УДК 004.032.26 + 004.93 Научная статья

Нейросетевая модель многомодального распознавания человеческой агрессии

М. Ю. Уздяев

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, лаборатория автономных робототехнических систем, 14 линия д. 39, г. Санкт-Петербург, 199178, Россия

E-mail: uzdyaev.m@iias.spb.su

Увеличение количества пользователей социокиберфизических систем, умных пространств, систем интернета вещей актуализирует проблему выявления деструктивных действий пользователей, таких как агрессия. При этом, деструктивные действия пользователей могут быть представлены в различных модальностях: двигательная активность тела, сопутствующее выражение лица, невербальное речевое поведение, вербальное речевое поведение. В статье рассматривается нейросетевая модель многомодального распознавания человеческой агрессии, основанная на построении промежуточного признакового пространства, инвариантного виду обрабатываемой модальности. Предлагаемая модель позволяет распознавать с высокой точностью агрессию в условиях отсутствия или недостатка информации какой-либо модальности. Экспериментальное исследование показало 81.8% верных распознаваний на наборе данных IEMOCAP. Также приводятся результаты экспериментов распознавания агрессии на наборе данных IEMOCAP для 15 различных сочетаний обозначенных выше модальностей.

Ключевые слова: распознавание агрессии, анализ поведения, нейронные сети, многомодальная обработка данных.

DOI: 10.26117/2079-6641-2020-33-4-132-149

Поступила в редакцию: 18.11.2020 В окончательном варианте: 10.12.2020

Для цитирования. Уздяев М. Ю. Нейросетевая модель многомодального распознавания человеческой агрессии // *Вестник КРАУНЦ. Физ.-мат. науки*. 2020. Т. 33. № 4. С. 132-149. DOI: 10.26117/2079-6641-2020-33-4-132-149

Финансирование. Работа выполнена при поддержке РФФИ (проект № 18-29-22061_мк

Введение

В последние годы все большее распространение получают социокиберфизические пространства, умные среды, системы интернета вещей. Вместе с этим, растет и число пользователей таких систем. Данные системы в первую очередь служат для облегчения деятельности большого количества людей, а также для облегчения взаимодействия людей друг с другом. Однако, с ростом количества пользователей увеличивается риск возникновения конфликтных ситуаций, которые зачастую сопровождаются деструктивными проявлениями, такими как агрессия. Кроме того, такие системы также могут быть подвергнуты умышленным воздействиям, провоцирующим пользователей на деструктивное поведение. Для обеспечения устойчивого и безопасного функционирования таких систем необходима разработка и внедрение новых эффективных методов автоматического выявления проявлений человеческой агрессии в сложных социокиберфизических пространствах.

Человеческая агрессия представляет собой сложный психологический и социальный феномен. Л. Берковиц определяет агрессию как мотивированную двигательную активность, которая направлена на причинение повреждений, ущерба или дискомфорта объекту нападения, который может быть одушевленным или неодушевленным [1]. С другой стороны, А. Бандура подчеркивает, что агрессия есть деструктивное поведение, нарушающее социальные нормы [2]. Отечественный исследователь С.Н. Ениколопов объединяет эти два определения: «агрессия – это целенаправленное деструктивное и наступательное поведение, нарушающее нормы и правила сосуществования людей в обществе, наносящее вред объектам нападения (одушевленным и неодушевленным), причиняющее ущерб или вызывающее у них психологический дискомфорт: отрицательные переживания, состояния напряженности, страха, подавленности и др.» [3]. Таким образом, приведенные определения подчеркивают поведенческую составляющую агрессии. А. Басс [4] предложил одну из наиболее полных классификаций, в которой он выделил основные виды агрессии, представляющие собой оппозиционные пары: физическая-вербальная, активная-пассивная, прямая-косвенная. При этом, данные виды агрессии могут сочетаться друг с другом, например, физическая активная, вербальная косвенная и т.д. Стоит отметить, что проявления прямой физической и вербальной агрессии можно оценивать на основании только лишь внешнего наблюдения за поведением индивида. В противоположность этому, скрытую и пассивную агрессию можно выявить только на основе сложных комплексных методик психодиагностики. Одним из наиболее удобных способов регистрации поведения пользователей социокиберфизических пространств является видеозапись. При этом, на видеозаписях могут быть представлены следующие модальности: двигательная активность тела, сопутствующее выражение лица, невербальное речевое поведение, вербальное речевое поведение.

Современные одномодальные методы распознавания эмоций и агрессии [5, 6, 7], позволяют достичь точности распознавания в 75-99%. Однако, применение таких методов в значительной степени ограниченно наличием поступающих данных соответствующей модальности: в случае, если по какой-либо причине прекращается поступление данных соответствующей модальности, то работоспособность одномодального метода распознавания будет нарушена. Кроме того, одной модальности может быть недостаточно не только для однозначной идентификации агрессии, но и для объективной трактовки некоторой ситуации

в контексте наличия агрессивного поведения. В частности, в работе [8] продемонстрирована важность учета контекста и предыстории (информации о наличии агрессии в предшествующие моменты времени) при рассмотрении некоторой ситуации. Таким образом, для выявления агрессии целесообразнее использовать многомодальные методики ее распознавания. Кроме того, отмечается [9], что в задаче распознавания человеческой агрессии методы многомодальной обработки показывают более высокие результаты, чем методы, обрабатывающие каждую модальность по отдельности.

Для эффективного решения задачи многомодального распознавания человеческой агрессии необходима разработка методов, учитывающих сложные взаимосвязи в гетерогенных данных, представленных в различных модальностях. В статье описывается нейросетевая модель многомодального распознавания человеческой агрессии, а также метод ее обучения, основанный на обработке гетерогенной информации, включающей в себя представленные на видео изменяющиеся мимические выражения, представленная на видео двигательная активность тела индивидов, аудиозапись невербального речевого поведения, текстовое представление вербального речевого поведения. Приводятся параметры модели и результаты моделирования, включая как результаты распознавания для различных сочетаний модальностей.

Известные подходы многомодального анализа при решении задачи распознавания агрессии

В работе [10] рассматриваются проявления агрессии на железнодорожных станциях. Для распознавания используется аудио и видео модальности, отражающие агрессивное поведение людей. Для анализа видео применялся особый дескриптор, выделяющий кинетическую энергию ключевых точек тела человека. Для анализа аудио использовались коэффициенты MFCC. Для связывания информации различных модальностей была использована динамическая Байесовская сеть. На аудио-визуальном наборе данных, собранном авторами работы, разработанная система показала 77% точности распознавания. Однако в данном подходе имеются следующие недостатки: набор данных имеет малый размер и содержит в себе весьма ограниченный спектр агрессивных действий и ситуаций. Исследователи в работе [11] расширили представленный в [10] метод при помощи формирования представления сцен, где происходит агрессия в 3D. На наборе данных из [10] данный метод показал 89% точности. В работе [12] рассматриваются подходы связывания информации различных модальностей при помощи обработки иерархической скрытой марковской моделью данных, полученных с различных сенсоров (давления, ультразвука, видеокамер и аудио). Для итоговой классификации были использованы Байесовские сети и SVM. Данный подход показал 92% точности распознавания на узкой специфической задаче распознавания агрессивного ажитированного поведения у пациентов, больных дименцией. Исследователи [13] предлагают подход, рассматривающий многомодальную детекцию и распознавания фрагментов голливудских фильмов, содержащих сцены насилия. Данный подход состоит из двух этапов. На первом этапе происходит выделение признаковых представлений аудио и видео модальностей с последующей классификацией при помощи метода k-NN. В качестве методов выделения признаков видео

были использованы следующие дескрипторы: дисперсия ориентации движений, усредненные дескрипторы движения, расширенный набор дескрипторов Хаара [14]. Для выделения признаков аудио был использован набор из 12 низкоуровневых дескрипторов, выполняющих выделение спектральных и статистических параметров аудиосигнала. Данный подход позволяет достичь 83% точности на наборе данных, собранных авторами. Ряд работ [8, 9, 15, 16, 17, 18, 19] исследовательской группы Технологического университета Делфта (Нидерланды) рассматривает различные аспекты многомоальной обработки данных в задаче распознавания человеческой агрессии. В основе этих работ лежат онтологические модели предметной области агрессии, проявляемой в железнодорожном транспорте в Нидерландах с учетом конкретной ситуации, методология составления набора данных, содержащего проявления агрессии, методы связывания данных различных модальностей, модели машинного обучения, выполняющие распознавание, результаты экспериментальной проверки разработанных моделей на собранном наборе данных. В работе [20] была предложена система распознавания агрессивных действий человека на основании обработки данных различных модальностей, полученных при помощи сенсора Kinect: координат ключевых точек тела человека, RGB кадров видео, а также аудиозаписей. В данном методе обработка данных происходит в два этапа: на первом этапе выполняется классификация признаков отдельных модальностей при помощи SVM, а на втором выполняется усреднение результатов SVM классификаторов отдельных модальностей. Данный подход показал 90% точности распознавания на наборе данных, собранном авторами. Однако, данный подход имеет свои ограничения: методы извлечения признаков, используемых для классификации, представляют собой сложные и не эффективные с вычислительной точки зрения процедуры, задаваемые вручную. В работах [21, 22, 23] были предложены концептуальные модели многомодальной обработки данных в задаче распознавания агрессии в социокиберфзических пространствах, однако в данных работах не представлены результаты экспериментальных исследований.

Все рассмотренные методы имеют следующие общие особенности и недостатки: набор модальностей обычно ограничивается двумя, максимум, тремя, не берется во внимание все обозначенное выше множество модальностей, которые могут быть представлены на видео; большинство методов требуют одновременной обработки информации всех модальностей; разработанные модели тестировались на данных, собранных самими авторами, не находящихся в публичном доступе и имеющих малый размер; не достаточное внимание уделяется распознавания в условиях недостатка или отсутствия той или иной модальности, а также при различных сочетаниях модальностей.

Описание предлагаемой модели

В рамках данного исследования предлагается развитие нейросетевой модели многомодальной обработки информации, впервые представленной в работе [23]. Рассматриваемая модель не требует одновременной обработки информации всех модальностей для выполнения распознавания. Это свойство позволяет применять ее для распознавания в условиях недостатка или отсутствия информации какой-либо модальности.

Важным аспектом обработки многомодальной информации является выделение такой значимой информации из исходных данных различных модальностей, которая бы была инвариантной к виду модальности и зависела бы только от той категории, которая была присвоена экземпляру в ходе разметки данных. Для выделения такой информации целесообразно формировать т.н. промежуточные признаковые представления, где векторы признаков будут удовлетворять требованию инвариантности относительно модальности экземпляра.

Рассмотрим подробнее принцип работы рассматриваемой модели. Обработка данных различных модальностей моделью выполняется в два этапа: сначала по отдельности обрабатываются данные X_j разных модальностей соответствующими нейронными сетями с параметрами θ_j, которые формируют на своем выходе промежуточное признаковое представление Y_j, где j – порядковый номер модальности. Промежуточное представление $\tilde{Y}_j$ необходимо для того, чтобы обеспечить инвариантность признакового представления относительно вида модальности и сделать его зависимым только от распознаваемого класса. Полученное промежуточное представление $\tilde{Y}_j$ подается на вход нейронной сети с параметрами $\bar{\theta}$, которая выполняет итоговую классификацию каждого экземпляра $\tilde{Y}_j$ соответствующей модальности, формируя на выходе вектор вероятностей прогнозов распознанных классов $\hat{Y}_i$. При этом инвариантность результатов распознавания относительно вида модальности обеспечивается в ходе процесса обучения, который заключается в том, что выходной нейросетевой классификатор настраивает свои параметры $\bar{\theta}$ на основании обработки всех экземпляров данных всех модальностей при помощи метода обратного распространения ошибки. В свою очередь, параметры нейронных сетей, выполняющих обработку отдельных модальностей θ_j также обновляются на основании ошибки, получаемой на выходе классификатора с параметрами $\bar{\theta}$ и распространяемой на эти нейронные сети с параметрами θ_j. Архитектура модели приведена на рис. 1.

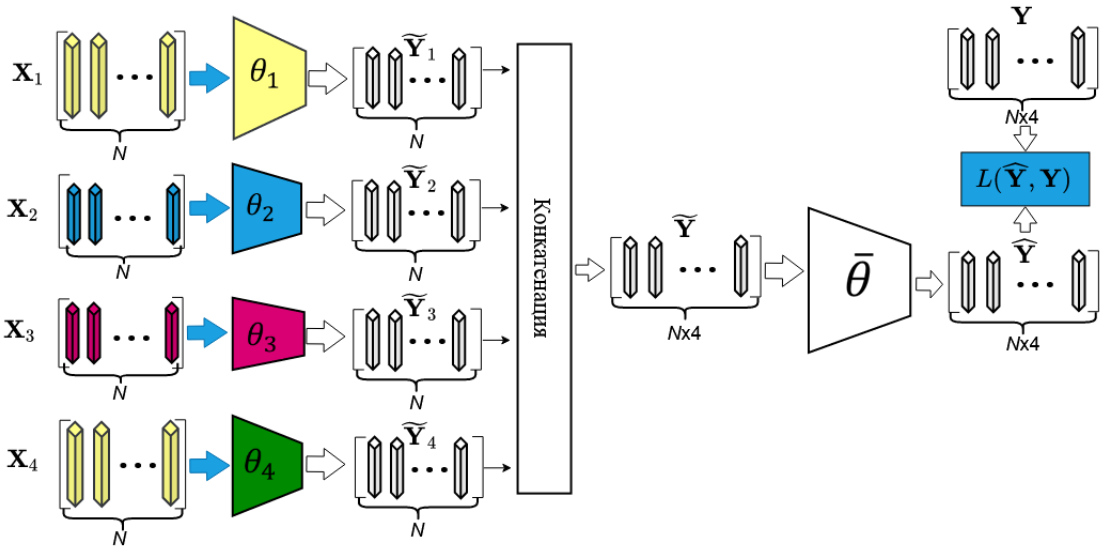

Рис. 1. Архитектура предлагаемой модели

Обучение описываемой модели выполняется при помощи пакетного стохастического градиентного спуска с пакетами малого размера (Mini-Batch Stochastic Gradient Descend). Это означает, что нейронной сетью обрабатывается

не отдельные экземпляры данных, а т.н. пакеты, состоящие из нескольких экземпляров, для каждого экземпляра в пакете формируется отдельный вывод, а значения градиентов функции потерь суммируются по обрабатываемому пакету данных. Именно применение такого типа организации обработки данных открывает возможности для формирования инвариантных относительно модальности промежуточных представлений $\tilde{Y}$.

Пакетный градиентный спуск подразумевает одновременную обработку нейросетью не отдельных экземпляров данных, а т.н. пакетов данных, состоящих из множества отдельных экземпляров данных, случайно отобранных из обучающей выборки. При этом, выходная ошибка нейронной сети, получаемая на выходе, усредняется для всего обрабатываемого пакета. Таким образом, при обработке данных таким образом происходит учет и сглаживание их неоднородности, что в итоге приводит к повышению устойчивости обучения на сложных выборках большого объема.

Обработка многомодальной информации выполняется следующим образом: на вход каждой нейронной сети поступает пакет данных соответствующей модальности X_j, который формируется путем конкатенации отдельных экземпляров данных этой модальности $X_j = [x_{j1}, x_{j2}, \ldots, x_{jN}]$, где $X_j \in R^{d_j N}$, j — индекс модальности, N — размер пакета, d_j — размерность данных j-й модальности. На выходе каждой нейронной сети θ_j формируется пакет соответствующих промежуточных представлений $\tilde{Y}_j = [\tilde{y}_{j1}, \tilde{y}_{j2}, \ldots, \tilde{y}_{jN}]$. Далее формируется общий пакет промежуточных представлений с помощью конкатенации пакетов промежуточных представлений отдельных модальностей $\tilde{Y} = [\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_M]$, где M — количество обрабатываемых модальностей. Далее Y подается на вход нейросетевого классификатора с параметрами $\bar{\theta}$, который формирует на выходе пакет $\hat{Y} = [\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_M]$, содержащий результаты распознавания для всех модальностей $\hat{Y}_j = [\hat{y}_{j1}, \hat{y}_{j2}, \ldots, \hat{y}_{jN}]$. Для $\hat{Y}$ вычисляется функция потерь перекрестная энтропия (cross-entropy loss function) по всем экземплярам результатов распознавания для всех модальностей:

$$L(Y,Y) = -\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{C} y_{ijk} \cdot \log(y_{ijk}) \tag{1}$$

где M — количество модальностей, N — размер пакета, C — количество классов. Именно учет всех модальной в функции потерь, которая определяет величину коррекции параметров как выходного классификатора $\bar{\theta}$, так и нейронной сети каждой отдельной модальности как раз и обеспечивает формирование инвариантного относительно модальности промежуточного признакового пространства $\tilde{Y}$. Подробнее процесс обучения предлагаемой нейросетевой модели приведен в псевдокоде ниже.

Ввод: K — количество пакетов, на которые разбивается обучающая выборка M — количество модальностей θ_j — веса нейронной сети, выполняющей обработку j-й модальности, $\bar{\theta}$ — веса выходного нейросетевого классифкатора.

while $L(y_{ij}, \hat{y}_{ij}) \neq \min$ **do**
  for $i = 1 \ldots K$ **do**
    for $j = 1 \ldots M$ **do**
      Получить пакет данных X_j
      Выполнить прямое распространение НС j-й модальности $\tilde{Y}_j = f(\theta_j | X_j)$
    end for
    Выполнить конкатенацию промежуточных представлений всех модальностей $\tilde{Y} = [\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_M]$

Выполнить прямое распространение для выходного классифкатора $\hat{Y} = f(\bar{\theta}|\tilde{Y})$
Вычислить ошибку $L(\hat{Y}, Y)$
Обратное распространение ошибки и обновление весов выходного классификатора $\bar{\theta} = \bar{\theta} - \eta \nabla_{\bar{\theta}} L(\hat{Y}, Y)$
for j = 1... M **do**
Выполнить обратное распространение ошибки $L(\hat{Y}, Y)$ и обновить веса НС j-й модальности $\theta_j = \theta_j - \eta \nabla_{\theta_j} L(\hat{Y}, Y)$
end for
end for
end while

Для выполнения многомодального распознавания выполняется усреднение оценок прогнозируемых меток классов $\hat{y}_i$ по всем представленным модальностям

$$\hat{y} = \frac{1}{M} \sum_{i=1}^{M} \hat{y}_i \tag{2}$$

где M — количество модальностей.

Рассмотрим теперь подробнее архитектурные особенности каждого элемента $(\theta_1 - \theta_4, \bar{\theta})$ представленной нейросетевой модели. Выделение изображений лиц выполнялось при помощи метода MTCNN [24], который способен выполнять локализацию лиц на цифровых изображениях с высокой точностью в реальном масштабе времени, а также способен работать с высокой точностью в условиях шума и прочих искажений цифровых изображений. Выделенные области цифровых изображений, содержащие изображения лиц далее обрабатываются сверточной нейронной сетью.

В качестве модели, использующейся для экстракции признаков мимических проявлений эмоций (нейронная сеть с параметрами θ_1 на рис. 1), была выбрана глубокая сверточная нейронная сеть, состоящая из двух частей: первая часть, называемая экстрактором признаков и служащая для выделения наиболее значимой информации о мимических проявлениях на кадрах видео. Она представляет собой глубокую трехмерную сверточную нейросеть архитектуры ShuffleNet [25], которая была предварительно обучена на базе изображений лиц AffectNet [26], содержащем более миллиона изображений, на которых представлены мимические проявления человеческих эмоций. В ходе обучения многомодальной системы параметры экстрактора признаков статичных кадров фиксируются и не изменяются в ходе обучения. Вторая часть состоит из четырех последовательно идущих друг за другом слоев трехмерной свертки (3D CNN), выполняющих пространственно-временной анализ динамики выделенных мимических признаков. При этом, слои свертки чередуются слоями пакетной нормализации [27] и активационными функциями ReLU [28]. Архитектура описываемой нейронной сети приведена на рис. 2.

В качестве модели, выполняющей выделение признаков двигательной активности тела на видео (нейронная сеть с параметрами θ_2 на рис. 1) применялась 3D CNN архитектуры Residual 3D Neural Network (R3D) [29, 30], которая является трехмерным развитием глубокой нейросетевой архитектуры ResNet [31], выполняющей классификацию визуальных объектов на цифровых изображениях. Основной архитектурной составляющей этой сети является т.н. Residual Block, главной отличительной особенностью которого является учет на выходе блока

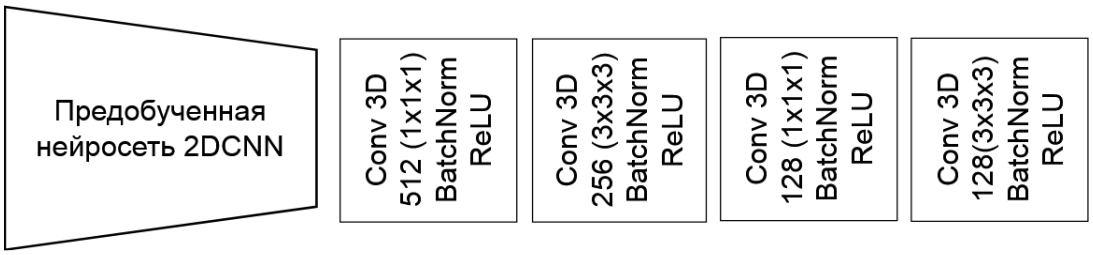

Рис. 2. Архитектура нейронной сети выделения мимических признаков в видео.

входной карты признаков при помощи суммирования входа и выхода. Пример Residual блока приведен на рис. 3.

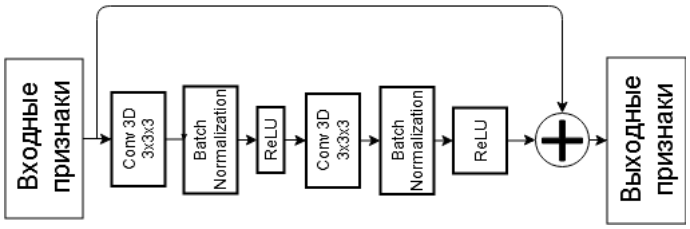

Рис. 3. Архитектурные особенности Residual блока.

Сверточная часть Residual блока состоит из последовательно двух блоков, состоящих из сверточного слоя, слоя пакетной нормализации [27]. На выходе блока расположена активационная функция ReLU. На рисунке 4 приведена архитектура R3D, состоящая из восьми последовательно идущих друг за другом Residual блоков. Количество каналов, генерируемое каждым блоком, также отражено на рис. 4.

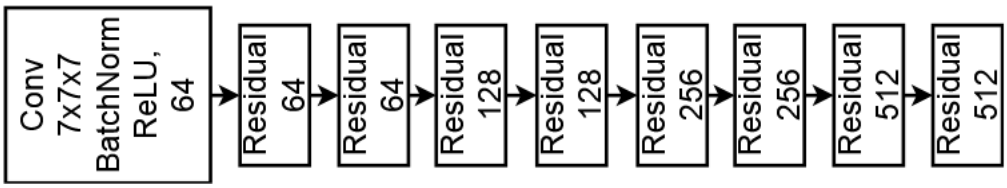

Рис. 4. Архитектура R3D для выделения признаков двигательной активности тела в видео.

Нейронная сеть данной архитектуры способна обрабатывать последовательности RGB кадров видео размером 112x112 пикселей произвольной длины. В качестве экстрактора признаков было взято 6 первых Residual блоков R3D, обученной на наборе данных Kinetics [32]. Параметры экстрактора признаков фиксировались и не изменялись в процессе обучения. Параметры остальных двух Residual блоков заново инициализировались и изменялись в ходе обучения.

В качестве данных, используемых для обработки модальности невербального речевого поведения были выбраны спектральные портреты или спектрограммы записей фраз, произнесенных испытуемыми. Спектрограмма представляет собой последовательность оконных преобразований Фурье, выполненных над исходным аудиосигналом речи человека. При этом, использовалось окно Хэмминга при оконном преобразовании с перекрытиями между окнами, составляющими половину временной длительности окна. Полученный спектральный портрет далее приводился

к мел-шкале, которая соответствует шкале восприятия человеческой речи слуховым анализатором. Далее выполнялся перевод амплитудных значений полученных спектрограмм в децибельные значения путем получения энергии спектра или квадрата его значения и последующего логарифмирования результата.

В качестве модели выделения признаков невербального речевого поведения (нейронная сеть с параметрами θ_1 на рис. 1) использовалась глубокая сверточная нейронная сеть архитектуры 11-слойной архитектуры VGG [33], архитектура которой приведена на рис. 5.

Рис. 5. Архитектура VGG11 для выделения признаков невербального речевого поведения в аудио.

Данная нейронная сеть способна выполнять обработку спектрограмм аудиосигнала речи человека. Спектрограмма представляет собой последовательность энергетических значений оконных преобразований Фурье, выполненных над исходным аудиосигналом и приведенных к мел-шкале. Обрабатываемая спектрограмма масштабируется к размеру 224x224 элемента для соблюдения соответствия входному размеру нейронной сети.

Для выделения признаков в тексте была использована двунаправленная рекуррентная нейронная сеть архитектуры длительной краткосрочной памяти (Bidirectional Long Term Short Memory – Bi-LSTM) [34, 35]. Данная нейронная сеть имеет следующие гиперпарапметры: размер входа равен 4039; 2 скрытых слоя; размер скрытого состояния равен 256.

Архитектура выходного классификатора (нейронная сеть с параметрами $\bar{\theta}$ на рисунке 1), представляет собой трехслойную полносвязную нейронную сеть прямого распространения. Первый слой имеет 512 входных и 128 выходных элементов, второй — 128 входных и 32 выходных элемента, третий — 32 входных и 2 выходных элемента. Архитектура нейронной сети приведена на рис. 6.

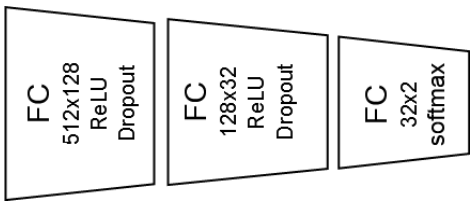

Рис. 6. Архитектура выходного классификатора.

Все слои, кроме выходного, имеют активационную функцию ReLU. После активационных функций каждого слоя расположены слои прореживания (dropout) [36] для борьбы с переобучением. Выходной слой нейронной сети имеет активационную функцию softmax:

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{3}$$

где $\hat{y}_i$ — это значение i-го нейрона после применения активационной функции z_i — это значение взвешенной суммы i-го нейрона, C — количество классов.

Описание обучающих данных

В качестве обучающих данных был выбран набор многомодальных данных IEMOCAP [37]. Данный набор был составлен в 2008 году в университете Южной Калифорнии. Этот набор состоит из аудио, визуальных и текстовых данных, содержащих текст реплик говорящих людей, записи их поведения на видео и аудио, а также данные о движении ключевых точек на лице, голове и руках субъектов. При этом использовались технологии и специализированное оборудование для захвата движений (motion capture). Данный набор данных состоит из пяти сессий записи. В состав каждой сессии входит от 9 до 15 записей диалогов пары актеров – мужчины и женщины. Всего в записях диалогов для рассматриваемого набора данных принимало участие 10 человек. Каждая запись представляет собой импровизацию или игру по сценарию того или иного тематического диалога, в котором проявлялись эмоции пользователей. Разметка эмоций выполнялась следующим образом: сначала видеозапись разбивалась на временные фрагменты, соответствующие отдельным репликам участников диалогов. Всего было выделено 10305 фрагментов, средняя продолжительность фрагментов составляет 4.5 с. Далее выполнялась экспертная оценка каждого отдельного фрагмента на соответствие эмоции из набора счастье, гнев, грусть, волнение, фрустрация и нейтральное состояние. Кроме того, учитывался также субъективный отчет испытуемых о своем эмоциональном состоянии. Окончательное решение о присвоении той или иной метки класса соответствующей реплике принималось на основании мажоритарного голосования экспертов. В результате было получено 7380 размеченных фрагмента, у которых наблюдалась достаточная степень согласованности мнений экспертов. На рис. 7. приведена диаграмма распределения размеченных экземпляров по меткам эмоций. На рис. 8. приведены примеры кадров из набора данных IEMOCAP.

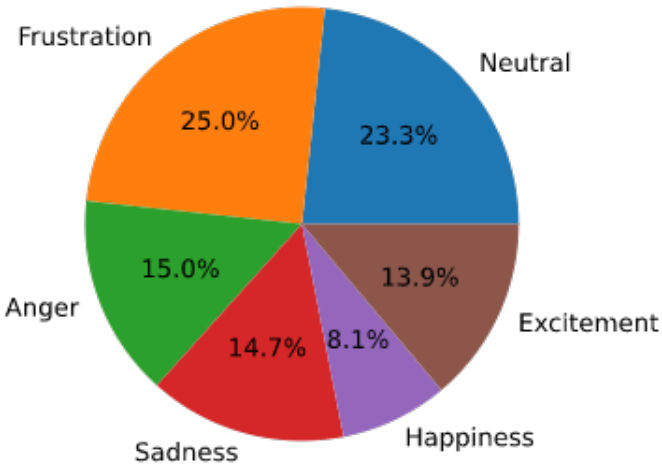

Рис. 7. Соотношение экземпляров размеченных данных по эмоциям.

Данный набор был изначально размечен для классификации эмоций. Однако, сценарии в этом наборе данных, содержащие проявления гнева, построены на конфликте, изображаемом актерами. Для того, чтобы использовать IEMOCAP для распознавания агрессии экземпляры, содержащие гнев, были дополнительно

Рис. 8. Примеры кадров из набора IEMOCAP.

проверены на наличие там такого поведения, которое может быть рассмотрено как агрессивное. В качестве такого поведения рассматривались оскорбления, насмешки, ирония, сарказм, угрозы, унижения, перебивания, сообщения в повелительном тоне, приказы замолчать и т.д. Всего было отобрано 1054 экземпляра, удовлетворяющих агрессивному поведению. Для представления не агрессивного поведения было также выбрано 1054 экземпляра, содержащих реплики не относящиеся к проявлениям агрессии. Из собранного набора в качестве обучающей выборки было отобрано случайным образом 1686 экземпляров. Остальные 422 экземпляра были взяты для тестирования модели.

Организация эксперимента

Эксперимент организован следующим образом. Сначала выполняется обучение предлагаемой модели на обучающей выборке, состоящей из данных всех модальностей, представленных в IEMOCAP: видео, содержащих мимические выражения (лицо), видео, содержащих двигательную активность тела (тело), невербального речевого поведения (звук), представленного в аудиосигнале, текста реплик пользователей (текст). Затем, на тестовой выборке выполняется тестирование обученной модели распознавать проявления агрессии на различных сочетаниях модальностей:

1) Лицо, тело, звук, текст
2) Лицо, тело, звук
3) Лицо, тело, текст
4) Лицо, звук, текст
5) Тело, звук, текст
6) Лицо, тело
7) Лицо, звук
8) Лицо, текст
9) Тело, звук

10) Тело, текст
11) Звук, текст
12) Лицо
13) Тело
14) Звук
15) Текст

Таким образом, покрываются все возможные сочетания, которые могут быть в данном наборе модальностей. Обучение выполнялось в течение 200 эпох. Размер пакета в алгоритме пакетного градиентного спуска было выбрано равным 16. Для тестирования модели использовались настроенные параметры модели, полученные после прохождения 47 эпохи обучения. Это было обусловлено тем, что после 47 эпохи было достигнуто минимальное значение функции потерь (2) на тестовой выборке. Обучение выполнялось с использованием библиотеки PyTorch [38].

Методы борьбы с переобучением

Для борьбы с переобучением выполнялась аугментация данных. При обработке нейронными сетями видео, содержащих изображения лиц и двигательную активность тела, над каждым кадром выполнялись следующие операции: аффинные геометрические преобразования, изменение пропорций изображения по горизонтальной или вертикальной, или обоим осям одновременно, масштабирование (уменьшение и увеличение) изображений в 0.75, 0.9, 1.15 и 1.25 раза, поворот на случайное количество градусов из промежутка от -15 до +15 градусов, вырезание из изображения случайного фрагмента. лиц. В качестве временной аугментации выполнялся выбор случайного фрагмента в записи, состоящего из последовательных кадров фиксированной длины. Для аугментации аудио фрагментов применялись такие приемы как выбор случайного фрагмента аудиосигнала, заполнение нулями фрагментов спектрограммы, заполнение спектрограммы последовательностями преобразованного речевого сигнала, длина которого меньше размера спектрограммы. При этом, конкретный вид аугментации выбирался случайным образом.

Применение аугментации в ходе процесса обучения глубоких нейронных сетей методом стохастического градиентного спуска расширяет обучающую выборку и позволяет увеличить репрезентативность обучающих данных, что в итоге делает процесс градиентного спуска более устойчивым, расширяет обобщающие способности нейронных сетей и является эффективным методом решения проблемы переобучения глубоких нейросетевых моделей.

Метрики

Так как в данной работе будет рассмотрено несколько моделей для изучения эффективности распознавания эмоциональной составляющей двигательной активности человека в видеопотоке, то необходимо определить метрики для оценки качества обучения и проведения сравнительного анализа.

В данной работе будут использоваться следующие метрики.

1) Доля корректных прогнозов (accuracy).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

2) Мера точности (precision).

$$precision = \frac{TP}{TP+FP} \tag{5}$$

3) Мера полноты (recall) или доли истинно положительных распознаваний (True Positive Rate – TPR)

$$recall = TPR = \frac{TP}{TP+FN} \tag{6}$$

4) F1-мера (F1-score) или среднее гармоническое precision и recall:

$$F1-score = 2\frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP+\frac{TP+FN}{2}} \tag{7}$$

где TP — количество истинно положительных распознаваний, TN — количество истинной отрицательных распознаваний, FP — количество ложноположительных распознаваний, FN — количество ложноотрицательных распознаваний. Результаты экспериментов сведены в таблицу.

Таблица

Результаты экспериментального исследования для различных конфигураций модальностей

Конфигурация (сочетание модальностей)	Accuracy	Recall (TPR)		Precision		F1-score	
		0	1	0	1	0	1
1. Лицо, тело, звук, текст	0.818	0.813	0.825	0.872	0.750	0.841	0.786
2. Лицо, тело, звук	0.801	0.798	0.806	0.859	0.729	0.827	0.765
3. Лицо, тело, текст	0.763	0.802	0.720	0.761	0.766	0.781	0.742
4. Лицо, звук, текст	0.801	0.795	0.801	0.863	0.723	0.828	0.764
5. Тело, звук, текст	0.806	0.809	0.801	0.850	0.750	0.829	0.775
6. Лицо, тело	0.737	0.773	0.695	0.744	0.729	0.758	0.712
7. Лицо, звук	0.794	0.763	0.853	0.910	0.649	0.830	0.737
8. Лицо, текст	0.732	0.769	0.690	0.739	0.723	0.754	0.706
9. Тело, звук	0.789	0.784	0.796	0.855	0.707	0.818	0.749
10. Тело, текст	0.742	0.802	0.684	0.709	0.782	0.753	0.730
11. Звук, текст	0.799	0.794	0.805	0.859	0.723	0.825	0.762
12. Лицо	0.680	0.700	0.651	0.739	0.606	0.719	0.628
13. Тело	0.701	0.745	0.653	0.701	0.702	0.722	0.677
14. Звук	0.773	0.748	0.819	0.889	0.628	0.812	0.711
15. Текст	0.718	0.775	0.662	0.692	0.750	0.731	0.703

Для метрик Recall, Precision, F1-score результаты были получены для распознавания отдельных классов: агрессивное поведение (обозначено в таблице меткой 0) и не агрессивное поведение (обозначено в таблице меткой 1).

Свои лучшие результаты распознавания модель показала при обработке всех четырех модальностей. При этом, алгоритм показал сравнимые результаты для обоих классов (recall для агрессии составил 0.813, а для другого поведения 0.825). При этом, доля истинно распознанных агрессивных действий среди всех распознанных действий как агрессивных заметно превышает (precision=0.872) обратную ситуацию (precision=0.750) и показывает самые высокие показатели для всех сочетаний модальностей.

Результаты распознавания моделью сочетаний из трех модальностей (конфигурации 2–5) достигли сравнимых результатов, за исключением, конфигурации «лицо, тело, текст» (accuracy=0.763). Также данное сочетание показало заметное снижение относительно других сочетаний в распознавании не агрессивного поведения. В то же самое время, сочетание «лицо, тело, текст» показывает сравнительную с другими сочетаниями результативность при распознавании агрессии (recall=0.802) при заметно более низких результатах распознавания не агрессивного поведения (recall=0.720). В целом, для конфигураций 2 – 5 модель показала большие значения precision для распознавания агрессивного, чем для распознавания не агрессивного поведения.

Для сочетаний из двух модальностей (конфигурации 6–11) наблюдается некоторое снижение общих результатов распознавания по метрике accuracy. При этом, особо выделяются лучшие показатели сочетаний модальностей, где присутствует невербальное речевое поведение («звук») по сравнению с другими. В то же самое время среди худших показателей выделяются сочетания модальностей, в которых присутствует модальность мимических выражений лица. Также стоит особо отметить, что сочетание «лицо, звук» показало максимальное значение precision для распознавания агрессии (0.91) при заметно более низких показателях для распознавания не агрессивных действий (precision=0.649). Также стоит отметить, что для конфигураций из двух модальностей, в тех модальностях, в которых присутствует «звук», значение recall для распознавания не агрессивного поведения превышает таковое для агрессивного.

Исследуемая модель показала высокие результаты при обработке отдельных модальностей (конфигурации 12–15). При этом, среди отдельных модальностей лучшие результаты показывает модальность «звук» (accuracy=0.773), а худшие – модальность «лицо» (accuracy=0.68). Также стоит отметить, что отдельные модальности лучше распознают агрессивное поведение, чем не агрессивное (значение recall выше для агрессивного, чем для не агрессивного поведения), за исключения модальности «звук», где ситуация обратная. При распознавании агрессивного поведения, значения precision у модальностей «лицо» и «звук» выше, чем при распознавании не агрессивного поведения. Обратная ситуация наблюдается для модальностей «тело» и «текст», где значение precision при распознавании агрессивного поведения ниже, чем при распознавании не агрессивного.

Высокие результаты распознавания модальности невербального речевого поведения обусловливают более высокие результаты в многомодальном распознавании, где присутствуют эта модальность в случае комбинации двух и более модальностей. Обратное можно сказать про модальность мимических выражений, которая показала худшие значения accuracy для одномодального распознавания. В сочетаниях, где она присутствует, в целом, наблюдаются более низкие результаты распознавания. Такие результаты могут быть обусловлены качеством данных той или иной модальности. Например, извлекаемые из видео в

наборе IEMOCAP изображения лиц имеют низкое разрешение, не самый удачный ракурс съемки, при котором может не быть представлено всей полноты мимических проявлений.

Также важно отметить, что распознавание моделью агрессивного поведения, в целом, выполняется лучше, чем не агрессивного, за исключением звуковой модальности, о чем было сказано выше. Это может быть вызвано балансом экземпляров различных классов в обучающей выборке, где половину составили экземпляры агрессивного поведения, которое, во многом, было ассоциировано с проявлениями гнева, а вторую половину составили случайно отобранные экземпляры, размеченные в оригинальном наборе IEMOCAP другими категориями.

Результаты распознавания отдельных модальностей наглядно демонстрируют, что используемая модель расширяет свое применение на различные сочетания обрабатываемых модальностей, включая и одномодальные случаи. При этом, существенное падение результатов распознавания наблюдается именно в одномодальных случаях.

Заключение

Рассматриваемая в статье модель в результате ее апробации на многомодальном наборе IEMOCAP показала высокие результаты распознавания агрессивного поведения людей при обработке данных четырех модальностей – видео, содержащие мимические выражения, видео, содержащие двигательную активность тела человека, аудиосигнал, на котором представлено невербальное речевое поведение, текста реплик пользователей, отражающего вербальное речевое поведение. При этом, распознавании агрессии по всем четырем модальностям было достигнуто 81.8% верно выполненных распознаваний. При исследовании классификации агрессии моделью во всех возможных комбинациях обозначенных модальностей были достигнуты результаты, которые существенно не ухудшаются относительно результатов распознавания по всем четырем модальностям. Особо стоит отметить следующие сочетания модальностей: двигательная активность тела, невербальное поведение, вербальное поведение (80.6% верно выполненных распознавания) и комбинацию невербально поведение, вербальное поведение (79.9% верно выполненных распознавания). Кроме того, отмечается достаточно высокие показатели распознавания агрессии по отдельным модальностям, где особо выделяется невербальное речевое поведение (77.3% верно выполненных распознаваний). Таким образом, эксперименты показали, что рассматриваемая модель не только позволяет эффективно распознавать проявления агрессии по всему набору модальностей, но также способна эффективно распознавать агрессию в условиях недостатка или отсутствия информации какой-либо модальности или даже нескольких модальностей одновременно.

В качестве дальнейших исследований планируется выполнить обучение предложенной модели других наборах данных, выполнить кросс-корпусные исследования для оценки обобщающих способностей модели, выполнить обучение модели для различных конфигураций модальностей. Рассматриваются также следующие возможности модификации самой модели: вычисление и обратное распространение ошибки по каждой модальности отдельно с балансировкой влияния

этих ошибок на обучающий процесс; применение сиамских нейронных сетей [39], генеративных состязательных моделей [40].

Конкурирующие интересы. Конфликтов интересов в отношении авторства и публикации нет.

Авторский вклад и ответсвенность. Автор участвовал в написании статьи и полностью несет ответственность за предоставление окончательной версии статьи в печать.

Список литература/References

[1] Berkowitz L., *Aggression: Its causes, consequences, and control*, Mcgraw-Hill Book Company, 1993, 158 pp.

[2] Bandura A., *Aggression: A social learning analysis.*, prentice-hall, 1973.

[3] Ениколопов С. Н., "Понятие агрессии в современной психологии", *Прикладная психология*, 2001, № 1, 60–72. [Enikolopov S. N., "Ponyatie agressii v sovremennoy psikhologii", *Prikladnaya psikhologiya*, 2001, № 1, 60–72 (in Russian)].

[4] Buss A. H., *The psychology of aggression.*, Wiley, 1961.

[5] El Ayadi M., Kamel M. S., Karray F., "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, **44**:3 (2011), 572–587.

[6] Trigeorgis G. et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network", *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*., IEEE, 2016, 5200–5204.

[7] De Souza F. D. M. et al., "Violence detection in video using spatio-temporal features", *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on.*, IEEE, 2010, 224–230.

[8] Lefter I., Rothkrantzac L.J.M., Burghoutsb G.J., "A comparative study on automatic audio–visual fusion for aggression detection using meta-information", *Pattern Recognition Letters*, 2010, 1953–1963.

[9] Lefter I. et al., "Addressing multimodality in overt aggression detection", *International Conference on Text, Speech and Dialogue*, Springer, Berlin, Heidelberg, 2010, 25–32.

[10] Zajdel W. et al., "CASSANDRA: audio-video sensor fusion for aggression detection", *2007 IEEE conference on advanced video and signal based surveillance*, IEEE, 2007, 200–205.

[11] Kooij J. F. P. et al., "Multi-modal human aggression detection", *Computer Vision and Image Understanding*, **144** (2016), 106–120.

[12] Qiu Q. et al., "Multimodal information fusion for automated recognition of complex agitation behaviors of dementia patients", *2007 10th International Conference on Information Fusion*, IEEE, 1–8.

[13] Giannakopoulos T. et al., "Audio-visual fusion for detecting violent scenes in videos", *Hellenic conference on artificial intelligence*, Springer, Berlin, Heidelberg, 2010, 91–100.

[14] Giannakopoulos T. et al., "An extended set of haar-like features for rapid object detection", *Proceedings. international conference on image processing*, **1**, IEEE, 2002, I–I.

[15] Yang Z., *Multi-modal aggression detection in trains*, 2009.

[16] Lefter I., Burghouts G. J., Rothkrantz L. J. M., "Learning the fusion of audio and video aggression assessment by meta-information from human annotations", *2012 15th International Conference on Information Fusion*, IEEE, 2012, 1527–1533.

[17] Lefter I., *Multimodal Surveillance: Behavior analysis for recognizing stress and aggression.*, 2014.

[18] Lefter I., et al., "NAA: A multimodal database of negative affect and aggression", *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2017, 21–27.

[19] Lefter I., Rothkrantz L. J. M., "Multimodal cross-context recognition of negative interactions", *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2017, 56–61.

[20] Patwardhan A., Knapp G., “Aggressive actions and anger detection from multiple modalities using Kinect”, 2016.

[21] Levonevskii D. et al., “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.

[22] Уздяев М. Ю. и др., “Методы детектирования агрессивных пользователей информационного пространства на основе генеративно-состязательных нейронных сетей”, *Информационно-измерительные и управляющие системы*, **17**:5 (2019), 60–68. [Uzdiaev M. et al., “Metody detektirovaniya agressivnykh pol’zovateley informatsionnogo prostranstva na osnove generativno-sostyazatel’nykh neyronnykh setey”, *Informatsionno-izmeritel’nye i upravlyayushchie sistem*, **17**:5 (2019), 60–68 (in Russian)].

[23] 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.

[24] Zhang K., et al., “Joint face detection and alignment using multitask cascaded convolutional networks”, *IEEE Signal Processing Letters*, **23**:10 (2016), 1499–1503.

[25] Zhang X., et al., “Shufflenet: An extremely efficient convolutional neural network for mobile devices”, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 6848–6856.

[26] Mollahosseini A., Hasani B., Mahoor M. H., “Affectnet: A database for facial expression, valence, and arousal computing in the wild”, *IEEE Transactions on Affective Computing*, **10**:1 (2017), 18–31.

[27] Ioffe S., Szegedy C., “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, 2015, arXiv: 1502.03167.

[28] Nair V., Hinton G. E., “Rectified linear units improve restricted boltzmann machines”, *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, 807–814.

[29] 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*, 2017, 3154–3160.

[30] 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.

[31] He K., et al., “Deep residual learning for image recognition”, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–778.

[32] Kay W., et al., “The kinetics human action video dataset”, 2017, arXiv: 1705.06950.

[33] Simonyan K., Zisserman A., “Very deep convolutional networks for large-scale image recognition”, 2014, arXiv: 1409.1556.

[34] Hochreiter S., Schmidhuber J., “Long short-term memory”, *Neural computation*, **9**:8 (1997), 1735–1780.

[35] Schuster M., Paliwal K. K., “Bidirectional recurrent neural networks”, *IEEE transactions on Signal Processing*, **45**:11 (1997), 2673–2681.

[36] Srivastava N., et al., “Dropout: a simple way to prevent neural networks from overfitting”, *The journal of machine learning research*, **15**:1 (2014), 1929–1958.

[37] Busso C., et al., “IEMOCAP: Interactive emotional dyadic motion capture database”, *Language resources and evaluation*, **42**:4 (2008), 335.

[38] *https://pytorch.org/*.

[39] Chicco D., “Siamese neural networks: An overview”, *Artificial Neural Networks*, 2020, 73–94.

[40] Goodfellow I., et al., “Generative adversarial nets”, *Advances in neural information processing systems*, 2014, 2672–2680.

Vestnik KRAUNC. Fiz.-Mat. Nauki. 2020. vol. 33. no. 4. P. 132–149. ISSN 2079-6641

MSC 62M45 Research Article

Neural network model for multimodal recognition of human aggression

M. Yu. Uzdyaev

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, Laboratory of autonomous robotic systems, 39, 14th Line, 199178, St. Petersburg, Russia.

E-mail: uzdyaev.m@iias.spb.su

Growing user base of socio-cyberphysical systems, smart environments, IoT (Internet of Things) systems actualizes the problem of revealing of destructive user actions, such as various acts of aggression. Thereby destructive user actions can be represented in different modalities: locomotion, facial expression, associated with it, non-verbal speech behavior, verbal speech behavior. This paper considers a neural network model of multi-modal recognition of human aggression, based on the establishment of an intermediate feature space, invariant to the actual modality, being processed. The proposed model ensures high-fidelity aggression recognition in the cases when data on certain modality are scarce or lacking. Experimental research showed 81.8% correct recognition instances on the IEMOCAP dataset. Also, experimental results are given concerning aggression recognition on the IEMOCAP dataset for 15 different combinations of the modalities, outlined above.

Key words: aggression recognition, behavior analysis, neural networks, multimodal data processing.

DOI: 10.26117/2079-6641-2020-33-4-132-149

Original article submitted: 19.11.2020 Revision submitted: 19.12.2020

For citation. Uzdyaev M. Yu. Neural network model for multimodal recognition of human aggression. *Vestnik KRAUNC. Fiz.-mat. nauki.* 2020, **33**: 4, 132-149. DOI: 10.26117/2079-6641-2020-33-4-132-149

Competing interests. The author declare that there are no conflicts of interest regarding authorship and publication.

Contribution and Responsibility. The author contributed to this article. The author is solely responsible for providing the final version of the article in print. The final version of the manuscript was approved by the author.

Funding. This work was supported by the Russian Foundation for Basic Research (project No. 18-29-22061_mk.