Оригинальная статья / Original article

https://doi.org/10.21869/2223-1560-2021-25-4-177-200

CC BY 4.0

Метод отслеживания процессов взаимодействия пользователей с объектами на видеопоследовательностях

Р. Н. Яковлев [1] ✉

[1] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербургский институт информатики и автоматизации Российской академии наук
14-я линия В.О., д. 39, г. Санкт-Петербург 199178, Российская Федерация

✉ e-mail: iakovlev.r@mail.ru

Резюме

Цель исследования. *При разработке киберфизических систем и интеллектуальных пространств, предназначенных для анализа пользовательской активности, актуальной является задача отслеживания взаимодействий пользователей с объектами.*

Методы. *В данной работе для решения задачи детектирования и отслеживания взаимодействий пользователей с объектами на видеопоследовательностях был разработан соответствующий метод, основанный на комбинированном использовании нейросетевых моделей детектирования объектов и сегментации пользователей, а также построения карт глубины по кадрам видеоряда. В исследовании представлены соответствующие алгоритмы и алгоритмические модели. Апробация и оценка качества функционирования разработанного метода производилась на основе тестового набора данных, включающего в себя 1000 видеопоследовательностей длительностью до 20 секунд.*

Результаты. *В ходе экспериментальных исследований были определены показатели точности (accuracy, recall, precision) детектирования взаимодействий для видеопоследовательностей с уровнем освещенности 100% и 50%, которые составили {0.82, 0.78, 0.76} и {0.70, 0.59, 0.70} соответственно, при этом усредненные доли корректно отслеженных взаимодействий для данных наборов видеопоследовательностей имели значения 81% и 71%. Согласно результатам проведенного тестирования, разработанный метод предоставляет возможность осуществлять детектирование и отслеживание взаимодействий пользователей с объектами в режиме реального времени, в том числе в условиях неполной освещенности сцены.*

Заключение. *По результатам апробации предложенного метода отслеживания взаимодействий пользователей с объектами на тестовом наборе из 1000 видеопоследовательностей, предложенное решение пока-зало довольно высокое качество детектирования и отслеживания взаимодействий для видеопоследовательностей с уровнями освещенности 100% и 50%. Таким образом, разработанный метод в определенной мере является устойчивым к изменению уровня освещенности сцены и обеспечивает успешное решение задачи детектирования и отслеживания взаимодействия пользователей с различными классами объектов по видеопоследовательности, не требуя при этом применения специализированного оборудования.*

Ключевые слова: *киберфизические системы; взаимодействие пользователя с объектом; трекинг взаимодействий; Mask R-CNN; Spine-Net; RetinaNet; FCRN-DepthPrediction; Rolo.*

Конфликт интересов: *Автор декларирует отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.*

Для цитирования*:* Яковлев Р. Н. Метод отслеживания процессов взаимодействия пользователей с объектами на видеопоследовательностях // Известия Юго-Западного государственного университета. 2021; 25(4): 177-200. https://doi.org/10.21869/2223-1560-2021-25-4-177-200.

Поступила в редакцию 14.10.2021 *Подписана в печать 02.11.2021* *Опубликована 20.12.2021*

© Яковлев Р. Н., 2021

A Method for Tracking the Processes of User Interaction with Objects in Video Sequences

Roman N. Iakovlev [1] ✉

[1] 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,
39, 14th Line, St. Petersburg 199178, Russian Federation

✉ e-mail: iakovlev.r@mail.ru

Abstract

Purpose of research. *When developing cyber-physical systems and intelligent environment designed to analyze user activity, the task of tracking user interactions with objects is topical.*

Methods. *In this paper, to solve the problem of detecting and tracking user interactions with objects in video sequences, an appropriate method based on the combined application of neural network models for detecting objects and segmenting users, as well as constructing depth maps based on video sequence frames was developed. The study presents corresponding algorithms and algorithmic models. The testing and assessment of the quality of functioning of the developed method were carried out on the basis of a test data set including 1,000 video sequences with a duration of up to 20 seconds.*

Results. *In the course of experimental studies, the indicators of accuracy (accuracy, recall, precision) of detecting interactions for video sequences with illumination levels of 100% and 50% were determined, which amounted to {0.82, 0.78, 0.76} and {0.70, 0.59, 0.70}, respectively, while the average proportions of correctly tracked interactions for these sets of video sequences had values of 81% and 71%. According to the results of the conducted testing, the developed method provides an opportunity to detect and track user interactions with objects in real time, including those in conditions of the lack of illumination of the scene.*

Conclusion. *Based on the results of the testing of the proposed method of tracking user interactions with objects on a test set of 1,000 video sequences, the proposed solution showed a fairly high quality of detecting and tracking interactions in video sequences with illumination levels of 100% and 50%. Thus, the developed method is to a certain extent resistant to changes in the level of illumination of the scene and provides a successful solution to the problem of detecting and tracking user interaction with various classes of objects in the video sequence, without requiring the use of specialized equipment.*

Keywords: *cyber physical systems; user interaction with an object; tracking interactions; Mask R-CNN; Spine-Net; RetinaNet; FCRN-DepthPrediction; Rolo.*

Conflict of interest. *The author declare the absence of obvious and potential conflicts of interest related to the publication of this article.*

For citation: Iakovlev R. N. A Method for Tracking the Processes of User Interaction with Objects in Video Sequences. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta = Proceedings of the Southwest State University*. 2021; 25(4): 177-200 (In Russ.). https://doi.org/10.21869/2223-1560-2021-25-4-177-200.

Received 14.10.2021 *Accepted 02.11.2021* *Published 20.12.2021*

***

Введение

Современные технологии и инструменты компьютерного зрения обеспечивают решение широкого спектра задач в данной области. Однако на сегодняшний день остается ряд проблем,

решение которых с использованием известных инструментов, оказывается достаточно затруднительным или даже полностью невозможным. К числу подобных задач относится задача отслеживания процессов взаимодействия пользователей с объектами. Программные инструменты решения данной задачи имеют широкий потенциал для использования в производственной сфере, в частности, контроля выполнения регламента работ и обнаружения ошибок сотрудников при работе с изделиями или инструментами. Большинство существующих подходов к решению рассматриваемой задачи предназначены для детектирования взаимодействия с небольшим количеством объектов, требуют использования специализированного оборудования, могут применяться в ограниченном наборе рабочих сред или не позволяют осуществлять трекинг соответствующих взаимодействий по видеопоследовательности, что значительно ограничивает область применения данных решений. Настоящее исследование посвящено разработке метода отслеживания процессов взаимодействия пользователей с объектами, обеспечивающего как детектирование, так и последующий трекинг соответствующего взаимодействия. Предложенный метод должен быть пригодным для использования в различных рабочих условиях с сохранением требуемого уровня точности детектирования и трекинга.

Материалы и методы

Разработка метода отслеживания процессов взаимодействия пользователей с объектами (human-object interaction) по видеопоследовательностям предполагает решение следующей группы задач: детектирование на кадрах видеопоследовательности пар вида «пользователь – объект взаимодействия», идентификация факта начала взаимодействия, трекинг соответствующих пар до завершения процесса взаимодействия.

В настоящее время существует большое количество подходов к решению задачи детектирования объектов на изображениях и видеопоследовательностях [1, 2, 3]. В табл. 1 представлена сравнительная характеристика качества решения задачи детектирования объектов на изображениях на примере известного набора данных COCO [4]. Оценки качества приведены в отношении ряда современных нейросетевых моделей, направленных на решение соответствующей задачи.

Согласно приведенным выше результатам, одним из наиболее перспективных решений в данной области, является предобученная нейронная сеть (NN) RetinaNet с экстрактором признаков SpineNet-190 [5].

Такая модель способна не только осуществлять детектирование объектов на изображении, но и определять классовую принадлежность детектируемых объектов. К типам объектов, которые данная сеть способная идентифицировать, относятся различные категории офисной мебели, иные предметы, а также непосредственно люди.

Таблица 1. Сравнительная характеристика качества решения задачи детектирования объектов на примере набора данных COCO

Table 1. Comparative characteristics of the quality of solving the problem of detecting objects using the example of the COCO dataset

Нейросетевая модель / Neural network model	AP	AP_{50}
RetinaNet (SpineNet-190)	52.1	71.8
FAIR Mask R-CNN	50.3	72.0
RetinaNet (ResNet-101-FPN)	39.1	59.1
Mask R-CNN (ResNet-101-FPN)	38.2	60.3
FAIRCNN	33.2	52.0

Таким образом, данная нейронная сеть может быть использована в качестве основы для решения задачи по идентификации пользователей и объектов взаимодействия. Однако формирование корректных пар «пользователь-объект взаимодействия» требует применения дополнительных алгоритмических моделей к результатам работы данной нейронной сети.

В контексте задачи детектирования начала взаимодействия пользователя с объектом существует большое количество методов, направленных на идентификацию типов взаимодействия. Авторами [6] предложена модель InteractNet, определяющая на фотографиях сложных повседневных сцен тип взаимодействия человека с объектами. Модель обучена на сложном наборе данных V-COCO и решает задачу распознавания триплетов (человек, действие, объект), локализуя контуры человека, связанного объекта взаимодействия и идентифицируя выполняемое действие. Аналогичный проект представили авторы статьи [7]. Они предложили модель HOI (Human-Object Interaction) для обнаружения взаимодействия человека с объектом, которая в отличие от сложных сквозных моделей использует внешние признаки от предварительно обученных детекторов объектов и кодирует макет с помощью созданных вручную ограничительных элементов координат (необязательно ключевых точек позы человека). Для определения на изображении 18 ключевых точек каждого человека авторы используют модель Open Pose [8], которая определяет ключевые точки скелета человека на изображении или в видеопотоке. Рассмотренные выше методы [6, 7] обладают достаточно высокой точностью определения вида взаимодействия человека с объектом на основе внешних показателей, но для их успешного применения необходимо, чтобы пользователь и объект в кадре были видны полностью. Недостатком обоих методов также является их ориентация на работу с одиночными изображениями без возможности работы с видеопотоком. Кроме того, в рамках данных методов предполагается, что на

входных данных взаимодействие между пользователем и объектом уже осуществляется, таким образом, данные решения не могут быть использованы для решения задачи детектирования взаимодействия между пользователем и объектом.

Альтернативный подход к идентификации факта начала взаимодействия пользователя с объектом может быть основан на идентификации соприкосновения пользователя с объектом. Для выявления данных ситуаций необходимо осуществлять оценку взаимного расположения объекта и рук пользователя в пространстве, при этом пересечение сегментов рук и объекта на изображении является необходимым, но недостаточным условием совпадения их пространственных положений в связи с отсутствием информации об удаленности объекта от пользователя по оси перпендикулярной плоскости изображения. Данная информация может быть получена, в частности, на основе анализа результатов работы различных методов построения карт глубины на соответствующем изображении. Таким образом, реализация надежной идентификации факта начала взаимодействия пользователя с объектом требует комбинированного решения задач по детектированию рук пользователя, детектированию объекта, а также анализа результатов работы выбранного метода построения карт глубины на соответствующем изображении для установления факта соприкосновения пользователя с объектом.

Результаты оценки качества в отношении ряда современных решений, полученные на примере набора данных COCO, представлены в табл. 2.

Таблица 2. Сравнительная характеристика качества решения задачи сегментации объектов на примере набора данных COCO

Table 2. Comparative characteristics of the quality of solving the object segmentation problem using the example of the COCO dataset

Модель NN / NN Model	AP	AP_{50}
Mask R-CNN (SpineNet-190)	46.3	70.9
HRNet	41.0	63.4
Mask R-CNN (ResNeXt-101-FPN)	37.1	60.0
RetinaMask (ResNet-101-FPN)	36.4	57.3
Mask R-CNN (ResNet-101-FPN)	35.7	58.0
Mask R-CNN (ResNet-50-FPN)	35.6	57.6

Большинство решений в области детектирования рук пользователя, сконцентрированы на проблеме определения той или иной позы/жеста руки [9–13], например, в работе [14] исследователи сделали акцент на обучении модели распознаванию языка жестов. С точки зрения непосредственного решения задачи сегментации рук пользователя актуальным представляется использование широкого инструментария сегментационных моделей NN.

В соответствии с приведенной выше информацией, наиболее подходящим решением в контексте настоящего исследования является модель Mask R-CNN (SpineNet-190) [5], поскольку она ориентирована на выделение сегментов различного рода объектов, требует небольшого количество данных для обучения и при этом характеризуется довольно высоким показателем точности детектирования.

Методы построения карт глубины могут быть классифицированы на две группы: аппаратные [15] и методы, основанные на использовании нейросетевых моделей [16–18]. Во избежание лишних трат на специальную аппаратуру, поддерживающую возможность съемки с собственной картой глубины, была выбрана модель FCRN-Depth-Prediction [18]. Данная модель позволяет строить карту глубины по 2D изображению. Преимуществом FCRN-DepthPrediction в сравнении с другими решениями данной группы является небольшое количество параметров обучения (train parameters), из чего следует, что для обучения требуется меньше данных, чем для аналогичных методов. Архитектура данной сети построена на основе остаточной нейронной сети ResNet-50 [19] и дополняет её Upsampling слоями, которые повышают точность предсказанных карт глубины.

Ожидается, что совместное использование выбранного метода, осуществляющего сегментацию рук пользователя, в совокупности с применением нейронной сети, реализующей детектирование объектов, и нейросетевой модели, обеспечивающей построение карты глубины изображений, позволит достоверно устанавливать факты соприкосновения пользователя с объектом и соответственно обеспечить решение задачи идентификации факта начала взаимодействия пользователя с объектами.

В контексте обеспечения трекинга пар пользователь-объект на видеопоследовательности до завершения процесса взаимодействия пользователя с объектом рассмотрим существующие методы трекинга объектов на видеопоследовательности. На сегодняшний день широкое распространение получили методы, основанные на использовании нейросетевых моделей [20–23]. По результатам проведенного анализа наиболее перспективным представляется решение, предложенное в работе [21]. Авторами была разработана модель, представляющая собой глубокую нейронную сеть, которая принимает в качестве входных данных видеокадры, а возвращает для каждого кадра координаты ограничивающего прямоугольника отслеживаемого объекта. Представленный метод расширяет глубокую сверточную нейронную сеть YOLO [24] в пространственно-временную область с использованием рекуррентных нейронных сетей (ROLO). Авторами была проведена обширная эмпирическая оценка эффективности ROLO в сравнении с 10 существующими решениями задачи трекинга. Сравнительная оценка проводилась

на наборе из 30 сложных общедоступных видеопоследовательностей. Результат оценки значений удачных построений за единичный проход (OPE) представленной в работе модели равен 0.458. В целом, все представленные выше методы демонстрируют достаточно высокую точность трекинга объектов на видеопоследовательности, однако наибольшей эффективностью при сохранении высокого уровня точности обладает метод [21]. Данное решение может быть использовано в режиме реального времени, что положительным образом сказывается на применимости метода к реальным условиям и обусловливает выбор данного решения для обеспечения трекинга пар пользователь-объект взаимодействия на видеопоследовательности.

Таким образом, за счет комбинирования и модернизации рассмотренных выше подходов, методов и моделей в рамках данной работы будет разработан метод отслеживания процессов взаимодействия пользователей с объектами по видеопоследовательностям, не требующий применения специализированных аппаратных решений и отличающийся возможностью использования в режиме реального времени при сохранении высокого уровня точности детектирования и трекинга.

В соответствии с результатами проведенного анализа связанных методов и подходов, для обеспечения отслеживания взаимодействия пользователей с объектами в рамках исследования предлагается авторский метод решения данной задачи, включающий в себя следующие основные этапы:

1. Детектирование на кадрах видеопоследовательности пар вида пользователь-объект взаимодействия:

a. Детектирование пользователей и объектов с использованием предобученной модели нейронной сети RetinaNet;

b. Определение потенциальных пар пользователь-объект взаимодействия с использованием разработанного алгоритма предварительной идентификации взаимодействия.

2. Идентификация фактов начала взаимодействия:

a. Получение данных о пользователе – детектирование и сегментация рук пользователя с использованием нейросетевой модели Mask R-CNN;

b. Формирование карты глубины для кадров видеопоследовательности с помощью нейронной сети FCRN-Depth-Prediction;

c. Идентификация фактов взаимодействия с использованием разработанного алгоритма, основанного на оценке пространственного положения рук пользователей относительно потенциальных объектов взаимодействия.

3. Анализ процессов взаимодействия пользователей с объектами:

a. Трекинг пар пользователь-объект в процессе взаимодействия за счет применения метода отслеживания объектов на основе рекуррентных нейронных сетей Rolo [21];

b. Идентификация завершения процессов взаимодействия.

Как можно заметить, предложенный метод содержит три ключевых этапа, каждый из которых направлен на решение отдельной группы подзадач. Важным требованием к разрабатываемому решению является обеспечение его работоспособности в режиме реального времени. Таким образом, первый этап применяется к поступающим кадрам видеоряда и направлен не только на идентификацию наборов пользователей и объектов на исследуемой сцене, но также и на снижение объема данных, для которых потребуется осуществлять более глубокую и ресурсоемкую обработку. Второй этап посвящен идентификации пользователей, взаимодействующих с теми или иными объектами на сцене. Важно подчеркнуть, что потенциально каждый пользователь может взаимодействовать с некоторым набором объектов, что выдвигает дополнительные требования к точности производимой оценки в отношении наличия взаимодействия в рамках каждой конкретной пары пользователь-объект. Третий этап связан с отслеживанием тех пар пользователь-объект, для которых было подтверждено наличие взаимодействия. В рамках данного этапа осуществляется трекинг соответствующих пар, а также производится периодическая проверка на завершение процессов взаимодействия. Рассмотрим детальнее каждый из этапов предложенного метода.

Детектирование на кадрах видеопоследовательности пар вида пользователь-объект взаимодействия

Для детектирования пользователей и объектов на кадрах видеоряда в качестве модели применяется предобученная нейронная сеть RetinaNet. Для определения объектов на сцене на вход нейронной сети подается одиночный кадр из видеопотока, предварительно очищенный от шумов за счет применения гауссовой фильтрации. Результатом работы нейронной сети является информация об объектах на изображении: координаты ограничивающих прямоугольников, вероятности принадлежности каждого объекта к определенному классу. Важно отметить, что к числу классов, распознаваемых RetinaNet, также относится и человек, таким образом, по результатам работы данной нейронной сети выполняется детектирование не только объектов, но и непосредственно пользователей.

Далее осуществляется определение потенциальных пар пользователь – объект взаимодействия, данные пары выявляются в соответствии со следующим алгоритмом:

Для каждого детектированного на некотором изображении пользователя U_i из набора U_{img}, вычислим центр соответствующего ему ограничивающего прямоугольника:

$$Center_i(Box_i) = \frac{(B_x + A_x, B_y + A_y)}{2},$$

$$\{Box_i \in Box_{img} | U_{img} \rightarrow Box_{img}\},$$

где Box_i – ограничивающий прямоугольник, соответствующий детектированному пользователю U_i; A и B – координаты диагональных вершин ограничивающего прямоугольника Box_i; U_{img} – набор детектированных на изображении пользователей.

Для каждого детектированного на некотором изображении объекта O_i из набора O_{img} вычислим центр соответствующего ему ограничивающего прямоугольника:

$$OCenter_i(OBox_i) = \frac{(B_x + A_x, B_y + A_y)}{2},$$

$$\{OBox_i \in OBox_{img} \mid O_{img} \to OBox_{img}\},$$

где $OBox_i$ – ограничивающий прямоугольник, соответствующий детектированному объекту O_j; A и B – координаты диагональных вершин ограничивающего прямоугольника $OBox_i$; O_{img} – набор детектированных на изображении объектов.

Далее для каждого элемента из наборов Box_{img} и $OBox_{img}$ произведем оценку размера соответствующих ограничивающих прямоугольников согласно следующему выражению:

$$Size(Box) = \frac{\left\|(B_x - A_x, B_y - A_y)\right\|}{2},$$

где A и B – координаты диагональных вершин соответствующего ограничивающего прямоугольника из наборов Box_{img} или $OBox_{img}$.

Для каждого пользователя U_i из набора U_{img} определим расстояния между данным пользователем и объектами $O_j \in O_{img}$, детектированными на исследуемой сцене. В качестве расстояния между пользователем U_i и объектом O_j будем использовать расстояния между центрами соответствующих им ограничивающих прямоугольников:

$$Dist(U_j, O_j) = \left\|Center_i - OCenter_j\right\|.$$

Определение потенциального набора пар вида пользователь – объект взаимодействия $PairSet_{pot}$ осуществляется в соответствии со следующим выражением:

$$Pair(U_i, O_j) = \begin{cases} 1, & \dfrac{Dist(U_i, O_j)}{Size(Box_i) + Size(OBox_j)} \le \beta \\ 0, & \dfrac{Dist(U_i, O_j)}{Size(Box_i) + Size(OBox_j)} > \beta, \end{cases}$$

$$Pair(U_i, O_j) \in PairSet_{pot}$$

где β – некоторое строго положительное пороговое значение, характеризующее предельную удаленность пользователя U_i от объекта O_j, при которой объект рассматривается в качестве потенциального объекта взаимодействия для данного пользователя.

По результатам работы данного алгоритма формируется набор потенциальных пар пользователь-объект взаимодействия, которые подлежат дальнейшему анализу на предмет фактического наличия взаимодействия.

Идентификация факта начала взаимодействия

Для идентификации факта взаимодействия пользователя с объектом необходимо предварительно произвести оценку пространственного положения рук каждого пользователя относительно соответствующих потенциальных объектов взаимодействия, что требует пред-

варительного решения задач по детектированию рук пользователя, детектированию объекта, а также построению карты глубины соответствующего изображения.

Для детектирования и сегментации рук пользователя на изображении была выбрана модель Mask R-CNN (SpineNet-190), характеризующаяся высокими показателями точности своей работы и относительно низкими требованиями с точки зрения вычислительных ресурсов для обеспечения собственного функционирования в реальном времени. Входными данными модели является кадр видеопоследовательности, а выходными – сегментированное изображение, с выделенными контурами рук пользователей, а также соответствующими ограничивающими прямоугольниками.

Полученные результаты применения модели масштабируются с целью формирования матрицы соответствия пикселей входного изображения выявленным сегментам рук пользователей. Данная матрица совмещается с результатами детектирования пользователей, полученными на предыдущем этапе, по принципу перекрытия областей с целью обеспечения связности между полученными контурами рук и сформированным ранее набором пользователей U_{img}. Таким образом, каждый выявленный на изображении сегмент руки Arm_k ассоциирован с конкретным пользователем U_i.

Так как детектирование потенциальных объектов взаимодействия было выполнено в рамках предыдущего этапа, далее для оценки относительного положения объектов и рук пользователей в трехмерном пространстве необходимо построить карту глубины текущего кадра видеопоследовательности. По результатам проведенного анализа существующих методов построения карты глубины было принято решение воспользоваться моделью нейронной сети FCRN-DepthPrediction, выходными данными которой является матрица значений глубины. Проведение операции масштабирования данной матрицы до размера исходного изображения позволяет обеспечить соответствие пикселей входного изображения и полученных значений глубины.

Следующий шаг заключается в оценке положения рук каждого пользователя относительно соответствующих потенциальных объектов взаимодействия в рамках определённых на предыдущем этапе пар пользователь – объект взаимодействия. Для каждой пары $P_n = =(U_i; O_j)$ из набора потенциальных пар пользователь – объект взаимодействия $PairSet_{pot}$ оценка пространственного положения объектов относительно рук пользователей производится в соответствии со следующим алгоритмом идентификации взаимодействия:

Для каждой руки $\mathrm{Arm_k} \in \mathrm{Arm_{img}}$, детектированной на кадре видеопоследовательности и ассоциированной с некоторым пользователем U_i, определим центр соответствующего ограничивающего прямоугольника:

$$ACenter_k(ABox_k) = \frac{(B_x + A_x, B_y + A_y)}{2},$$

$$\{ABox_k \in ABox_{img} \mid Arm_{img} \rightarrow ABox_{img}\},$$

где $ABox_k$ – ограничивающий прямоугольник, соответствующий детектированной руке Arm_k; A и B – координаты диагональных вершин ограничивающего прямоугольника $ABox_k$; Arm_{img} – набор задетектированных на изображении рук.

Определим вектор, характеризующий положение центра объекта O_j относительно центра руки ACenter_k, следующим образом:

$$direct_n = OCenter_j - ACenter_k .$$

В общем случае карту глубины можно рассматривать как дискретно заданную функцию F, определенную на множестве пар (x, y), где x и y – значения индексов пикселей исследуемого кадра видеопоследовательности по горизонтальной и вертикальной осям соответственно. В качестве характеристики изменения удаленности руки Arm_k и объекта O_j в направлении перпендикулярном плоскости изображения возьмем двустороннюю разностную оценку производной данной функции dF по направлению direct_n. Произведем расчет данных оценочных значений на отрезке [ACenter_k, OCenter_j] в соответствии со следующим выражением:

$$\frac{dF}{d(direct_n)}(Point_q) =$$

$$= \frac{F(Point_0 + \gamma \cdot \frac{direct_n}{|direct_n|}) - F(Point_0 - \gamma \cdot \frac{direct_n}{|direct_n|})}{2 \cdot \gamma},$$

где γ – параметр шага, а $Point_q = (x_q, y_q)$ – q-ый пиксель изображения, который принадлежит отрезку [ACenter_k, OCenter_j].

Результирующий вывод относительно наличия взаимодействия между пользователем и объектом осуществляется в соответствии со следующим выражением:

$$\forall\ Point_q \in Points, \left|\frac{dF}{d(direct_n)}(Point_q)\right| \le \theta,$$

где $Points$ – набор пикселей изображения, которые принадлежат отрезку [ACenter_k, OCenter_j]; θ – некоторое строго положительное пороговое значение, характеризующее предельное изменение величины удаленности при смещении вдоль отрезка [ACenter_k, OCenter_j] и выполняющее роль критерия идентификации резких изменений значений глубины в данном направлении. В случае, если представленное выше выражение истинно, делается вывод о наличии взаимодействия пользователя с объектом в рамках пары $P_n = (U_i; O_j)$.

Таким образом, по итогам данного этапа для каждой пары P_n из набора $\text{PairSet}_\text{pot}$ делается окончательный вывод о наличии взаимодействия между пользователями и объектами, в результате чего формируется набор пар вида пользователь – объект взаимодействия $PairSet_{res}$.

Анализ процессов взаимодействия пользователей с объектами

Как было сказано ранее, данный этап связан с отслеживанием тех пар пользователь-объект, для которых было подтверждено наличие взаимодействия. В

рамках данного этапа осуществляется трекинг соответствующих пар, а также производится периодическая проверка на завершение процессов взаимодействия. Трекинг пар пользователь-объект в процессе их взаимодействия обеспечивается за счет применения метода отслеживания объектов на основе рекуррентных нейронных сетей Rolo [21]. На вход данному методу подается информация об отслеживаемом объекте – соответствующий ограничивающий прямоугольник, а также поступающие кадры исследуемой видеопоследовательности. Для каждого поданного кадра данный метод возвращает обновленные параметры ограничивающего прямоугольника для отслеживаемого объекта. Таким образом, обеспечивается отслеживание всех пар P_n из результирующего набора $PairSet_{res}$. Проверка на завершение взаимодействия для каждой пары P_n инициируется на каждом j-ом кадре видеопоследовательности и реализуется в соответствии с алгоритмом идентификации взаимодействия, рассмотренном ранее в рамках второго этапа. По результатам выполнения данной проверки определяется актуальный набор взаимодействующих пар пользователь-объект $PairSet_{final}$.

Обобщенная алгоритмическая модель разработанного метода отслеживания процессов взаимодействия пользователей с объектами представлена ниже на рис. 1. Важно отметить, что представленная ниже алгоритмическая модель предполагает анализ лишь каждого m-го кадра видеопоследовательности. При этом доля анализируемых кадров может быть как увеличена для достижения большей точности отслеживания, так и уменьшена с целью минимизации затрачиваемых вычислительных ресурсов. Кроме того, данная алгоритмическая модель учитывает возможность возникновения случаев множественного взаимодействия как некоторого пользователя с группой объектов, так и нескольких пользователей с одним объектом за счет предложенной реализации процесса выявления потенциальных пар пользователь-объект.

Ниже на рис. 2 представлены поэтапные результаты работы сформированной модели на некотором кадре видеопоследовательности.

Где а) исходный кадр видеопоследовательности; б) результаты детектирования пользователей и объектов; в) результаты выявления потенциальных пар пользователь-объект взаимодействия; г) результаты детектирования и сегментации рук пользователей; д) результаты формирования карты глубины для данного кадра видеоряда; е) результаты оценки производной функции dF в соответствии с алгоритмом идентификации взаимодействий; ж) результат идентификации взаимодействий между потенциальными парами пользователь-объект.

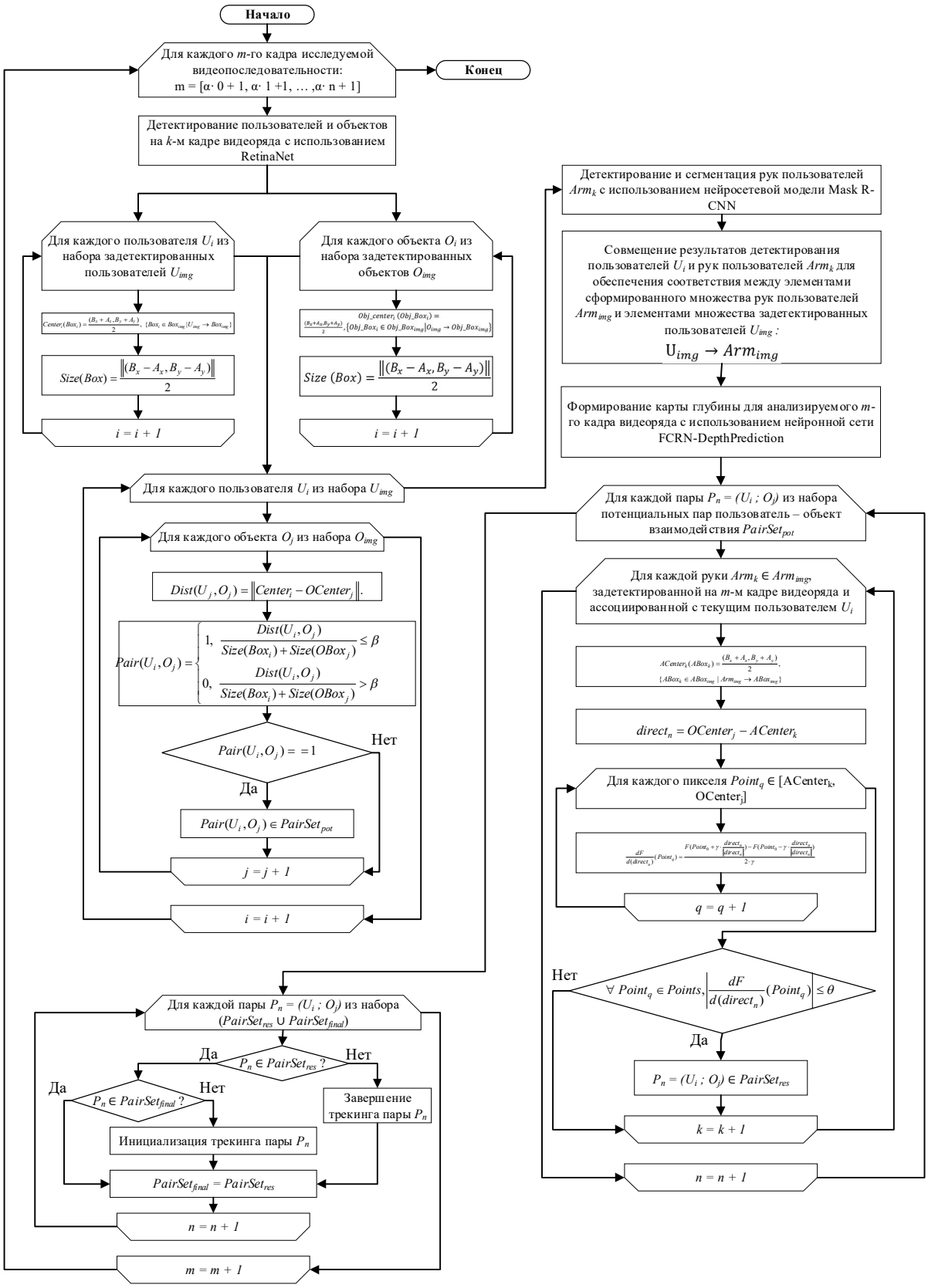


Рис. 1. Обобщенная алгоритмическая модель разработанного метода отслеживания процессов взаимодействия пользователей с объектами на видеопоследовательности

Fig. 1. Generalized algorithmic model of the developed method for tracking the processes of user interaction with objects in a video sequence

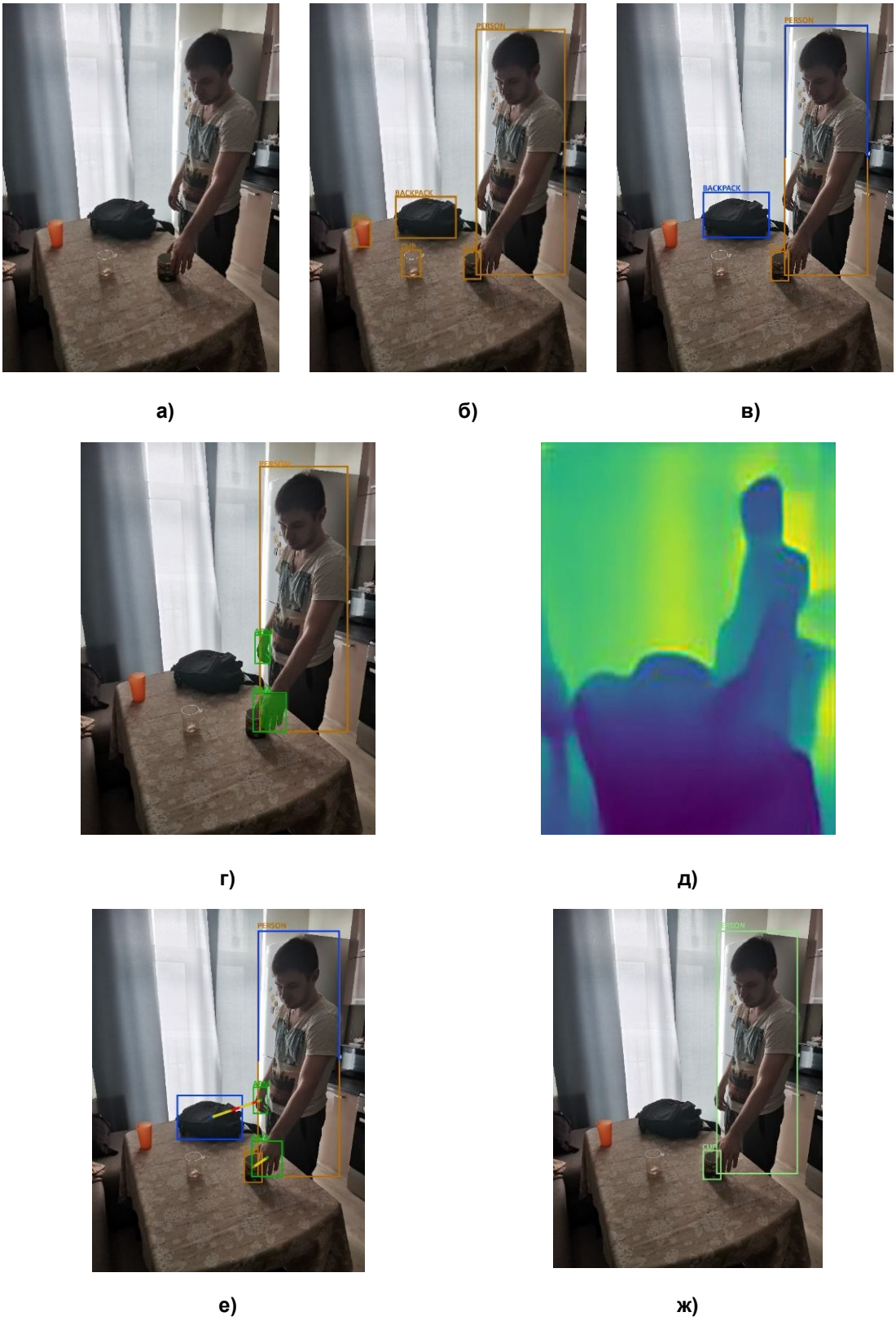


Рис. 2. Поэтапные результаты работы предложенной алгоритмической модели на некотором кадре видеопоследовательности

Fig. 2. Step-by-step results of the proposed algorithmic model for a certain frame of the video sequence

На рис. 2,а представлен обрабатываемый кадр видеопоследовательности. Согласно разработанной модели, в первую очередь осуществляется детектирование пользователей и объектов с использованием нейросетевой модели RetinaNet (рис. 2,б). На данном кадре был детектирован один пользователь и четыре объекта: три чашки и один рюкзак. Далее, в соответствии с разработанным алгоритмом, выполняется поиск потенциальных пар пользователь-объект взаимодействия, результаты работы данного алгоритма отражены на рис. 2,в. Было определено 2 потенциальные пары: пользователь-рюкзак, пользователь-чашка (ограничивающий прямоугольник отображен на рисунке). После того, как потенциальные пары взаимодействующих пользователей и связанных с ними объектов определены, выполняется детектирование и сегментация рук пользователей с использованием нейросетевой модели Mask R-CNN. На рис. 2,г представлены найденные на данном кадре видеопоследовательности сегменты рук пользователей. Следующим шагом является построение карты глубины для соответствующего изображения, полученные с помощью нейронной сети FCRN-DepthPrediction результаты представлены на рис. 2,д. Далее производится анализ отдельных участков полученной карты глубины, и, в частности, осуществляется двусторонняя разностная оценка производной функции dF по направлению, связывающему центр руки пользователя и центр соответствующего объекта. Результаты, полученные для данного кадра видеопоследовательности, представлены на рис. 2,е. и отражены соответствующими линиями, при этом цвет линии по умолчанию является желтым, но на участках, где оценка производной функции *dF* превосходит заданное пороговое значение, цвет линии становится красным. Для потенциальной пары пользователь- рюкзак данная линия имеет красные участки, что, в соответствии с логикой разработанной модели, свидетельствует об отсутствии взаимодействия пользователя с данным объектом. Таким образом, в качестве фактически взаимодействующих пар, моделью выделена лишь пара пользователь-чашка (ограничивающий прямоугольник отображен на рис. 2,ж). Данный результат работы алгоритма идентификации взаимодействия следует считать корректным, так как в действительности взаимодействие пользователя с рюкзаком или иным объектом сцены, помимо чашки, отсутствует.

Далее перейдем к оценке реализации предложенного метода в контексте качества детектирования и отслеживания взаимодействий пользователей с объектами на видеопоследовательностях.

Результаты и их обсуждение

Апробация и оценка качества функционирования разработанного метода отслеживания процессов взаимодействия пользователей с объектами производилась на основе тестового набора данных, включающего в себя 1000 ви-

деопоследовательностей длительностью до 20 секунд. В тестовом наборе данных присутствовали видеопоследовательности как с одним, так и с несколькими пользователями, взаимодействующими с объектами сцены. На каждой тестовой видеопоследовательности представлено как минимум одно взаимодействие одного из следующих типов: один пользователь – один объект, один пользователь – несколько объектов, несколько пользователей – один объект. Объекты взаимодействия представляют собой объекты 15 различных классов, подлежащих детектированию нейронной сетью RetinaNet. Тестовый набор данных состоит из видеопоследовательностей, сформированных при раз-личных условиях освещенности: 30%, 50% и 100%, где за 100% взят нормативный уровень освещенности для офисных помещений[1].

На основе данных, полученных в результате применения разработанного метода к набору тестовых видеопоследовательностей, были сформированы различные количественные оценки точности работы предложенного решения. В частности, были определены усредненные показатели точности (accuracy, recall, precision) детектирования взаимодействий в разрезе классов объектов взаимодействия и уровней освещенности. Значения соответствующих показателей были вычислены следующим образом:

Для каждой видеопоследовательности на основе анализа результатов применения двух первых этапов разработанного метода были сформированы матрицы ошибок идентификации взаимодействия M_{ij}. Далее для каждой матрицы M_{ij} были определены показатели accuracy A_{ij}, recall R_{ij} и precision Pr_{ij} в соответствии со следующими формулами:

$$A_{ij}=\frac{TP_{ij}+TN_{ij}}{TP_{ij}+FP_{ij}+FN_{ij}+TN_{ij}},$$

$$R_{ij}=\frac{TP_{ij}}{TP_{ij}+FN_{ij}},$$

$$Pr_{ij}=\frac{TP_{ij}}{TP_{ij}+FP_{ij}}.$$

На заключительном шаге осуществлялось усреднение данных показателей в разрезе отдельных классов объектов $j = \{1 \dots 15\}$ и уровней освещенности сцены $L = \{30\%, 50\%, 100\%\}$. С этой целью из общего набора видеопоследовательностей формировались специфические наборы видеопоследовательностей V_set_{jL}, включающие в себя лишь видеопоследовательности V_k с заданным уровнем освещенности L и содержащие на сцене объекты j-го класса. Таким образом, усредненные показатели точности A_{jL}, R_{jL}, Pr_{jL} были определены в соответствии со следующими выражениями:

$$A_{jL}=\frac{1}{n_{jL}}\sum_{k=1}^{n_{jL}} A_{jL}(V_k), \quad V_k \in V\_set_{jL};$$

[1] ГОСТ Р 55710-2013. Освещение рабочих мест внутри зданий. Нормы и методы измерений: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 8 ноября 2013 г. N 1364-ст. URL: http://docs.cntd.ru/document/1200105707 (дата обращения: 27.04.2020). Текст: электронный.

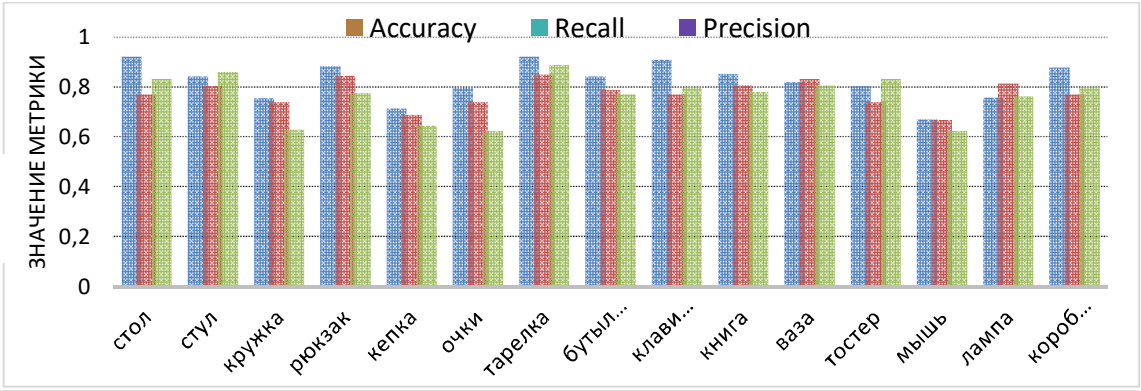
$$R_{jL}=\frac{1}{n_{jL}}\sum_{k=1}^{n_{jL}} R_{jL}(V_k),\quad V_k\in V_{set_{jL}};$$

$$Pr_{jL}=\frac{1}{n_{jL}}\sum_{k=1}^{n_{jL}} Pr_{jL}(V_k),\quad V_k\in V\_set_{jL}.$$

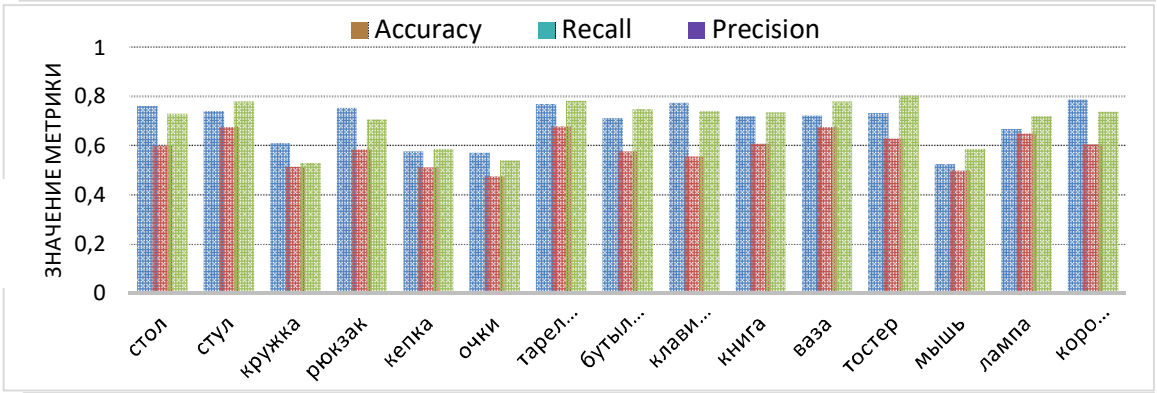
Ниже на рис. 3 представлены диаграммы полученных значений показателей A_{jL}, R_{jL}, Pr_{jL} для различных классов объектов j при разных уровнях освещенности сцены L.

Из представленных выше рисунков можно заключить, что точность работы метода существенным образом зависит от уровня освещенности и значительно снижается при понижении уровня освещенности.

а)



б)



в)

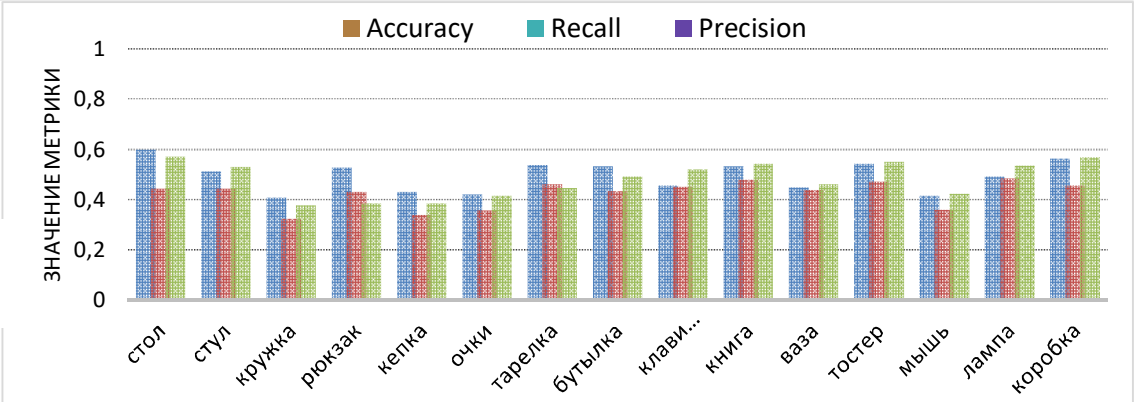


Рис. 3. Диаграммы полученных значений показателей A_{jL}, R_{jL}, Pr_{jL} для различных классов объектов j при уровнях освещенности сцены L: 100% (**а**), 50% (**б**), 30% (**в**)

Fig. 3. Diagrams of the obtained values of the indicators AjL, RjL, PrjL for various classes of objects j at scene illumination levels L: 100% (**а**), 50% (**б**), 30% (**в**)

При этом стоит отметить, что при переходе от уровня освещенности в 100% к 50% наибольшее снижение испытывает показатель R_{jL}, при следующем переходе к 30% уровню освещенности все показатели испытывают примерно равную степень снижения. Полученные результаты могут быть объяснены тем, что понижение уровня освещенности сказывается в первую очередь на качестве детектирования объектов. Таким образом, пропускается часть потенциальных пар пользователь-объект, при этом качество идентификации взаимодействий ухудшается в меньшей степени, что обуславливает меньшую долю ошибок вида FP. Однако при еще больших снижениях уровня освещенности помимо ошибок детектирования, возникают также значительные погрешности при построении карт глубины, что существенным образом сказывается на качестве идентификации взаимодействий. Данная ситуация наблюдается при применении разработанного метода на видеопоследовательностях с уровнем освещенности в 30%. Важно отметить, что класс объекта также влияет на качество детектирования взаимодействий, в частности рассчитанные показатели имеют более низкие значения для объектов следующих классов: кружка, кепка, очки, мышь. Предполагается, что подобные результаты связаны с меньшим размером данных объектов, а также с их геометрическими особенностями.

Усредненные по классам значения показателей A_{jL}, R_{jL}, Pr_{jL} для видеопоследовательностей с различными уровнями освещенности равны {0.82, 0.78, 0.76}, {0.70, 0.59, 0.70}, {0.49, 0.43, 0.48} для 100%, 50% и 30% уровней освещенности соответственно. В целом на основе полученных результатов можно заключить, что разработанный метод демонстрирует приемлемое качество детектирования взаимодействий для видеопоследовательностей с уровнями освещенности в 100% и 50%.

Для осуществления оценки непосредственно качества отслеживания процессов взаимодействия пользователей с объектами было принято решение в качестве основной метрики воспользоваться долей корректно отслеженных взаимодействий. Взаимодействие рассматривалось как корректно отслеженное в случае, если:

Взаимодействие было задетектировано не позднее чем через время $\Delta t=0.05T$ от фактического начала взаимодействия, где T соответствует интервалу времени, когда взаимодействие фактически имело место;

Временной интервал фиксации взаимодействия посредством разработанного метода соответствует реальному не менее, чем на 70% с нулевым смещением;

Момент времени завершения взаимодействия, полученный с использованием разработанного метода, соответствует фактическому с отклонением не более чем $\pm 0.05T$.

Ниже на рис. 4 представлена диаграмма доли корректно отслеженных взаимодействий для каждой видеопоследовательности из тестового набора.

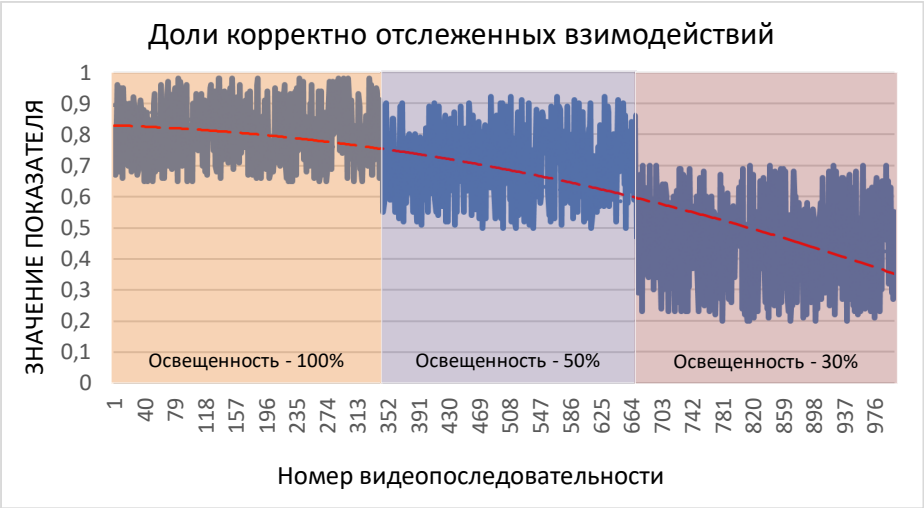


Рис. 4. Диаграмма доли корректно отслеженных взаимодействий для тестового набора видеопоследовательностей

Fig. 4. Diagram of the proportion of correctly tracked interactions for a test set of video sequences

Разработанный метод демонстрирует довольно высокую точность отслеживания взаимодействий для видеопоследовательностей со 100% уровнем освещенности. При этом по аналогии с качеством детектирования взаимодействий при использовании предложенного метода на видеопоследовательностях с более низким уровнем освещенности точность отслеживания взаимодействий становится существенно ниже, что проиллюстрировано на рисунке выше соответствующей линией тренда. Усредненная доля корректно отслеженных взаимодействий для наборов видеопоследовательностей с уровнем освещенности 100%, 50% и 30% составила соответственно 0.81, 0.71 и 0.45, а оценка среднеквадратического отклонения данной величины 0.09, 0.12, 0.15. Полученные результаты свидетельствуют о том, что со снижением уровня освещенности имеет место не только снижение точности работы предложенного решения, но и повышение вариабельности результатов. Тем не менее, качество работы метода остается на приемлемом уровне для видеопоследовательностей даже с 50% уровнем освещенности. Однако при дальнейшем снижении освещенности наблюдается критическое снижение качества отслеживания взаимодействий и уже при 30% уровне освещенности метод становится непригоден к практическому применению. Таким образом, предложенное решение позволяет успешно детектировать и отслеживать взаимодействия пользователей с различными классами объектов по видеопоследовательностям и в определенной мере является пригодным к использованию в условиях неполной освещенности сцены.

Выводы

По результатам апробации предложенного метода отслеживания взаимодействий пользователей с объектами на

тестовом наборе из 1000 видеопоследовательностей, предложенное решение показало довольно высокое качество детектирования и отслеживания взаимодействий для видеопоследовательностей с уровнями освещенности 100% и 50%. Усредненные показатели точности (accuracy, recall, precision) детектирования взаимодействий для соответствующих наборов видеопоследовательностей имеют значение {0.82, 0.78, 0.76} и {0.70, 0.59, 0.70}, а усредненные доли корректно отслеженных взаимодействий для данных наборов видеопоследовательностей составили 81% и 71%. Однако при более низких уровнях освещенности точность работы разработанного решения продемонстрировала значительное снижение – на наборе видеопоследовательностей с уровнем освещенности 30% усредненные показатели точности составили {0.49, 0.43, 0.48}, а доля корректно отслеженных взаимодействий снизилась до 45%. Полученные результаты могут быть объяснены тем, что при снижении уровня освещенности также существенно снижается качество функционирования заложенных в основу предложенного метода моделей. Таким образом, разработанный метод в определенной мере является устойчивым к изменению уровня освещенности сцены и обеспечивает успешное решение задачи детектирования и отслеживания взаимодействия пользователей с различными классами объектов по видеопоследовательности, не требуя при этом применения специализированного оборудования.

В рамках дальнейших исследований предполагается осуществить модернизацию разработанного метода с точки зрения потребляемых вычислительных ресурсов с целью его последующей интеграции в киберфизическую систему, ориентированную на анализ пользовательской активности.

Список литературы

1. Masking salient object detection, a mask region-based convolutional neural network analysis for segmentation of salient objects / B.A. Krinski, D.V. Ruiz, G.Z. Machado, E. Todt // 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE). IEEE, 2019. P. 55-60.

2. He K., Girshick R., Dollár P. Rethinking imagenet pre-training // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. P. 4918-4927.

3. Coco common objects in context [Электронный ресурс] // Detection Leaderboard [сайт]. URL: https://cocodataset.org/#detection-leaderboard.

4. Lin T. Y. et al. Microsoft coco: Common objects in context //European conference on computer vision. 2014. С. 740-755.

5. SpineNet: Learning scale-permuted backbone for recognition and localization / X. Du, T.Y. Lin, P. Jin, G. Ghiasi, M. Tan, Y Cui., X. Song // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. P. 11592-11601.

6. Detecting and recognizing human-object interactions / G. Gkioxari, R. Girshick, P. Dollár, K. He // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. P. 8359-8367.

7. Scaling human-object interaction recognition through zero-shot learning / L. Shen, S. Yeung, J. Hoffman, G. Mori, L. Fei-Fei // 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018. P. 1568-1576.

8. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields / Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, Y. Sheikh // IEEE transactions on pattern analysis and machine intelligence. 2019. Vol. 43. №. 1. P. 172-186.

9. 3d hand pose detection in egocentric rgb-d images / G. Rogez, M. Khademi, III Supancic, J., J.M.M. Montiel, D. Ramanan // European Conference on Computer Vision. Springer, Cham, 2014. P. 356-371.

10. Hand pose estimation and hand shape classification using multi-layered randomized decision forests / C. Keskin, F. Kıraç, Y.E. Kara, L. Akarun // European Conference on Computer Vision. Springer, Berlin, Heidelberg. 2012. P. 852-863.

11. Oberweger M., Wohlhart P., Lepetit V. Hands deep in deep learning for hand pose estimation // Proceedings of 20th Computer Vision Winter Workshop (CVWW). 2015. P. 21-30. arXiv preprint arXiv:1502.06807.

12. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations / G. Garcia-Hernando, S. Yuan, S. Baek, T. K. Kim // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. P. 409-419.

13. Moon G., Chang J.Y., Lee K.M. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map // Proceedings of the IEEE conference on computer vision and pattern Recognition. 2018. P. 5079-5088.

14. Redmon J., Farhadi A. Yolov3: An incremental improvement //arXiv preprint arXiv:1804.02767. 2018.

15. Murawski K., Murawska M., Pustelny T. Optimizing the light source design for a sensor to measure the stroke volume of the artificial heart // 13th Conference on Integrated Optics: Sensors, Sensing Structures, and Methods. International Society for Optics and Photonics, 2018. Vol. 10830. P. 1083006.

16. Karsch K., Liu C., Kang S. B. Depth extraction from video using non-parametric sampling // European conference on computer vision. Springer, Berlin, Heidelberg, 2012. P. 775-788.

17. Eigen D., Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture // Proceedings of the IEEE international conference on computer vision. 2015. P. 2650-2658.

18. Deeper depth prediction with fully convolutional residual networks / I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab // 2016 Fourth international conference on 3D vision (3DV). IEEE, 2016. P. 239-248.

19. Deep residual learning for image recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. P. 770-778.

20. Cheong Y. Z., Chew W. J. The Application of Image Processing to Solve Occlusion Issue in Object Tracking // MATEC Web of Conferences. EDP Sciences, 2018. Vol. 152. P. 03001.

21. Spatially supervised recur-rent convolutional neural networks for visual object tracking / G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, Z. He // 2017 IEEE International Symposium on Circuits and Systems (IS-CAS). IEEE, 2017. P. 1-4.

22. Real-time visual object tracking with natural language description / Q. Feng, V. Ablavsky, Q. Bai, G. Li, S. Sclaroff // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020. P. 700-709.

23. Deep reinforcement learning for visual object tracking in videos / D. Zhang, H. Maei, X. Wang, Y.F. Wang // arXiv preprint arXiv:1701.08936. 2017.

24. You only look once: Unified, real-time object detection / J. Redmon, S. Divvala, R. Girshick, A. Farhadi // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. P. 779-788.

References

1. Krinski B.A., Ruiz D.V., Machado G.Z., Todt E. Masking salient object detection, a mask region-based convolutional neural network analysis for segmentation of salient objects. *Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE), IEEE*, 2019, pp. 55-60.

2. He K., Girshick R., Dollár P. Rethinking imagenet pre-training. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4918-4927.

3. Coco common objects in context [Electronic resource]. Detection Leaderboard [site]. Available at: https://cocodataset.org/#detection-leaderboard.

4. Lin T. Y. et al. Microsoft coco: Common objects in context. *European conference on computer vision*, 2014, pp. 740-755.

5. Du X., Lin T.Y., Jin P., Ghiasi G., Tan M., Cui Y., Song X. SpineNet: Learning scale-permuted backbone for recognition and localization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11592-11601.

6. Gkioxari G., Girshick R., Dollár P., He K. Detecting and recognizing human-object interactions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359-8367.

7. Shen L., Yeung S., Hoffman J., Mori G., Fei-Fei L. Scaling human-object interaction recognition through zero-shot learning. *IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE*, 2018, pp. 1568-1576.

8. Cao Z., Hidalgo G., Simon T., Wei S. E., Sheikh Y. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 2019, vol. 43, no. 1, pp. 172-186.

9. Rogez G., Khademi M., Supancic III, J., Montiel J.M.M., Ramanan D. 3d hand pose detection in egocentric rgb-d images. *European Conference on Computer Vision, Springer, Cham*, 2014, pp. 356-371.

10. Keskin C., Kıraç F., Kara Y.E., Akarun L. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. *European Conference on Computer Vision. Springer*. Berlin, Heidelberg, 2012, pp. 852-863.

11. Oberweger M., Wohlhart P., Lepetit V. Hands deep in deep learning for hand pose estimation. *Proceedings of 20th Computer Vision Winter Workshop (CVWW)*, 2015, pp. 21-30. arXiv preprint arXiv:1502.06807.

12. Garcia-Hernando G., Yuan S., Baek S., Kim T. K. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. *Proceedings of the IEEE conference on computer vision and pattern recognit*ion, 2018, pp. 409-419.

13. Moon G., Chang J.Y., Lee K.M. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. *Proceedings of the IEEE conference on computer vision and pattern Recognition*, 2018, pp. 5079-5088.

14. Redmon J., Farhadi A. Yolov 3: An incremental improvement. *arXiv preprint arXiv*:1804.02767. 2018.

15. Murawski K., Murawska M., Pustelny T. Optimizing the light source design for a sensor to measure the stroke volume of the artificial heart. *13th Conference on Integrated Optics: Sensors, Sensing Structures, and Methods. International Society for Optics and Photonics*, 2018, vol. 10830, 1083006 p.

16. Karsch K., Liu C., Kang S. B. Depth extraction from video using non-parametric sampling. *European conference on computer vision*. Springer, Berlin, Heidelberg, 2012, pp. 775-788.

17. Eigen D., Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650-2658.

18. Laina I., Rupprecht C., Belagiannis V., Tombari F., Navab N. Deeper depth prediction with fully convolutional residual networks. *2016 Fourth international conference on 3D vision (3DV). IEEE*, 2016, pp. 239-248.

19. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

20. Cheong Y. Z., Chew W. J. The Application of Image Processing to Solve Occlusion Issue in Object Tracking. *MATEC Web of Conferences, EDP Sciences*, 2018, vol. 152, 03001 p.

21. Ning G., Zhang Z., Huang C., Ren X., Wang H., Cai C., He Z. Spatially supervised recur-rent convolutional neural networks for visual object tracking. *2017 IEEE Inter-national Symposium on Circuits and Systems (IS-CAS), IEEE*, 2017, pp. 1-4.

22. Feng Q., Ablavsky V., Bai Q., Li G., Sclaroff S. Real-time visual object tracking with natural language description. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 700-709.

23. Zhang D., Maei H., Wang X., Wang Y.F. Deep reinforcement learning for visual object tracking in videos. *arXiv preprint arXiv*:1701.08936. 2017.

24. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.

Информация об авторе / Information about the Author

Яковлев Роман Никитич, младший научный сотрудник лаборатории технологий больших данных социокиберфизических систем, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук,
г. Санкт-Петербург, Российская Федерация,
e-mail: iakovlev.r@mail.ru,
ORCID: https://orcid.org/0000-0002-6721-9707

Roman N. Iakovlev, Junior Researcher of Laboratory of Big Data Technologies in Socio-Cyberphysical Systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
St. Petersburg, Russian Federation,
e-mail: iakovlev.r@mail.ru,
ORCID: https://orcid.org/0000-0002-6721-9707