УДК 004.931

АНАЛИЗ МЕТОДОВ КЛАССИФИКАЦИИ ДЕЙСТВИЙ ЧЕЛОВЕКА НА ВИДЕОИЗОБРАЖЕНИИ

Статья поступила в редакцию 21.01.2021, в окончательном варианте – 20.02.2021.

Марьенков Александр Николаевич, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а,
кандидат технических наук, доцент, ORCID: 0000-0003-1378-3553, e-mail: marenkovan17@gmail.com
Приходько Александр Александрович, Астраханский государственный университет, 414052, Российская Федерация, г. Астрахань, ул. Камышинская, 2,
магистрант, e-mail: alexsandr_498@mail.ru

В работе обоснованы актуальность и практическая значимость разработки новых методов анализа видеоизображения с целью классификации действий человека для дальнейшего выявления потенциально опасных инцидентов на объекте информатизации. Рассмотрены классификаторы на основе модели нейронной сети 3D ResNet, а также подходы, использующие векторную модель тела с применением библиотеки OpenPose. Первый эксперимент проведен с использованием модели нейронной сети 3D ResNet. Для обучения был использован датасет от Kinetic, включающий порядка 400 действий, среди которых присутствовали движения из единоборств. В тестовом наборе были использованы примеры из хоккейных драк и боевых приемов из фильмов. Следующий эксперимент заключался в классификации действия на базе анализа векторной модели тела человека. Kinect предоставляет данные о движении в виде иерархии основных узлов скелета человека, где вращение одних суставов относительно других представлено в виде кватернионов. Итоговое обучение модели происходило с применением датасета RGBU-D с 432 аннотированными действиями. В заключительном эксперименте для представления формализованного движения был выбран формат BVH. Переобучение модели проводилось на RGBU-D датасете, в связи с чем описание всех кадров пришлось изменить с 20 ключевых точек стандарта OpenPose до 17 из стандарта BVH, которые использовались в последующей работе с моделью. За основу конечного модуля по классификации действий, имеющихся на экране, была взята структура нейронной сети с LSTM-слоем с изменением входных данных – вместо набора фреймов из видео стал подаваться набор векторов тел людей в кадре. Обучение данной нейронной сети было проведено с использованием датасета в 2000 видеофайлов (1000 опасных ситуаций [в основном драки] и 1000 обычных действий в жизнедеятельности человека, не представляющие угрозы). Были проанализированы полученные результаты, сделаны выводы о применимости рассмотренных подходов для задачи распознавания действия человека на видеоизображении.

Ключевые слова: распознавание, глубокое обучение, нейронные сети, распознавание и классификация действий человека, выявление инцидентов, анализ видеоизображения

ANALYSIS OF METHODS FOR CLASSIFYING HUMAN ACTIONS ON A VIDEO IMAGE

The article was received by the editorial board on 21.01.2021, in the final version – 20.02.2021.

Marienkov Alexander N., Astrakhan State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation,
Cand. Sci. (Engineering), ORCID: 0000-0003-1378-3553, e-mail: marenkovan17@gmail.com
Prikhodko Alexander A., Astrakhan State University, 2 Kamyshinskaya St., 414052, Astrakhan, Russian Federation,
undergraduate student, e-mail: alexsandr_498@mail.ru

The work justifies the relevance and practical significance of developing new methods for analyzing video images with the aim of classifying human actions for further identification of potentially dangerous incidents at the informatization facility. Classifiers based on model of neural network 3D ResNet, as well as approaches using vector model of body with application of library OpenPose are considered. The first experiment is made with use of model of neural network 3D ResNet. Dataset from Kinetic was used for training. That dataset is including about 400 actions, among which there were movements from martial arts. Examples from hockey fights and combat techniques from films were used in the testing set. The next experiment was to classify the action based on an analysis of a vector model of a human body. Kinect provides motion data in the form of a hierarchy of the main nodes of the human skeleton, where the rotation of some joints relative to others is represented in the form of quaternions. The final training of the model occurred using the RGBU-D dataset with 432 annotated actions. The BVH format was chosen to represent the formalized movement in the final experiment. Model retraining was carried out on the RGBU-D dataset, and therefore the description of all frames had to be changed from 20 key points of the standard OpenPose to 17 from the BVH standard, which were used in subsequent work with the model. The structure of the neural network with the LSTM layer with a change in input data was taken as the basis of the final module for classifying the actions available on the screen – instead of a set of frames, a set of vectors of people's bodies in the frame began to be supplied

from the video. Training of this neural network was carried out using a dataset in 2000 video files (1000 dangerous situations [mainly fights] and 1000 ordinary actions in human life that are not a threat). The results were analyzed as well as conclusions were made about the applicability of the approaches considered for the task of recognizing the action of a person on a video image.

Keywords: recognition, deep learning, neural networks, recognition and classification of human actions, incident detection, video image analysis

Graphical annotation (Графическая аннотация)

Введение. Системы видеонаблюдения являются одним из основных средств обеспечения безопасности на объекте информатизации. Видеонаблюдение позволяет организовать круглосуточный контроль за объектом с записью видеоархива для последующего анализа произошедших инцидентов. Однако конечный анализ происходящего на контролируемой территории проводит сотрудник службы безопасности, непосредственно наблюдающий за происходящим на экране монитора. В условиях большого количества видеокамер на сотрудника службы безопасности ложится трудновыполнимая задача анализа огромного потока видеоданных, непрерывно поступающих с объекта наблюдения [1].

В связи с этим становится актуальной задача автоматизации процесса анализа видеопотока с целью выявления инцидентов, возникающих на контролируемом объекте. В настоящее время уже существует множество разработок, направленных на предварительный анализ происходящего на видеоизображении: забытые предметы, оружие, проход людей в запретную зону (например, выход на железнодорожные пути), драки и т.п.

Несмотря на то, что проблемой анализа видеоряда занимается большое количество исследователей, данное направление все еще остается перспективным для совершенствования методик и разработки новых функциональных возможностей по анализу происходящего на экране монитора системы видеонаблюдения [8]. В рамках этой задачи основной проблемой является распознавание движений человека и их классификация. При этом стоит отметить, что на качество распознавания и классификацию движений влияют факторы, изменяющие походку человека (одежда, скрывающая человека; переносимые предметы: сумки, рюкзаки; неудобная обувь) или входные параметры изображения (ракурс, освещение, разрешение камеры, расстояние от человека до камеры) [3].

Цель данной работы – изучить и сравнить эффективность методов распознавания движений человека с применением различных нейросетевых технологий. Были рассмотрены классификаторы на основе модели нейронной сети 3D ResNet, а также подходы, использующие векторную модель тела с применением библиотеки OpenPose.

Проведение экспериментальных исследований, часть 1. Первый эксперимент был проведен с использованием модели нейронной сети 3D ResNet. Идентификация совершаемого действия человека в кадре основывается на анализе всего кадра [9].

Для обучения модели был использован датасет от Kinetic с большим количеством действий. Действия в датасете уже размечены и включают в себя порядка 400 наименований, среди которых также есть действия из единоборств. Наличие примеров с единоборствами стало причиной отбора данного датасета для обучения модели (так как является уникальным фактором относительно других подобных публичных датасетов для распознаваний действий) на поиск потенциально опасных инцидентов на контролируемом объекте. Для тестирования обученной модели была использована выборка, включающая хоккейные драки и боевые приёмы из фильмов.

После окончания обучения модели на 10 эпохе и проверки кроссвалидации на выборке из «живых ситуаций», были получены неудовлетворительные результаты, а именно, было выявлено, что данная модель не может верным способом стабильно определять предполагаемые опасные действия человека на видеоизображении.

Итоговые результаты по первичному эксперименту с определением действий человека представлены в таблице ниже (табл. 1), а динамика во времени представлена на изображении ниже (рис. 1).

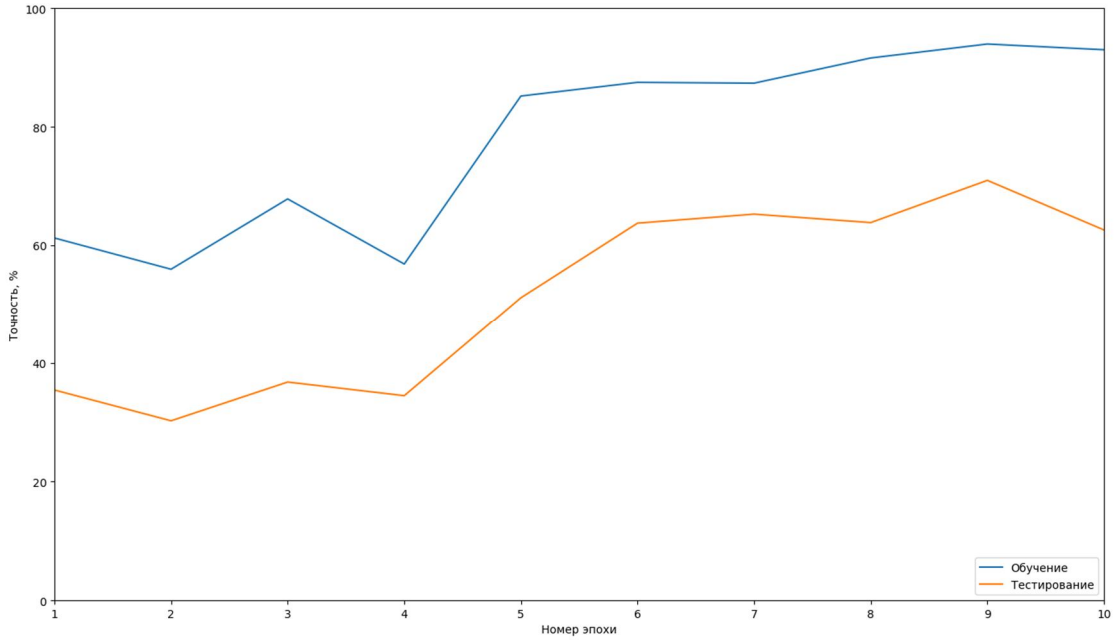

Рисунок 1 – Изменение точности во время обучения

Таблица 1 – Результаты обучения модели на основе нейронной сети 3D ResNet

Количество эпох	Фаза обучения		Фаза валидации	
	Ошибки	Точность	Ошибки	Точность
10	0,2920	0,9059	0,1703	0,5633

Проведенный анализ эксперимента показал, что причиной некорректного выявления действия на конечной выборке являлся тот фактор, что в кадре присутствовало несколько людей, выполняющих разные действия, а также для первой модели важную роль играло окружение, что в случае решения задачи не является опорным пунктом выполнения предсказания об обстановке на объекте информатизации, так как обстановка там по большей части будет неизменной.

Проведение экспериментальных исследований, часть 2. Второй эксперимент заключался в классификации действия на базе анализа векторной модели тела человека.

Векторная модель тела человека есть формализованное представление движения человека, где в виде векторов представлены кости человеческого скелета, а углам между ними соответствуют углы поворота основных узлов человеческого тела друг относительно друга.

Kinect предоставляет данные о движении в виде иерархии основных узлов скелета человека, где вращение одних суставов относительно других представлено в виде кватернионов (роль вращающихся векторов выполняют кости скелета), а смещение представлено в виде трехмерных векторов в локальной для каждого узла системе координат.

С целью улучшения результатов и ухода от прошлых ошибок последующий эксперимент проходил в направлении получения векторной модели тела человека и его анализа для классификации действия по вектору. Пример практической реализации данного подхода представлен на рисунке ниже (рис. 2) [8].

Рисунок 2 – Работа библиотеки OpenPose

Данный подход основан на предположении, что данные о позе человека, извлеченные из видеороликов, содержат достаточную информацию для обучения классификатора, способного распознавать отдельные действия и получать представления для оценки всего состояния в целом [4].

Таким образом, сперва необходимо получить вектор, описывающий текущее движение человека на видеоизображении. Дальнейшим шагом является сопоставление векторов по методу ближайшего соседа для выявления дескриптора движения и классификации действия объекта. Сама классификация действия происходит за счёт использования метода ближайшего соседа на эталонных векторах положения суставов для конкретных действий и вектора, полученного с изображения. При нахождении наименьшего сдвига с эталоном выбирается итоговое действие в классификаторе, пример эталонных векторов для бега представлен на рисунке ниже (рис. 3) [10].

Рисунок 3 – Временно-пространственное представление вектора тела

Итоговое обучение модели происходило с применением датасета RGBU-D с 432 аннотированными действиями. Процесс обучения проходил вплоть до 10 эпох с итоговой точностью 47,32 %, динамика которого представлена на рисунке ниже (рис. 4), а результаты представлены в таблице ниже (табл. 2).

Полученный результат показал, что переход к пространственно-временной модели классификации действий благоприятно повлиял на накопительную часть в обучении, но, к сожалению, не дал ожидаемых результатов, как и описанный ранее эксперимент.

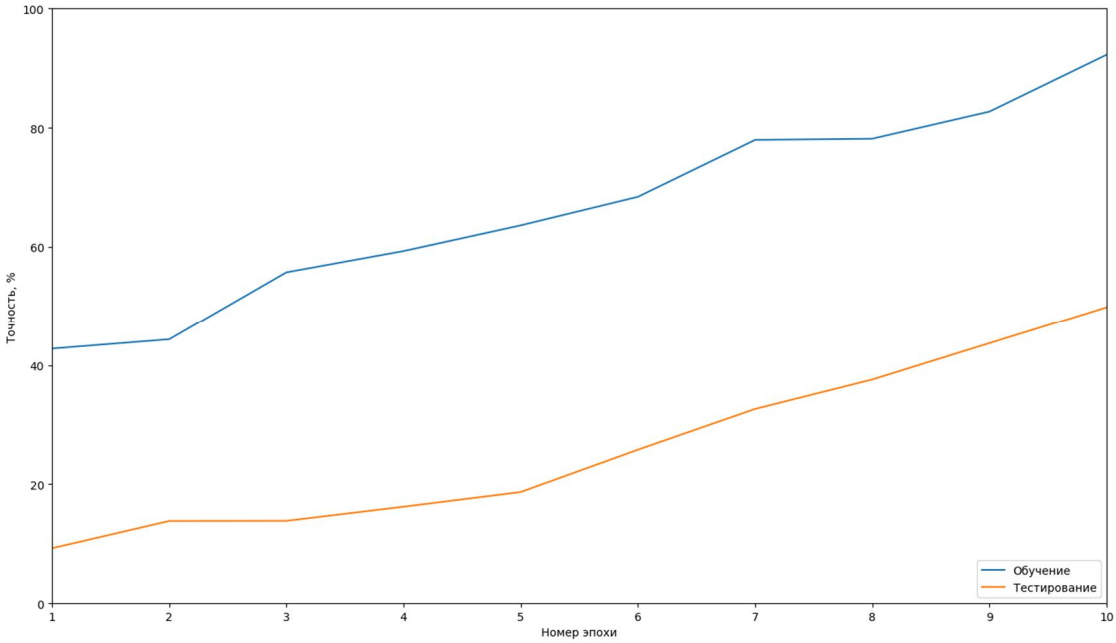

Рисунок 4 – Изменение точности во время обучения

Таблица 2 – Результаты по ошибкам первого и второго рода на обучающей выборке

	Верная гипотеза	
	H_0 Действие есть	H_1 Действия нет
Количество принятых решений	Действие распознано верно / 50 раз	Распознано несуществующее действие / 7 раз
	Действие не распознано / 0 раз	Не найдено несуществующих действий / 43 раза

Проведение экспериментальных исследований, часть 3. В третьем эксперименте для представления формализованного движения был выбран формат BVH, как наиболее распространенный и наиболее полно описывающий структуру человеческого тела. BVH обозначает данные Bio Vision Hierarchical. Этот формат предоставляет возможность представления информации об иерархии каркаса тела человека в добавление к данным о движении. Каждый элемент скелета, визуализация которого представлена на рисунке ниже (рис. 5), содержит в себе информацию о смещении и вращении относительно родительского элемента. Вращение представляется в углах Эйлера.

Переобучение модели проводилось на RGBU-D датасете [5], в связи с чем описание всех кадров пришлось изменить с 20 ключевых точек стандарта OpenPose до 17 из стандарта BVH, которые использовались в последующей работе с моделью.

За основу конечного модуля по классификации действий, имеющихся на экране, была взята структура нейронной сети с LSTM-слоем с изменением входных данных – вместо набора фреймов из видео стал подаваться набор векторов тел людей в кадре.

Рисунок 5 – Построение векторной модели тела человека

Подобный подход призван ускорить скорость обработки, а следовательно, и обучения в целом, а также уменьшить требуемое потребление ресурсов во время обучения и использования модели в программном продукте конечным пользователем.

Обучение данной нейронной сети был проведено с использованием датасета в 2000 видеофайлов (1000 опасных ситуаций [в основном драки] и 1000 обычных действий в жизнедеятельности человека, не представляющие угрозы). Отличием данного датасета от предыдущих является включение данных, полученных в рамках исследования Kinect-датчиков, совместно с видеофрагментами драк из фильмов и спортивных матчей, также постановочных драк и видео с публичных камер наружного видеонаблюдения [7].

Для поиска человека в кадре была использована предобученная модель YOLOv3 на датасете COCO, которая уже способна выявлять человека на кадре «из коробки» [2]. Поиск человека в кадре обоснован уменьшением математических вычислений для случаев, когда людей в кадре нет.

На последующем шаге понадобилось обучить модель LightTrack [6], визуализация обучения которой представлена на рисунке ниже (рис. 6), для плавного связывания объектов при переходе между кадрами. Бонусным эффектом было получено присуждение уже имеющегося номера объектам, которые возвращались в кадр из того же положения, с которого они пропали.

Обучение модели LightTrack было закончено на 23 эпохе. Суммарное время обучения заняло 4 часа. Конечная точность модели остановилась на 96,32 %.

Затем потребовалось переобучить модель по построению векторной модели структуры человека, что заняло ещё 22 часа на 15 эпох, но для дальнейшей работы использовалась резервная копия от результата 14 эпохи как наиболее стабильной.

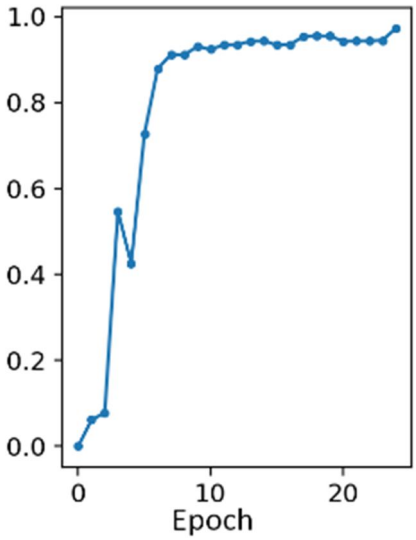

Рисунок 6 – Изменение точности во время обучения

Заключительный этап заключался в обучении модели для классификации векторных моделей людей с целью определения выявления инцидента (например, драки) на видеоизображении. Обучение данной модели заняло 18 часов, визуализация которого представлена на рисунке ниже (рис. 7), со стабильным потреблением 4 Гб оперативной памяти и 2 Гб видеопамяти на всё время обучения на площадке Google Colaboratory. Дальнейшее обучение с целью повышения точности возможно, но является трудоёмким процессом по причине имеющихся ограничений со стороны площадки.

В конечном итоге точность выявления инцидентов на видеоизображении составила 77,56 % на 15 эпохе обучения.

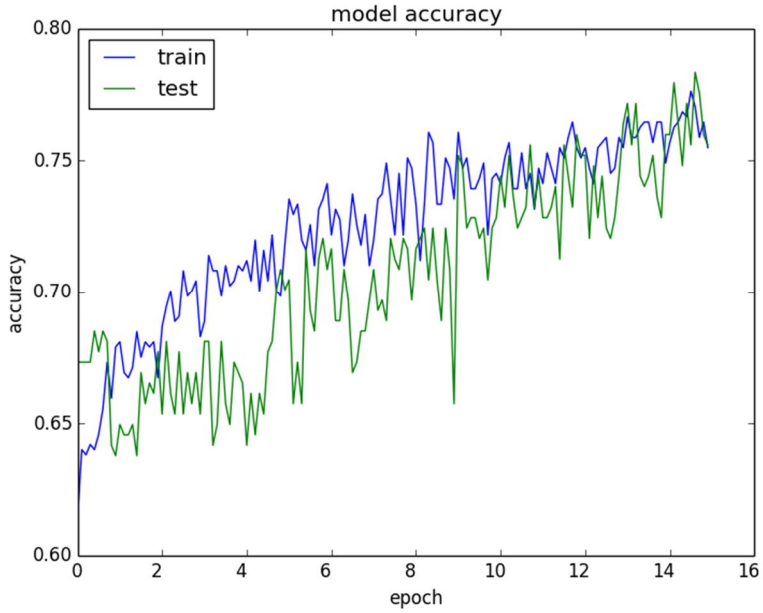

Рисунок 7 – Изменение точности модели во время обучения

Заключение. В работе был проведен обзор методов распознавания движений человека на видеоизображении. Представлены результаты экспериментов, демонстрирующих возможности данных методов:

- распознавание действий с помощью сети 3D ResNet (итоговая точность 56,33 %);
- распознавание действий человека с помощью анализа векторной модели человека (итоговая точность 47,32 %);
- распознавание действий нейронной сетью с LSTM-слоем с изменением входных данных – вместо набора фреймов из видео стал подаваться набор векторов тел людей в кадре (итоговая точность 77,56 %).

В результате проведенных экспериментов были выявлены возможности представленных в работе подходов по распознаванию и классификации действий человека в кадре. На основании полученных результатов был сделан выбор в пользу нейронной сети с LSTM-слоем, получающей на вход вектора людей для распознавания действий.

Дополнительно стоит отметить, что во всех экспериментах использовались различные датасеты, что также повлияло на итоговую точность. Однако полученные результаты позволяют интерпретировать общую эффективность методов и позволяют сделать выводы о возможностях рассмотренных методов распознавания и классификации движений человека при анализе видеоизображения.

Библиографический список

1. Дуленко В. А. Анализ подходов к обеспечению безопасности на городских территориальных объектах в рамках реализации Концепции «Безопасный город» / В. А. Дуленко, В. А. Пестриков // Вестник ВЭГУ. – 2011. – № 4 (54). – С. 22–27.

2. Белясников С. А. Методы обнаружения движущихся объектов в видеопотоке / С. А. Белясников, Р. С. Дорофеев // Молодежный вестник ИрГТУ. – 2016. – № 2. – С. 4.

3. Бокова О. И. К вопросу о внедрении механизмов интеллектуального анализа в информационную среду АПК «Безопасный город» / О. И. Бокова, В. С. Дунин, Н. С. Хохлов // Моделирование, оптимизация и информационные технологии. – 2015. – № 4. – С. 18.

4. Буйко А. Ю. Выявление действий на видео с помощью рекуррентных нейронных сетей / А. Ю. Буйко, А. Н. Виноградов // Программные системы: теория и приложения. – 2017. – Т. 8, № 4 (35). – С. 327–345.

5. RGBU-D Датасет. – Режим доступа: https://rgbd-dataset.cs.washington.edu/, свободный. – Заглавие с экрана. – Яз. англ.

6. CMU-Perceptual-Computing-Lab. – Режим доступа: https://github.com/CMU-Perceptual-Computing-Lab/openpose, свободный. – Заглавие с экрана. – Яз. англ.

7. Li Y. Online human action detection using joint classification-regression recurrent neural networks / Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, J. Liu. – 2016. – Режим доступа: https://arxiv.org/abs/1604.05633v2, свободный. – Заглавие с экрана. – Яз. англ.

8. Francois Chollet. Deep Learning with Python / Francois Chollet. – MANNING Shelter Island, 2018. – С. 361.

9. Ормонейт Д. Циклическое обучение и отслеживание человеческого движения / Д. Ормонейт, Х. Сиденбладш, М. Блэк, Т. Хасти // Достижения в области обработки информации нейронными системами. – 2011. – № 13. – С. 894–900.

10. Тройже Н. Декомпозиция биологического движения: Фреймворк анализа и синтеза человеческой походки / Н. Тройже // Журнал зрения, нейронауки и психологии систем визуализации. – 2002. – № 5. – С. 371–387.

References

1. Dulenko V. A., Pestrikov V. A. Analiz podkhodov k obespecheniyu bezopasnosti na gorodskikh territorialnykh obektakh v ramkakh realizatsii Koncepcii «Bezopasnyy gorod» [Analysis of approaches to ensuring security at urban territorial sites in the framework of the implementation of the Safe City Concept]. *Vestnik VEGU* [VEGU Bulletin], 2011, no. 4 (54), pp. 22–27.

2. Belyasnikov S. A., Dorofeev R. S. Metody obnaruzheniya dvizhushchikhsya obektov v videopotoke [Methods of detecting moving objects in a video stream]. *Molodezhnyy vestnik IrGTU* [ISTU Bulletin of Youth], 2016, vol. 2, p. 4.

3. Bokova O. I., Dunin V. S., Khokhlov N. S. K voprosu o vnedrenii mekhanizmov intellectual nogo analiza v informatsionnuyu sredu APK "Bezopasnyy gorod" [On the introduction of mechanisms of intellectual analysis in the information environment of the agro-industrial complex "Safe City"]. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii* [Modeling, optimization and information technologies], 2015, vol. 4, p. 18.

4. Buyko A. Yu., Vinogradov A. N. Vyyavlenie deystviy na video s pomoshchyu rekurrentnykh neyronnykh setey [Identification of actions on video using recurrent neural networks]. *Programmnye sistemy: teoriya i prilozheniya* [Software systems: theory and applications], 2017, vol. 8, no. 4, pp. 327–345.

5. *RGBU-D Dataset*. Available at: https://rgbd-dataset.cs.washington.edu/

6. *CMU-Perceptual-Computing-Lab*. Available at: https://github.com/CMU-Perceptual-Computing-Lab/openpose.

7. Li Y., Lan C., Xing J., Zeng W., Yuan C., Liu J. *Online human action detection using joint classification-regression recurrent neural networks*, 2016. Available at: https://arxiv.org/abs/1604.05633v2.

8. Francois Chollet. *Deep Learning with Python*. MANNING Shelter Island, 2018, p. 361.

9. Ormoneit D., Sidenbladh H., black M., Hastie, T. Vyyavlenie deystviy na video s pomoshchyu rekurrentnykh neyronnykh setey [Cyclic training and tracking human motion]. *Programmnye sistemy: teoriya i prilozheniya* [Advances in information processing by neural systems], 2011, vol. 13, pp. 894–900.

10. Troige N. Dekompozitsiya biologicheskogo dvizheniya: Freymvork analiza i sinteza chelovecheskoy pokhodki [Decomposition of biological movement: A framework for analyzing and synthesizing human gait]. *Zhurnal zreniya, neyronauki i psikhologii sistem vizualizatsii* [Journal of Vision, Neuroscience and Psychology of Visualization Systems], 2002, vol. 5, pp. 371–387.