения пространственных признаков визуальных объектов используются сверточные нейронные сети, а для анализа временной динамики этих признаков – рекуррентные, второй подход предполагает одновременный анализ пространственно-временной информации, содержащейся на кадрах видеопотока. Представителями такого подхода являются трехмерные сверточные нейронные сети (3D CNN). Такие нейросетевые архитектуры обеспечивают высокую точность как в общей задаче распознавания большого количества действий человека [16, 17, 34, 35], так и в более специфичных задачах, например, распознавание агрессивных действий человека [36]. На Рисунке 2 ниже представлена обобщенная архитектура 3D CNN, реализующая данный подход.

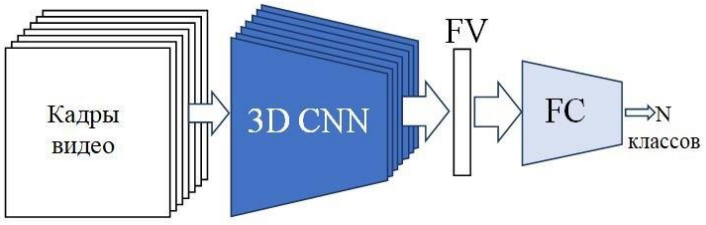


Рисунок 2 – Обобщенная архитектура 3D CNN
Figure 2 – High-level architecture of 3D CNN

Представленная выше архитектура принимает на вход последовательность кадров заданной длины. Данная последовательность обрабатывается 3D CNN, которая выполняет пространственно-временной анализ динамики визуальных объектов на кадрах видеоряда. На своем выходе 3D CNN формирует вектор признаков FV, содержащий в себе наиболее значимую информацию о пространственно-временной динамике визуальных объектов. Далее FV обрабатывается многослойной полносвязной нейронной сетью, которая выполняет итоговую классификацию. В рамках данного подхода также возможно применение методов переноса обучения, которое заключается в использовании экстракторов 3D CNN, предварительно обученных задаче классификации большого количества действий на больших репрезентативных наборах данных, для выделения FV и последующем дообучении классификатора FC специфической задаче классификации эмоций по двигательной активности тела на этих выделенных данных.

В данной работе в качестве 3D CNN экстракторов были использованы следующие архитектуры 3D CNN – Inception 3D (I3D) [16] и Residual 3D (R3D) [17, 35]. Обе архитектуры были предварительно обучены общей задаче классификации 400 действий человека на наборе данных Kinetics [37]. Основным структурным блоком в архитектуре I3D является т.н. 3D Inception блок [38], который выполняет параллельную обработку карты признаков, полученной на предыдущем слое, с помощью четырех ветвей, содержащих в себе операции свертки различного размера. Выход Inception блока формируется при помощи конкатенации карт признаков, генерируемых этими четырьмя ветвями. Сама нейронная сеть I3D состоит из пяти входных трехмерных сверточных и пулинговых слоев, затем следуют три каскада Inception блоков, содержащих в себе 2, 5 и 2 Inception блока соответственно. Эти каскады разделяются слоями трехмерного пулинга. Выход нейросетевой модели I3D является классификатором и представляет собой сверточный слой с ядрами размером (1x1x1) и активационной функцией softmax. В качестве обучаемой части в работе был взят последний каскад Inception блоков, содержащий в себе 2 блока и выходной сверточный классификатор. Веса остальных блоков не изменялись в процессе обучения.

Архитектура R3D [17, 35] является аналогом 2D архитектуры ResNet [31]. В R3D также используются блоки Residual, с тем лишь отличием, что вместо операций двумерной свертки и двумерного пулинга, используются их трехмерные аналоги. В данной работе была использована модель R3D, состоящая из 18 слоев. Первый слой данной нейронной сети представляет собой трехмерный сверточный слой, выполняющий выделение низкоуровневых пространственно-временных признаков. Далее следует четыре каскада Residual блоков, с двумя Residual блоками в каждом каскаде. Выход R3D представлен однослойной полносвязной нейросетью с активационной функцией softmax. В рамках настоящего исследования в качестве обучаемой части был взят последний каскад Residual блоков, содержащий в себе 2 блока и выходной полносвязный классификатор. Веса всех остальных блоков не изменялись в процессе обучения. Далее перейдем к оценке результатов обучения, представленных выше моделей (ResNet+LSTM, ResNet+GRU, I3D, R3D) с точки зрения их применимости к решению задачи распознавания эмоционального состояния человека по двигательной активности на видеопоследовательностях.

Результаты обучения и их анализ

Для сравнения эффективности рассматриваемых нейросетевых моделей в качестве экспериментального набора данных был выбран FABO (Face and Body Gesture Database)

[39], который содержит в себе большое количество размеченных экземпляров проявлений эмоций различными людьми, выраженных в мимике и двигательной активности тела индивидов. FABO состоит из 550 видеозаписей, содержащих 11 различных классов проявлений эмоций. При этом, данная база размечалась на основе экспертных оценок, а также субъективного отчета испытуемых. Данный набор содержит в себе эмоциональные проявления 23 людей (11 мужчин и 12 женщин) различной этнической принадлежности четырех возрастных групп – 18-24 года (9 человек), 25-29 года (7 человек), 30-40 года (4 человека), 40-50 (1 человек). При этом, на каждом видео присутствует несколько движений тела и мимических выражений индивида, отражающих проявление той или иной эмоции. На Рисунке 3 изображены примеры из набора данных FABO.

На этапе обучения рассматриваемых нейросетевых моделей, выполнялась аугментация видеопотока с целью расширения обучающей выборки и регуляризации данных. При этом, для обработки экземпляров исходного набора данных, использовалась как пространственная, так и временная аугментация. В качестве временной аугментации применялся выбор случайного фрагмента последовательных кадров длительностью 32 кадра. В качестве пространственной аугментации использовались следующие методы: вырезание случайного фрагмента кадра необходимого размера, равного 224x224 пикселя, масштабирование изображения со случайным коэффициентом, взятым из промежутка от -1.5 до 1.5; поворот на случайный угол из промежутка от -10 до 10 градусов; случайное зеркальное отображение кадра. На этапе валидации модели аугментация данных не выполнялась.

Рисунок 3 – Примеры кадров из набора данных FABO
Figure 3 – Sample frames from the FABO dataset

Для проведения экспериментальной оценки моделей изначальный набор данных был дополнительно расширен путем фрагментирования исходных видео посредством разбиения исходных видеозаписей на фрагменты длиной 64 кадра. Итоговый набор данных таким образом включал в себя 1596 видеозаписей. Для всех моделей данный

набор видеопоследовательностей был разбит на две части - обучающую выборку и валидационную выборку размером 1276 и 320 видеофрагментов соответственно. Обучение выполнялось методом стохастического пакетного градиентного спуска. При этом, для трехмерных сверточных нейронных сетей были выбраны размеры пакетов, равные 16 видеороликов по 32 кадра каждый, а для сверточно-рекуррентных архитектур – 32 видеоролика по 32 кадра каждый. В качестве алгоритма оптимизации параметров нейросетей был выбран алгоритм Adam [40]. В качестве функции потерь была выбрана многоклассовая логарифмическая перекрестная энтропия:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i,$$

где N – количество классов, y_i эталонное значение класса, $\hat{y}_i$ – актуальное значение класса, сгенерированное нейронной сетью. В данной работе для оценки качества работы моделей были использованы следующие метрики: доля верно распознанных экземпляров *acc*, точность *pr*, полнота *rec*:

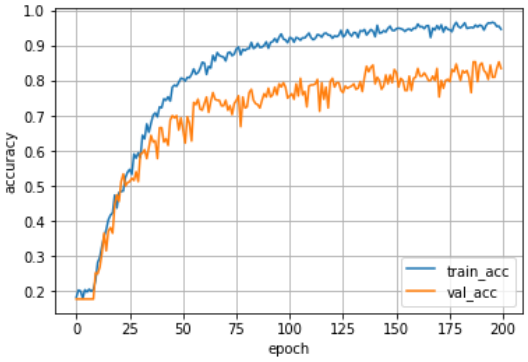
$$acc = \frac{t_p + t_n}{t_p + t_n + f_p + f_n},$$

$$pr = \frac{t_p}{\left(t_p + f_p\right)},$$

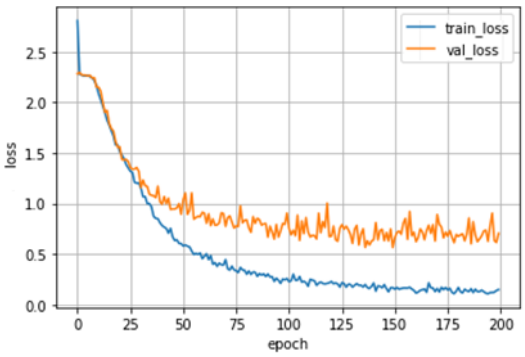
$$rec = \frac{t_p}{\left(t_p + f_n\right)}.$$

где t_p – количество верно классифицированных положительных экземпляров, t_n – количество верно классифицированных отрицательных экземпляров, f_p – количество неверно классифицированных положительных экземпляров, f_n – количество неверно классифицированных отрицательных экземпляров.

На Рисунке 4 представлены графики зависимости точности (accuracy) распознавания в процессе обучения и графики зависимости величины ошибки классификации от количества пройденных эпох: *а* – точность модели I3D; *б* – величина ошибки модели I3D; *в* – точность модели R3D; *г* – величина ошибки модели R3D; *д* – точности модели ResNet+GRU; *е* – величина ошибки модели ResNet+GRU; *ж* – точность модели ResNet+LSTM; *з* – величина ошибки для модели ResNet+LSTM в зависимости от числа пройденных эпох обучения.



а)



б)

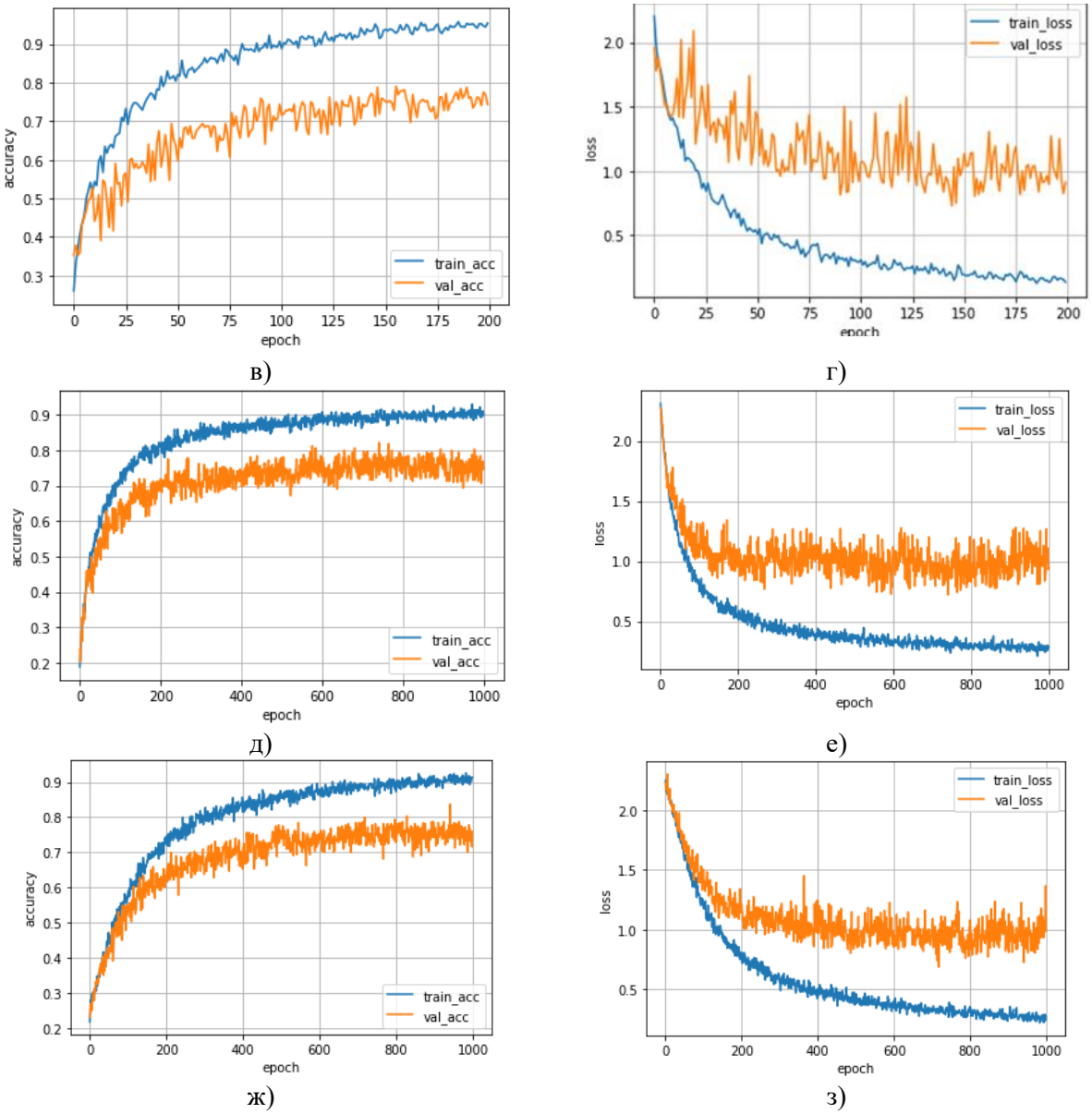


Рисунок 4 – Результаты экспериментов
Figure 4 – Experimental results

Стоит отметить, что значения функций ошибок на этапе валидации не растут по мере увеличения числа пройденных эпох, что свидетельствует о том, что переобучения у моделей не наступает. По результатам тестирования можно заключить, что итоговыми значениями точности для каждой модели стали: I3D – 88%, R3D – 91%, ResNet+GRU – 80%, ResNet+LSTM – 80%.

В Таблице 1 приведены результаты сравнительной оценки точности рассматриваемых в настоящем исследовании моделей, а также других моделей, при обучении которых использовался набор данных FABO [39], в контексте решения задачи распознавания эмоционального состояния человека по двигательной активности.

Таблица 1 – Точность различных моделей машинного обучения в задаче распознавания эмоций
Table 1 – Accuracy of different machine learning models in emotion recognition problem

Модель	Точность
SVM [41]	0.65
Random Forest [41]	0.76
Temporal Normalization [42]	0.67
Bag of words [42]	0.65
MHI [43]	0.53
MCCNN [43]	0.58
keyframes HMI + CNN + ConvLSTM [13]	0.73
I3D	0.88
R3D	**0.91**
ResNet+GRU	0.80

Как можно заметить, рассмотренные в данной работе модели демонстрируют значительно более высокие показатели точности в сравнении с другими решениями. При этом наибольшей точности достигает решение на основе трехмерной сверточной нейронной сети R3D: данный показатель для модели R3D составил 0.91.

В Таблице 2 представлены значения мер точности (precision, *pr*) и полноты (recall, *rec*) для каждой из предложенных в работе моделей. Полученные результаты позволяют оценить, как разработанные модели способны распознавать истинно положительные экземпляры классов среди положительно распознанных экземпляров, а также общую долю истинно положительных экземпляров соответственно.

Таблица 2– Поклассовые значения мер точности и полноты для моделей I3D, R3D, ResNet+GRU, ResNet +LSTM
Table 2 – Per-class values of precision and recall metrics for the models I3D, R3D, ResNet+GRU, ResNet +LSTM

Метка класса	I3D		R3D		ResNet +GRU		ResNet +LSTM	
	pr	*rec*	*pr*	*rec*	*pr*	*rec*	*pr*	*rec*
Счастье	1.00	1.00	0.72	0.95	0.82	0.85	0.88	0.81
Грусть	0.73	0.92	0.79	0.92	0.81	0.93	0.88	1.00
Скука	0.87	0.87	0.93	0.91	0.84	0.93	0.79	0.93
Страх	0.89	0.89	0.94	0.89	0.82	0.88	0.74	0.88
Негативное удивление	0.80	0.84	1.00	0.84	0.82	0.58	0.94	0.67
Позитивное удивление	0.67	0.57	0.80	0.57	1.00	0.70	0.70	0.70
Отвращение	0.95	0.82	0.95	0.95	0.88	0.67	0.94	0.71
Гнев	0.93	0.94	0.95	0.87	0.83	0.85	0.74	0.85
Неуверенность	0.88	0.75	0.76	0.95	0.78	0.69	0.81	0.50
Смущение	0.81	0.93	0.98	0.93	0.74	0.93	0.77	0.93
Тревога	0.96	0.79	0.89	0.97	0.62	0.43	0.92	0.52

Исходя из полученных значений метрик качества precision и recall можно сделать следующие выводы: модель I3D демонстрирует высокие результаты при выделении

истинно положительно распознанных экземпляров всех классов, за исключением класса позитивного удивления (precision=0.67, recall=0.57). Модель R3D также демонстрирует высокие показатели precision и recall, демонстрируя, однако низкую способность распознавания истинно положительных экземпляров класса «позитивное удивление» (recall=0.57). У сверточно-рекуррентных моделей ResNet+GRU и ResNet+LSTM также наблюдаются высокие показатели precision и recall. Однако, модель ResNet+GRU испытывает затруднения при распознавании истино положительных экземпляров классов «отвращение» и «тревога» (значения recall 0.58 и 0.43 соответственно). Кроме того, эта модель также имеет затруднения при распознавании релевантных экземпляров класса «тревога» (precision=0.62). Модель ResNet+LSTM имеет затруднения при распознавании истино положительных экземпляров классов «неуверенность» и «тревога» (значения recall 0.50 и 0.42 соответственно).

Таким образом, согласно полученным результатам экспериментальной оценки предложенных нейросетевых моделей, наиболее предпочтительными для использования при решении задачи распознавания эмоционального состояния человека по двигательной активности с точки зрения совокупности показателей точности классификации эмоций являются трехмерные сверточные модели I3D и R3D.

Заключение

Результаты апробации предложенных в работе нейросетевых моделей на наборе данных FABO показали высокое качество распознавания эмоций по двигательной активности тела человека. Так, модель R3D показала лучшую долю верно распознанных экземпляров, равную 91%. Другие предложенные модели: I3D, ResNet+LSTM, ResNet+GRU показали 88%, 80% и 80% точность распознавания соответственно. Показатели качества работы предложенных моделей существенно превосходят результаты других моделей, методов и систем, апробированных на наборе данных FABO. При этом, предложенные модели, в отличие от большинства существующих решений позволяют реализовывать распознавание эмоций на основе анализа RGB кадров видеопотока без выполнения их предварительной ресурсозатратной обработки. Стоит отметить, что для обучения моделей ResNet+LSTM и ResNet+GRU требуются большие временные затраты по сравнению с трехмерными сверточными нейронными сетями R3D и I3D. Таким образом, наиболее предпочтительными для использования при решении задачи распознавания эмоционального состояния человека по двигательной активности являются трехмерные сверточные модели I3D и R3D. Дальнейшее развитие настоящего исследования может заключаться в усовершенствовании предложенных моделей за счет добавления различных архитекурных элементов, таких как нейросетевой механизм внимания [44], а также посредством расширения пространства признаков обрабатываемых данных и обучения предложенных моделей на других наборах данных.

Благодарности

Работа выполнена при поддержке РФФИ (18-29-22061_мк).

ЛИТЕРАТУРА

1. Ватаманюк И.В., Яковлев Р.Н. Алгоритмическая модель распределенной системы корпоративного информирования в рамках киберфизической системы организации. Моделирование, оптимизация и информационные технологии. 2019;7(4). Доступно по: https://moit.vivt.ru/wp-content/uploads/2019/11/VatamanukSoavtori_4_19_1.pdf. DOI: 10.26102/2310-6018/2019.27.4.026 (дата обращения: 20.10.2020).
2. Letenkov M., Levonevskiy D. Fast Face Features Extraction Based on Deep Neural Networks for Mobile Robotic Platforms. *International Conference on Interactive*

Collaborative Robotics. Springer, Cham. 2020:200-211. DOI: 10.1007/978-3-030-60337-3_20

3. Ватаманюк И.В., Яковлев Р.Н. Обобщенные теоретические модели киберфизических систем. Известия Юго-Западного государственного университета. 2019;23(6):161-175. Доступно по: https://science.swsu.ru/jour/article/view/666/489. DOI: 10.21869/2223-1560-2019-23-6-161-175 (дата обращения: 20.10.2020).
4. Frijda N.H. Emotions and action. *Feelings and emotions: The Amsterdam symposium.* 2004:158-173.
5. He G., Liu X., Fan F., You J. Image2Audio: Facilitating Semi-supervised Audio Emotion Recognition with Facial Expression Image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020:912-913.
6. Kalsum T., Anwar S.M., Majid M., Khan B., Ali S.M. Emotion recognition from facial expressions using hybrid feature descriptors. *IET Image Processing*. 2018;12(6):1004-1012.
7. Levonevskii D., Shumskaya O., Velichko A., Uzdiaev M., Malov D. Methods for Determination of Psychophysiological Condition of User Within Smart Environment Based on Complex Analysis of Heterogeneous Data. *Proceedings of 14th International Conference on Electromechanics and Robotics «Zavalishin's Readings». Springer, Singapore*. 2020:511-523.
8. Уздяев М.Ю., Левоневский Д.К., Шумская О.О., Летенков М.А. Методы детектирования агрессивных пользователей информационного пространства на основе генеративно-состязательных нейронных сетей. *Информационно-измерительные и управляющие системы.* 2019;17(5):60-68.
9. Uzdiaev M. Methods of Multimodal Data Fusion and Forming Latent Representation in the Human Aggression Recognition Task. *2020 IEEE 10th International Conference on Intelligent Systems* (*IS*). IEEE. 2020:399-403.
10. Thakur N., Han C.Y. A complex activity based emotion recognition algorithm for affect aware systems. *2018 IEEE 8th Annual Computing and Communication Workshop and Conference* (*CCWC*). IEEE. 2018:748-753.
11. Wu J., Zhang Y., Ning L. The Fusion Knowledge of Face, Body and Context for Emotion Recognition. *2019 IEEE International Conference on Multimedia & Expo Workshops* (*ICMEW*). IEEE. 2019:108-113.
12. Piana S., Staglianò A., Odone F., Camurri A. Adaptive body gesture representation for automatic emotion recognition. *ACM Transactions on Interactive Intelligent Systems* (*TiiS*). 2016;6(1):1-31.
13. Ly S.T., Lee G.S., Kim S.H., Yang H.J. Emotion Recognition via Body Gesture: Deep Learning Model Coupled with Keyframe Selection. *Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence*. 2018:27-31.
14. Shen Z., Cheng J., Hu X., Dong Q. Emotion Recognition Based on Multi-View Body Gestures. *2019 IEEE International Conference on Image Processing* (*ICIP*). IEEE, 2019:3317-3321.
15. Targ S., Almeida D., Lyman K. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:*1603.08029. 2016.
16. Carreira J., Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017:6299-6308.
17. Hara K., Kataoka H., Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. *Proceedings of the IEEE International Conference on Computer Vision Workshops.* 2017:3154-3160.

18. Deng J., Dong W., Socher R., Li L. J., Li K., Fei-Fei L. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition.* IEEE. 2009:248-255.
19. Vinyals O., Toshev A., Bengio S., Erhan D. Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015:3156-3164.
20. Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhudinov R., Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning.* 2015:2048-2057.
21. Yao L., Torabi A., Cho K., Ballas N., Pal C., Larochelle H., Courville A. Describing videos by exploiting temporal structure. *Proceedings of the IEEE international conference on computer vision.* 2015:4507-4515.
22. Hori C., Hori T., Lee T. Y., Zhang Z., Harsham B., Hershey J. R., Sumi K. Attention-based multimodal fusion for video description. *Proceedings of the IEEE international conference on computer vision.* 2017:4193-4202.
23. Yue-Hei Ng, J., Hausknecht M., Vijayanarasimhan S., Vinyals O., Monga R., Toderici G. Beyond short snippets: Deep networks for video classification. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015:4694-4702.
24. Ullah A., Ahmad J., Muhammad K., Sajjad M., Baik S. W. Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access*. 2017;6:1155-1166.
25. Girshick R. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision.* 2015:1440-1448.
26. Ren S., He K., Girshick R., Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*. 2015:91-99.
27. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:779-788.
28. Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.Y., Berg A.C. Ssd: Single shot multibox detector. *European conference on computer vision*. Springer, Cham, 2016:21-37.
29. Pan S.J., Yang Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*. 2009;22(10):1345-1359.
30. Weiss K., Khoshgoftaar T.M., Wang D.D. A survey of transfer learning. *Journal of Big data.* 2016;3(1):9.
31. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:770-778.
32. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural computation.* 1997;9(8):1735-1780.
33. Chung J., Gulcehre C., Cho K., Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:*1412.3555. 2014.
34. Tran D., Bourdev L., Fergus R., Torresani L., Paluri M. Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision.* 2015:4489-4497.
35. Hara K., Kataoka H., Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 2018:6546-6555.

36. Saveliev A., Uzdiaev M., Dmitrii M. Aggressive Action Recognition Using 3D CNN Architectures. *2019 12th International Conference on Developments in eSystems Engineering* (*DeSE*). IEEE. 2019:890-895.
37. Kay W., Carreira J., Simonyan K., Zhang B., Hillier C., Vijayanarasimhan S., Suleyman M. The kinetics human action video dataset. *arXiv preprint arXiv:*1705.06950. 2017.
38. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Rabinovich A. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015:1-9.
39. Gunes H., Piccardi M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. *18th International Conference on Pattern Recognition* (*ICPR'06*). IEEE. 2006;1:1148-1153.
40. Kingma D. P., Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv*:1412.6980. 2014.
41. Gunes H., Piccardi M. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* (*Cybernetics*). 2008;39(1):64-84.
42. Chen S., Tian Y., Liu Q., Metaxas D.N. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing.* 2013;31(2):175-185.
43. Barros P., Jirak D., Weber C., Wermter S. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks.* 2015;72:140-151.
44. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:*1409.0473. 2014.

REFERENCES

1. Vatamaniuk I.V., Yakovlev R.N. Algorithmic model of a distributed corporate notification system in context of a corporate cyber-physical system. Modeling, optimization and information technology. 2019;7(4). Available at: https://moit.vivt.ru/wp-content/uploads/2019/11/VatamanukSoavtori_4_19_1.pdf. (In Russ) DOI: 10.26102/2310-6018/2019.27.4.026 (accessed 20.10.2020).
2. Letenkov M., Levonevskiy D. Fast Face Features Extraction Based on Deep Neural Networks for Mobile Robotic Platforms. *International Conference on Interactive Collaborative Robotics. Springer, Cham.* 2020:200-211. DOI: 10.1007/978-3-030-60337-3_20
3. Vatamaniuk I. V., Iakovlev R. N. Generalized Theoretical Models of Cyberphysical Systems. Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta = Proceedings of the Southwest State University. 2019;23(6):161-175 (In Russ.). Available at: https://science.swsu.ru/jour/article/view/666/489. DOI: 10.21869/2223-1560-2019-23-6-161-175 (accessed 20.10.2020).
4. Frijda N.H. Emotions and action. *Feelings and emotions: The Amsterdam symposium.* 2004:158-173.
5. He G., Liu X., Fan F., You J. Image2Audio: Facilitating Semi-supervised Audio Emotion Recognition with Facial Expression Image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020:912-913.
6. Kalsum T., Anwar S.M., Majid M., Khan B., Ali S.M. Emotion recognition from facial expressions using hybrid feature descriptors. *IET Image Processing*. 2018;12(6):1004-1012.
7. Levonevskii D., Shumskaya O., Velichko A., Uzdiaev M., Malov D. Methods for Determination of Psychophysiological Condition of User Within Smart Environment

Based on Complex Analysis of Heterogeneous Data. *Proceedings of 14th International Conference on Electromechanics and Robotics «Zavalishin's Readings». Springer, Singapore*. 2020:511-523.

8. Uzdiaev M., Levonevskii D., Shumskaya O., Letenkov M. Methods for detecting aggressive users of the information space based on generative-competitive neural networks. *"Informatsionno-izmeritelnye i upravlyayushchie sistemy" (Information-measuring and Control Systems).* 2019;17(5):60-68. (In Russ)
9. Uzdiaev M. Methods of Multimodal Data Fusion and Forming Latent Representation in the Human Aggression Recognition Task. *2020 IEEE 10th International Conference on Intelligent Systems* (*IS*). IEEE. 2020:399-403.
10. Thakur N., Han C.Y. A complex activity based emotion recognition algorithm for affect aware systems. *2018 IEEE 8th Annual Computing and Communication Workshop and Conference* (*CCWC*). IEEE. 2018:748-753.
11. Wu J., Zhang Y., Ning L. The Fusion Knowledge of Face, Body and Context for Emotion Recognition. *2019 IEEE International Conference on Multimedia & Expo Workshops* (*ICMEW*). IEEE. 2019:108-113.
12. Piana S., Staglianò A., Odone F., Camurri A. Adaptive body gesture representation for automatic emotion recognition. *ACM Transactions on Interactive Intelligent Systems* (*TiiS*). 2016;6(1):1-31.
13. Ly S.T., Lee G.S., Kim S.H., Yang H.J. Emotion Recognition via Body Gesture: Deep Learning Model Coupled with Keyframe Selection. *Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence*. 2018:27-31.
14. Shen Z., Cheng J., Hu X., Dong Q. Emotion Recognition Based on Multi-View Body Gestures. *2019 IEEE International Conference on Image Processing* (*ICIP*). IEEE, 2019:3317-3321.
15. Targ S., Almeida D., Lyman K. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:*1603.08029. 2016.
16. Carreira J., Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017:6299-6308.
17. Hara K., Kataoka H., Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. *Proceedings of the IEEE International Conference on Computer Vision Workshops.* 2017:3154-3160.
18. Deng J., Dong W., Socher R., Li L. J., Li K., Fei-Fei L. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition.* IEEE. 2009:248-255.
19. Vinyals O., Toshev A., Bengio S., Erhan D. Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015:3156-3164.
20. Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhudinov R., Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning.* 2015:2048-2057.
21. Yao L., Torabi A., Cho K., Ballas N., Pal C., Larochelle H., Courville A. Describing videos by exploiting temporal structure. *Proceedings of the IEEE international conference on computer vision.* 2015:4507-4515.
22. Hori C., Hori T., Lee T. Y., Zhang Z., Harsham B., Hershey J. R., Sumi K. Attention-based multimodal fusion for video description. *Proceedings of the IEEE international conference on computer vision.* 2017:4193-4202.

23. Yue-Hei Ng, J., Hausknecht M., Vijayanarasimhan S., Vinyals O., Monga R., Toderici G. Beyond short snippets: Deep networks for video classification. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015:4694-4702.
24. Ullah A., Ahmad J., Muhammad K., Sajjad M., Baik S. W. Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access*. 2017;6:1155-1166.
25. Girshick R. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision.* 2015:1440-1448.
26. Ren S., He K., Girshick R., Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*. 2015:91-99.
27. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:779-788.
28. Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.Y., Berg A.C. Ssd: Single shot multibox detector. *European conference on computer vision*. Springer, Cham, 2016:21-37.
29. Pan S.J., Yang Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*. 2009;22(10):1345-1359.
30. Weiss K., Khoshgoftaar T.M., Wang D.D. A survey of transfer learning. *Journal of Big data.* 2016;3(1):9.
31. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016:770-778.
32. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural computation.* 1997;9(8):1735-1780.
33. Chung J., Gulcehre C., Cho K., Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:*1412.3555. 2014.
34. Tran D., Bourdev L., Fergus R., Torresani L., Paluri M. Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision.* 2015:4489-4497.
35. Hara K., Kataoka H., Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 2018:6546-6555.
36. Saveliev A., Uzdiaev M., Dmitrii M. Aggressive Action Recognition Using 3D CNN Architectures. *2019 12th International Conference on Developments in eSystems Engineering* (*DeSE*). IEEE. 2019:890-895.
37. Kay W., Carreira J., Simonyan K., Zhang B., Hillier C., Vijayanarasimhan S., Suleyman M. The kinetics human action video dataset. *arXiv preprint arXiv:*1705.06950. 2017.
38. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Rabinovich A. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015:1-9.
39. Gunes H., Piccardi M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. *18th International Conference on Pattern Recognition* (*ICPR'06*). IEEE. 2006;1:1148-1153.
40. Kingma D. P., Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv*:1412.6980. 2014.
41. Gunes H., Piccardi M. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* (*Cybernetics*). 2008;39(1):64-84.

42. Chen S., Tian Y., Liu Q., Metaxas D.N. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing.* 2013;31(2):175-185.
43. Barros P., Jirak D., Weber C., Wermter S. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks.* 2015;72:140-151.
44. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:*1409.0473. 2014.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Уздяев Михаил Юрьевич, младший научный сотрудник лаборатории технологий больших данных социокиберфизических систем, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Санкт-Петербург, Российская Федерация
e-mail: *m.y.uzdiaev@gmail.com*
ORCID: 0000-0002-7032-0291

Mikhail Yu. Uzdiaev, junior researcher of Laboratory of big data in socio-cyberphysical systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russian Federation

Дударенко Дмитрий Михайлович, младший научный сотрудник лаборатории технологий больших данных социокиберфизических систем, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Санкт-Петербург, Российская Федерация
e-mail: dmitry@dudarenko.net
ORCID: 0000-0002-9509-178X

Dmitry M. Dudarenko, junior researcher of Laboratory of big data in socio-cyberphysical systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russian Federation

Миронов Виктор Николаевич, программист лаборатории технологий больших данных социокиберфизических систем, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Санкт-Петербург, Российская Федерация
e-mail: *vmn20@mail.ru*

Viktor N. Mironov, software engineer of Laboratory of big data in socio-cyberphysical systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russian Federation