

DOI: 10.12731/2227-930X-2022-12-4-96-110
УДК 004.932.2

РАСПОЗНАВАНИЕ АКТИВНОСТИ ЧЕЛОВЕКА ПО ВИДЕОДАННЫМ

А.В. Пятаева, М.А. Мерко,
В.А. Жуковская, А.А. Казакевич

Настоящая работа посвящена решению задачи классификации вида физической активности человека по визуальным данным. Авторами предложено использование глубинных нейронных сетей с целью определения типа активности. Системы распознавания человеческой активности по видеоданным или отдельному изображению в настоящее время находят активное применение в различных областях человеческой деятельности от приложений для обучения занятиям спортом до системы контроля эффективности сотрудников предприятия, поэтому решение задачи распознавания действий человека по визуальным данным является актуальной задачей. Авторами разработан алгоритм определения типа физической активности по визуальным данным на основе моделей DenseNet121 и MobileNetV2, а затем самостоятельно построена модель глубокой нейронной сети, так как предварительно обученные сети не давали необходимой точности обнаружения типа физической активности, выполнен подбор гиперпараметров. Программная реализация модели выполнена в среде IDLE на языке программирования Python. Экспериментальные исследования, выполненные на специализированном наборе данных UCF50, содержащем 50 различных видов действий человека, подтверждают эффективность использования предложенного подхода для решения поставленной задачи. Дополнительно репрезентативность тестового набора данных увеличена с помощью видеопоследовательностей, полученных с YouTube.

***Цель** – разработка алгоритма определения физической активности человека по визуальным данным.*

Метод или методология проведения работы: *в работе использованы методы компьютерного зрения; методы глубокого обучения, а также методы объектно-ориентированного программирования.*

Результаты: *разработан алгоритм отслеживания физической активности человека по визуальным данным с применением технологий глубокого обучения.*

Область применения результатов: *применение полученных результатов целесообразно в системах мониторинга деятельности человека, например, при отслеживании преступной деятельности в работе правоохранительных органов, в медицинской диагностике, для отслеживания активности сотрудников офиса и др.*

Ключевые слова: *распознавание физической активности человека; глубокие нейронные сети; классификация действий*

RECOGNITION OF HUMAN ACTIVITY BY VIDEO DATA

A.V. Pyataeva, M.A. Merko, V.A. Zhukovskaya, A.A. Kazakevich

The paper considers the problem solution of classifying the type of physical activity of a person according to visual data. The authors propose using of deep neural networks to determine the type of activity. The recognizing human activity from video data or a single image systems are currently actively used in various areas of human activity. As the example we can take the system for monitoring the effectiveness of enterprise employees. So solving the problem of recognizing human actions from visual data is an actual task. The authors developed an algorithm for determining the physical activity type by visual data based on the DenseNet121 and MobileNetV2 models. Then the deep neural network model was built and hyperparameters were selected, because pre-trained networks did not provide the required accuracy of detecting the type of physical activity. The software implementation of the model

is made in the IDLE environment in the Python programming language. Experimental studies performed on a specialized UCF50 dataset containing 50 different types of human actions confirm the effectiveness of using the proposed approach to solve the problem. Additionally, the representativeness of the test data set was increased with the help of video sequences obtained from YouTube.

***Purpose** – development of an algorithm for determining a person's physical activity based on visual data.*

***Methodology:** in the work the methods of computer vision, deep learning methods and object-oriented programming methods were used.*

***Results:** an algorithm for tracking a person's physical activity based on visual data using deep learning technologies has been developed.*

***Practical implications:** the obtained results can be used in human activity monitoring systems, for example, in tracking criminal activity, in medical diagnostics, in tracking the activity of office employees, etc.*

***Keywords:** human physical activity recognition; deep neural networks; actions classification*

Глубинные нейронные сети
для классификации активности человека

Задача распознавания физической активности отличается от других задач классификации – распознавания объектов на изображениях тем, что необходим ряд экземпляров данных, чтобы предсказать правильное действие. Для классификации вида человеческой активности требуется обработка видеоданных, так как само понятие активности предполагает действие, продолжающиеся во времени. Кроме того, по одному кадру часто невозможно отделить один вид действия от другого, например, действие «игра в теннис» состоит из различных двигательных паттернов, в одном кадре человек бежит, в другом прыгает, в третьем заносит ракетку для удара и т.п. Особенностью обработки данных видеоряда является наличие признаков действия определенной продолжительности, а, значит, базовые классификаторы компьютерного зрения,

ориентированные на обработку одного изображения для распознавания активности как действия будут неэффективны. Рассмотрим модели глубоких нейронных сетей, позволяющие выполнять обработку видеоданных.

В работе [6] для классификации видеоданных модель запускается для каждого отдельно взятого кадра видеоролика с последующим усреднением вероятностей присутствия каждого класса действий на исследуемом видео. Такой подход назван авторами однокадровой сверочной нейронной сетью. CNN. В статье [7] авторы рассмотрели два подхода – позднее и раннее слияния, где первый подход отличался тем, что метод объединения был встроен в саму сеть, а второй, противоположный позднему слиянию, так как временная размерность канала видео объединялись до передачи в модель, что позволило первому слою обучаться определению локальных движений пикселей между соседними кадрами.

Идея модели CNN + LSTM [8], описанной в статье заключается в использовании сверточных сетей, где их выходные сигналы передаются в многослойную сеть LSTM «многие к одному» для извлечения локальных признаков каждого кадра. Другая идея [9] использования также сети LSTM вместе с готовой моделью обнаружения позы, чтобы получить ключевые точки тела человека для каждого кадра в видео, а затем передать их в рекуррентную нейронную сеть для определения действия, выполняемого в видео.

Еще один подход, предложенный в работе [10], объединяет оптические потоки с CNN для захвата движения и пространственного контекста в видео используются два параллельных потока. Пространственный поток берет один кадр из видео, после чего запускает на нем несколько ядер CNN, а затем на основе пространственной информации делает прогноз. Временной поток принимает оптические потоки каждого соседнего кадра после их слияния с использованием Early Fusion, а затем использует информацию о движении для прогнозирования. В конце выполняется усреднение по обеим предсказанным вероятностям, чтобы получить окончательные вероятности.

Таким образом, применение глубоких нейронных сетей позволяет решать задачу обработки видеоданных для оценки вида деятельности продолжающегося в некоторый период времени.

Распознавание типа физической активности

Первым этапом алгоритма распознавания типа физической активности по видеоданным является выбор модели глубокого обучения [1]. Среди множества подходящий архитектур на этом этапе использованы DenseNet121 [12] и MobileNetV2 [13], так как эти модели используют меньшее количество памяти среди других моделей в своем классе и при этом обеспечивают приемлемую точность, а в контексте решаемой задачи модели должны быть легковесными, обеспечивая эффективное время выполнения и занимать минимум памяти. В качестве входных данных каждая из моделей принимает тензоры вида «высота изображения, ширина изображения, цветовые каналы». В работе использовано цветовое пространство – *RGB*. Таким образом, размерность входного тензора модели– (64, 64, 3), где высота изображения – 64, ширина изображения – 64, каналов цвета – 3. Для проверки качества работы моделей DenseNet121 и MobileNetV2 использован набор данных UCF50, подробно описанный в разделе экспериментальных исследований настоящей статьи. Для достоверности проверки качества работы предобученных сетей модели запускались с одними и теми же гиперпараметрами, а эффективности их работы оценивалась с помощью классической метрики Accuracy, также произведен расчет потерь при обучении. Количество эпох обучения для каждой модели составило 50, однако, для обучения DenseNet121 потребовалось 33 эпохи, для MobileNetV2 – 28 эпох. Размер мини-выборки составил 4 экземпляра, в процесс обучения интегрирована возможность ранней остановки для того, чтобы постоянно отслеживать величину ошибки на каждой эпохе. Если величина ошибки не уменьшается после 15 последовательных эпох, то обучение будет остановлено с сохранением последних значений весов сети. Модель MobileNetV2 дополнена слоем сглаживания и полносвяз-

ным слоем с 256 нейронами и функцией активации ReLU. Выходной слой модели имел сигноидную функцию активации. Набор данных разделен на тренировочный, тестовый и валидационный. На рисунках 1 и 2 показаны результаты тестирования моделей DenseNet121 и MobileNetV2 соответственно. Визуальные графики построены при помощи библиотеки Matplotlib.

На рисунках 1-2 приведены графики изменения точности и потерь для каждой из предварительно обученных моделей – DenseNet121 и MobileNetV2 – с ростом числа эпох.

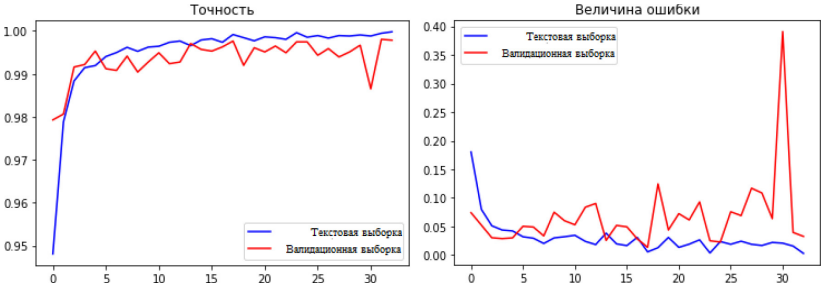


Рис. 1. График accuracy и loss модели DenseNet121

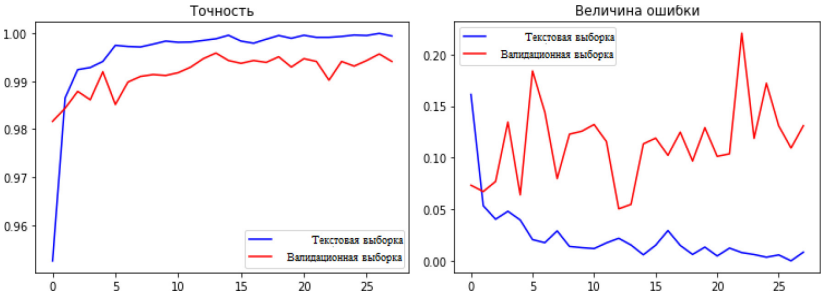


Рис. 2. График accuracy и loss модели MobileNetV2

Не смотря на высокую эффективность, продемонстрированную предобученными моделями для распознавания человеческой активности по визуальным данным, они не позволяют работать в режиме реального времени, поэтому требуется разработка модели, требующей меньше вычислительных ресурсов.

Вторым этапом работы алгоритма классификации виде активности является самостоятельное конструирование модели. Схема построенной модели изображена на рисунке 3.

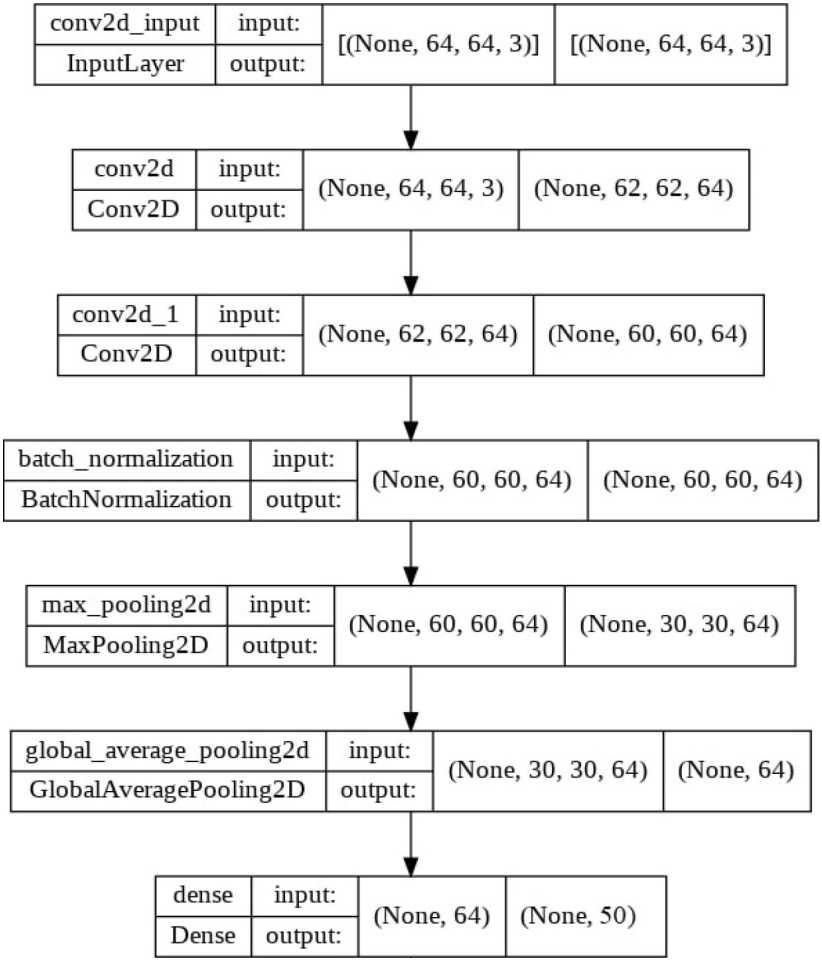


Рис. 3. Схема построенной модели сверточной нейронной сети

Разработанная модель содержит три подряд сверточных блока, позволяющих выделить значимые признаки для различных видов де-

ятельности, наблюдаемых по видеоданным. Первые два сверточных слоя модели содержат по 64 нейрона с ядром свертки 3 x 3. На третьем сверточном слое выполняется свертка с 64 фильтрами. Размеры входного слоя определяются размерами видеокадра, его высотой и шириной, img_h × img_w, а также количеством каналов изображения. Функция активации – ReLU [4]. Каскад сверточных слоев завершается слоем пакетной нормализации, за которым идут слои подвыборки: max-пулинга и average-пулинг. За ними идет полносвязный слой с 256 нейронами и функцией активации ReLU, далее полносвязный слой из 50ти нейронов по количеству видов, распознаваемых действия, с функцией активации Softmax [5]. Программная реализация модели выполнена в среде IDLE [14] на языке программирования Python [2-3].

Для обучения самостоятельно построенной модели потребовалось 28 эпох из 50 с достижением значения Accuracy в 0,99. Качество модели было проверено на валидационной выборке исходного набора данных, тех данных, которые модели в процессе обучения не видела. Как следует из рисунка 4, величина Accuracy модели на неизученных ранее данных составила 0,9855.

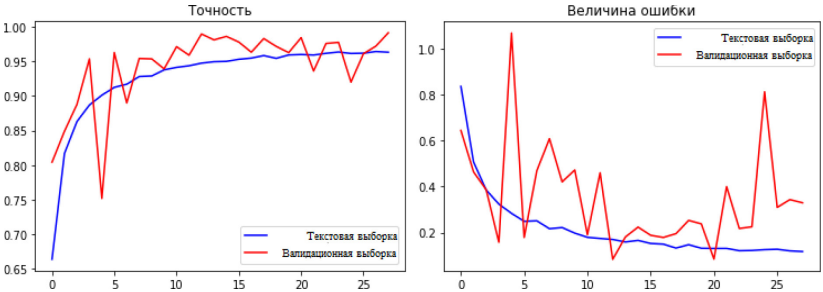


Рис. 4. График изменения точности модели и величины ошибки с ростом числа эпох обучения

Таким образом, точность предложенной модели сопоставима с рассмотренными ранее (разница не превышает 1%), а ее размер в десятки раз меньше рассмотренных моделей глубокого обучения DenseNet121 и MobileNetV2, что делает ее более предпочтительной в качестве классификатора физической активности.

Экспериментальные исследования

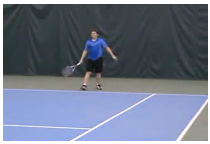
Для обучения моделей и проведения экспериментальных исследований использован набор данных UCF50 [11], содержащий 50 видов физической активности. Размер видеокадров составляет 320 x 240, число кадров каждого видеоролика различается, но в среднем составляет 150 кадров. Каждая видеопоследовательность содержит только одно действие. Дополнительно для тестирования моделей использованы видеоролики, загруженные с видеохостинга YouTube [15]. В наборе данных содержались следующие виды действий: игра в бейсбол, баскетбольная стрельба (серия ударов по корзине), жим лежа, езда на велосипеде, бильярдный удар, брасс, толчок, ныряние, игра на барабанах, фехтование, гольф, игра на гитаре, прыжок в высоту, скачки, верховая езда, вращение хулахупа, метание копья, жонглирование мячами, прыжки со скакалкой, прыжки на месте вверх, переправа на байдарках, выпады (упражнение для укрепления мышц ног), маршировование, замешивание теста, манипуляции с нунчаками, игра на фортепиано, приготовление пиццы, прыжки с шестом, прыжки на гимнастическом коне, подтягивания, удары, отжимания, скалолазание, гребля, сальса, скейтбординг, катание на лыжах, катание на гидроцикле, жонглирование футбольным мячом, качели, игра на индийской перкуссии, тай-чи (китайские боевые искусства), теннисные качели, прыжки на батуте, игра на скрипке, игра в волейбол, прогулки с собакой и игра в йо-йо. Примеры кадров использованных видеоданных и их описание приведено в таблице 1.

Таблица 1.

Описание тестовых данных

Описание тестовой видеопоследовательности	Образец кадра	Описание тестовой видеопоследовательности	Образец кадра
Категория I			
UCF50/ v_Walking-WithDog_g01_c01.avi, Количество кадров: 240 Действие: Прогулка с собакой		UCF50/ v_PlayingPiano_g01_c01.avi, Количество кадров: 210 Псевдоним: Игра на пианино	

Окончание табл. 1.

UCF50/ v_Walking-WithDog_g01_c03. avi, Количество кадров: 240 Псевдоним: Прогулка с собакой		UCF50/ v_PlayingPia-no_g16_c02.avi, Количество кадров: 210 Псевдоним: Игра на пианино	
UCF50/ v_Basket-ball_g01_c02.avi, Количество кадров: 180 Псевдоним: Баскет-бол		UCF50/ v_PlayingGu-itar_g01_c01.avi , Количество кадров: 250 Псевдоним: Игра на гитаре	
UCF50/ v_Basket-ball_g04_c01.avi, Количество кадров: 90 Псевдоним: Баскет-бол		UCF50/ v_PlayingGu-itar_g03_c07.avi, Количество кадров: 250 Псевдоним: Игра на гитаре	
UCF50/ v_TennisS-wing_g02_c01.avi, Количество кадров: 60 Псевдоним: Теннис		UCF50/ v_Bench-Press_g02_c01.avi, Количество кадров: 75 Псевдоним: Жим лежа	
UCF50/ v_TennisS-wing_g11_c07.avi, Количество кадров: 90 Псевдоним: Теннис		UCF50/ v_Bench-Press_g04_c03.avi, Количество кадров: 200 Псевдоним: Жим лежа	

На каждое действие собрано не менее четырех различных видеоклипов, предпочтение при отборе отдавалось наиболее отличающимся между собой видеорядам, содержащим одно и то же действие. Таким образом, общее количество использованных видеороликов составило 6681. На тестовых данных присутствовали люди различного пола и возраста, выполнявшие одно и тоже действие с существенной вариативностью, так в видеоролике UCF50/ v_WalkingWithDog_g01_c01.avi, и видео UCF50/ v_WalkingWithDog_g01_c03.avi, людьми выполняется действие «прогулка с собакой» со значительными различиями. Кроме того, в наборе данных объекты интереса имели различный масштаб,

точку и угол обзора, загроможденный фон, различные условия освещения.

Результаты экспериментальных исследований показывают, что предложенная нейронная сеть эффективно справляется с распознаванием действий по видеоданным. Даже такие сложные действия, как жонглирование футбольным мячом, гребля и игра в баскетбол, которые состоят из множества атомарных поддействий, обнаруживаются с высокой точностью.

Заключение

В работе реализовано решение классификации вида человеческой деятельности по видеоданным. Первоначально для распознавания действий применены модели DenseNet121 и MobileNetV2, затем самостоятельно сконструирована и обучена глубокая нейронная сеть. Для проведения экспериментальных исследований построена программная реализация модели в среде IDLE на языке программирования Python. Проверка качества модели выполнена с помощью набора данных UCF50 и полученных с ресурса YouTube видеороликов, содержащих различные действия человека. Экспериментальные исследования показывают высокую эффективность распознавания действий человека по визуальным данным. Самостоятельно разработанная модель нейронной сети незначительно превосходит по качеству работы предварительно обученные сети, однако требует значительно меньше вычислительных ресурсов, что позволяет в дальнейшем применять алгоритм распознавания активности человека в режиме реального времени, например, при контроле за работой сотрудников в офисе или для обнаружения внештатных ситуаций на производстве.

Список литературы

1. Николенко С. Глубокое обучение / С. Николенко, А. Кадурин, Е. Архангельская. СПб.: Питер, 2018. 480 с.
2. Андреас М. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. М.: Альфа-книга, 2017. 487 с.

3. Плас Д. Python для сложных задач. Наука о данных и машинное обучение. Руководство. М.: Питер, 2018. 759 с.
4. Chen Y., Guo M., Wang Z. An improved algorithm for human activity recognition using wearable sensors // 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI). IEEE, 2016. С. 248-252.
5. Dong Y. et al. Dezert-Smarandache theory-based fusion for human activity recognition in body sensor networks // IEEE Transactions on Industrial Informatics. 2020. Т. 16. № 11. С. 7138-7149.
6. Pigou L. et al. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video // International Journal of Computer Vision. 2018. Т. 126. № 2. С. 430-439.
7. Gadzicki K., Khamsehashari R., Zetzsche C. Early vs late fusion in multimodal convolutional neural networks // 2020 IEEE 23rd International Conference on Information Fusion (FUSION). IEEE, 2020. С. 1-6.
8. Ullah A. et al. Action recognition in video sequences using deep bi-directional LSTM with CNN features // IEEE access. 2017. Т. 6. С. 1155-1166.
9. Luo Y. et al. Lstm pose machines //Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. С. 5207-5215.
10. Sargano A. B., Angelov P., Habib Z. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition //Applied sciences. 2017. Т. 7. №. 1. С. 110.
11. UCF50 – Action Recognition Data [Электронный ресурс]. https://www.crcv.ucf.edu/data/UCF50.php (дата обращения: 12.10.2022)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. Densely connected convolutional networks // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition – 2017.
13. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. MobilenetV2: inverted residuals and linear bottlenecks // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition – 2018.
14. IDLE // Python 3.11.0 Documentation [Электронный ресурс]. https://docs.python.org/3/library/idle.html (дата обращения: 30.10.2022).

15. YouTube [Электронный ресурс]. https://www.youtube.com/ (дата обращения: 30.10.2022)

References

1. Nikolenko S., Kadurin A., Arkhangel'skaya E. *Glubokoe obuchenie* [Deep learning]. SPb.: Piter, 2018, 480 p.
2. Andreas M. *Vvedenie v mashinnoe obuchenie s pomoshch'yu Python. Rukovodstvo dlya spetsialistov po rabote s dannymi* [Introduction to Machine Learning with Python. A guide for data scientists]. M.: Al'fa-kniga, 2017, 487 p.
3. Plas D. *Python dlya slozhnykh zadach. Nauka o dannykh i mashinnoe obuchenie. Rukovodstvo* [Python for complex tasks. Data Science and Machine Learning. Guide]. M.: Piter, 2018, 759 p.
4. Chen Y., Guo M., Wang Z. An improved algorithm for human activity recognition using wearable sensors. *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, 2016, pp. 248-252.
5. Dong Y. et al. Dezert-Smarandache theory-based fusion for human activity recognition in body sensor networks. *IEEE Transactions on Industrial Informatics,* 2020, vol. 16, no. 11, pp. 7138-7149.
6. Pigou L. et al. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision,* 2018, vol. 126, no. 2, pp. 430-439.
7. Gadzicki K., Khamsehashari R., Zetzsche C. Early vs late fusion in multimodal convolutional neural networks. *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, 2020, pp. 1-6.
8. Ullah A. et al. Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE access,* 2017, vol. 6, pp. 1155-1166.
9. Luo Y. et al. Lstm pose machines. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2018, pp. 5207-5215.
10. Sargano A. B., Angelov P., Habib Z. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Applied sciences,* 2017, vol. 7, no. 1, p. 110.
11. UCF50 – Action Recognition Data. https://www.crcv.ucf.edu/data/UCF50.php

12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. Densely connected convolutional networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition – 2017.*
13. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. MobileNetV2: inverted residuals and linear bottlenecks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition – 2018.*
14. IDLE // Python 3.11.0 Documentation. https://docs.python.org/3/library/idle.html
15. YouTube. https://www.youtube.com/

ДАННЫЕ ОБ АВТОРАХ

Пятаева Анна Владимировна, кандидат технических наук, доцент кафедры Систем искусственного интеллекта, Институт космических и информационных технологий
СФУ
ул. Академика Киренского, 26Б, г. Красноярск, 660074, Российская Федерация anna4u@list.ru

Мерко Михаил Алексеевич, доцент кафедры Систем искусственного интеллекта, кандидат технических наук, Институт космических и информационных технологий
СФУ
ул. Академика Киренского, 26Б, г. Красноярск, 660074, Российская Федерация mmerko@sfu-kras.ru

Жуковская Владислава Андреевна, студентка 1 курса магистратуры, Институт космических и информационных технологий
СФУ
ул. Академика Киренского, 26Б, г. Красноярск, 660074, Российская Федерация
zhukovskaya.vlada00@mail.ru

Казакевич Алена Александровна, студентка 2 курса магистратуры, Институт космических и информационных технологий

СФУ
ул. Академика Киренского, 26Б, г. Красноярск, 660074, Российская Федерация

DATA ABOUT THE AUTHORS

Anna V. Pyataeva, Associate Professor of the Department of Artificial Intelligence Systems, Candidate of Technical Sciences
Siberian Federal University
26B, Academician Kirensky, Krasnoyarsk, 660074, Russian Federation
anna4u@list.ru
SPIN-code: 2498-2148
ORCID: https://orcid.org/0000-0002-0140-263X

Mikhail A. Merko, Associate Professor of the Department of Artificial Intelligence Systems, Candidate of Technical Sciences
Siberian Federal University
26B, Academician Kirensky, Krasnoyarsk, 660074, Russian Federation
mmerko@sfu-kras.ru
SPIN-code: 2305-6520

Vladislava A. Zhukovskaya, 1st year master's student
Siberian Federal University
26B, Academician Kirensky, Krasnoyarsk, 660074, Russian Federation
zhukovskaya.vlada00@mail.ru
ORCID: https://orcid.org/0000-0002-6113-3128

Alena A. Kazakevich, 2nd year master's student
Siberian Federal University
26B, Academician Kirensky, Krasnoyarsk, 660074, Russian Federation

Поступила 08.11.2022
После рецензирования 15.11.2022
Принята 21.11.2022

Received 08.11.2022
Revised 15.11.2022
Accepted 21.11.2022