DOI: 10.15514/ISPRAS-2020-32(1)-7

Разработка алгоритма распознавания движений человека методами компьютерного зрения в задаче нормирования рабочего времени

С.Е. Штехин, ORCID: 0000-0003-2866-4864 <sergei.shtekhin@ocrv.ru>
Д.К. Карачёв, ORCID: 0000-0002-1008-2535 <denis.karachev@ocrv.ru>
Ю.А. Иванова, ORCID: 0000-0003-1575-6882 <yustina.ivanova@ocrv.ru>

ОАО Отраслевой центр разработки и внедрения,
Россия, г. Сочи, Триумфальный проезд, д.1

Аннотация. Цель исследования заключается в разработке и тестировании алгоритмов для распознавания по видео людей и инструментов, с которыми они работают в конкретный момент времени. В рамках исследования в качестве базового решения был предложен и реализован алгоритм, состоящий из нескольких этапов: распознавание в видео-кадрах людей и определение координат краевых точек прямоугольника, в котором находится человек; определение в видео кадрах координат ключевых точек обнаруженных людей; распознавание в видео-кадрах инструментов и определение координат их краевых точек; определение инструментов, с которыми человек работает в конкретный момент времени (время считается по номеру кадра из видео). Для реализации алгоритма было проведено исследование, в ходе которого было протестировано дообучение существующих моделей компьютерного зрения для следующих задач компьютерного зрения: детекция объектов (Object detection) и людей, в частности, определение ключевых точек людей (Pose estimation), наложение объектов (Object Overlaying). В качестве метрики для мультиклассификационной задачи определения инструментов, которые находятся в руках у человека в каждом кадре (Object Overlaying), использовались следующие показатели: точность, чувствительность и f1-мера. Алгоритм запущен на web-сервисе и протестирован специалистами.

Ключевые слова: нейронные сети; компьютерное зрение; распознавание движений человека; машинное зрение; машинное обучение; наложение объектов; детектирование объектов

Для цитирования: Штехин С.Е., Карачёв Д.К., Иванова Ю.К. Разработка алгоритма распознавания движений человека методами компьютерного зрения в задаче нормирования рабочего времени. Труды ИСП РАН, том 32, вып. 1, 2020 г., стр. 121-136. DOI: 10.15514/ISPRAS–2020–32(1)–7

Благодарности: Авторы благодарны ОАО РЖД.

Computer vision system for Working time estimation by Human Activities detection in video frames

S.E. Shtekhin, ORCID: 0000-0003-2866-4864 <ivanov@ispras.ru>
D.K. Karachev, ORCID: 0000-0002-1008-2535 <denis.karachev@ocrv.ru>
Yu.A. Ivanova, ORCID: 0000-0003-1575-6882 <yustina.ivanova@ocrv.ru>

Industry Center for Information Systems' Development and Deployment,
1, Triumphalny, Sochi, 109004, Russia

Abstract. The goal of the research is to develop and to test methods for detecting people, parametric points for their hands and their current working tools in the video frames. The following algorithms are implemented: humans bounding boxes coordinates detection in video frames; human pose estimation: parametric points detection for each person in video frames; detection of the bounding boxes coordinates of the defined tools in video frames; estimation of which instrument the person is using at the certain moment. To implement algorithms, the existing computer vision models are used for the following tasks: Object detection, Pose estimation, Object overlaying. Machine learning system for working time detection based on computer vision is developed and deployed as a web-service. Recall, precision and f1-score are used as a metric for multi-classification problem. This problem is defined as what type of tool the person uses in a certain frame of video (Object Overlaying). Problem solution for action detection for the railway industry is new in terms of work activity estimation from video and working time optimization (based on human action detection). As the videos are recorded with a certain positioning of cameras and a certain light, the system has some limitations on how video should be filmed. Another limitation is the number of working tools (pliers, wrench, hammer, chisel). Further developments of the work might be connected with the algorithms for 3D modeling, modeling the activity as a sequence of frames (RNN, LSTM models), Action Detection model development, time optimization for the working process, recommendation system for working process from video activity detection.

Keywords: neural networks; computer vision; pose estimation; computer vision; machine learning; object overlaying; object detection; work optimization.

For citation: Shtekhin S.E., Karachev D.K., Ivanova Yu.A. Computer vision system for working time estimation by human activities detection in video frames. Trudy ISP RAN/Proc. ISP RAS, vol. 32, issue 1, 2020. pp. 121-136 (in Russian). DOI: 10.15514/ISPRAS-2020-32(1)-7

Acknowledgements: The authors are graceful to Russian Railway Company.

1. Введение

Для изучения и распространения передовых методов труда, затрат рабочего времени и установления нормативных величин, а также при исследовании трудовых процессов с быстрыми движениями и малыми отрезками времени, которые трудно или невозможно охватить методом визуальных наблюдений, используется видеонаблюдение [1].

В среднем, в год по всей сети железных дорог инженерами по организации и нормированию труда структурных предприятий функциональных филиалов ОАО «РЖД» пересматривается или разрабатывается вновь порядка 837 производственных процессов (порядка 18-22 нормативных сборников норм времени), на которые должна быть проведена видеосъемка. Видеосъемка рабочего времени является методом исследования производственных процессов, трудовых операций и фактических затрат рабочего времени. Этот метод не только обеспечивает высокую точность измерения всех фактических затрат рабочего времени, любых трудовых операций, движений, действий, но и позволяет фиксировать и демонстрировать их содержание.

Отличие видеосъемки от классических методов исследования рабочего времени заключается в том, что процесс замеров времени и анализ полученных результатов отделены друг от друга. Рабочая операция анализируется после съемки. С помощью видео можно наблюдать не только за ходом выполнения рабочего задания, но и за

используемыми средствами производства и материалами, приемами труда, рабочим местом. Результаты видеосъемки служат основой для проектирования рациональных трудовых процессов, нормативов на подготовительно-заключительные действия, обслуживание рабочего места, регламентированных перерывов для отдыха и питания, уточнения (проверки) или разработки норм времени, получения данных для проведения специальной оценки условий труда. Видеосъемка позволяет проводить обучение работников предприятий железнодорожной отрасли передовым приемам и методам труда.

Прогнозируется в среднем в год более 100 тысяч видеосъемок, с учетом проведения 3-х замеров по каждому производственному процессу со всех железных дорог и с учетом того, что процессом видеосъемки будет охвачено только 25-30% всех работ.

В связи с таким большим количеством видеосъемок в год, является актуальной задача автоматической разметки и анализа видео средствами машинного обучения (компьютерного зрения). На первом этапе были выделены технологические операции, в которых производится работа сотрудника с инструментами. Для таких технологических операций задача была поставлена следующим образом: «Произвести автоматическую разметку видеосъемки, определяя алгоритмами компьютерного зрения кадры, на которых сотрудник, выполняющий определенные работы, держит конкретные виды инструментов в руках».

Задача настоящего исследования относится к задаче распознавания движений человека (Human Action Detection). Задача решалась на данных по железнодорожной тематике, которая заключается в определении события, происходящего с человеком на обрабатываемом кадре (изображении). Всего использовалось 10 видео длительностью около 10 минут, 25 кадров на каждую секунду времени. Размер видео 1920 px & 1040 px.

2. Распознавание в видео кадре людей и определение их координат

Для исследования алгоритмов детекции людей проведено тестирование Object Detection на датасете COCO [2]. В основном, модели детекции людей показывают высокую точность (более 80%). Наибольшую точность показывает RetinaNet (0.99).

Первым этапом в рамках данной задачи является определение наличия сотрудника в видео кадре и его координаты. В дальнейшем, если сотрудник определен в видео кадре, то данный кадр передается на следующие этапы алгоритма, если сотрудника в кадре нет, то кадр дальше не обрабатывается. Это условие показывает важность точного определения наличия человека в кадре для следующего этапа – поиск ключевых точек сотрудника.

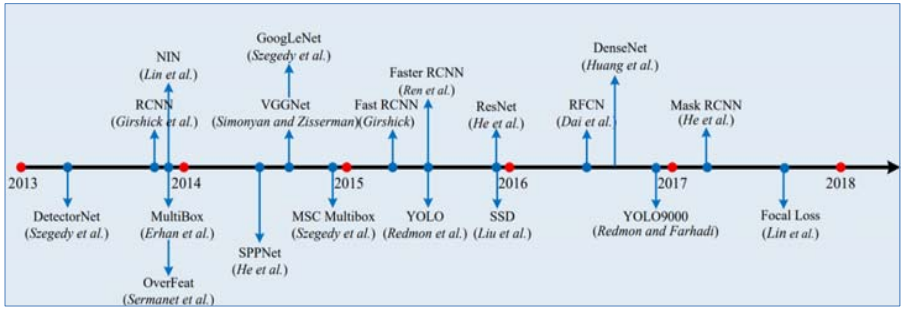


Рис. 1. Основные вехи развития моделей сверточных нейронных сетей, решающих задачу распознавания объектов [3]

Fig. 1. The main milestones of the development of convolutional neural network models that solve the problem of object recognition [3]

Данная задача относится к следующему классу задач компьютерного зрения – распознавание объектов (object detection). **Object detection** – обнаружение всех объектов указанных классов и определение охватывающей рамки (bounding box) для каждого из них. Современные методы решения задач данного класса представляют собой глубокие нейронные сети (Deep Learning). Основные вехи развития моделей сверточных нейронных сетей, решающих задачу распознавания объектов показаны на рис. 1.

Табл. 1. Сравнение современных сверточных нейросетей для методов Object Detection на COCO датасете [12]. Были проведены измерения при различных коэффициентах Жаккара (AP50 — при коэффициенте 50%, AP75 - при коэффициенте 75%, APS — для объектов площадью менее 32 квадратных пикселя, APM — для объектов площадью более 32, но менее 96 квадратных пикселя, APL — для объектов большей площади)

Table 1. Comparison of modern convolutional neural networks for Object Detection methods on COCO dataset [12]. Measurements were taken at various Jacquard coefficients (AP50 - at a coefficient of 50%, AP75 - at a coefficient of 75%, APS - for objects with an area of less than 32 square pixels, APM - for objects with an area of more than 32 but less than 96 square pixels, APL - for objects larger area)

	backbone	*AP*	*AP*$_{50}$	*AP*$_{75}$	*AP*$_S$	*AP*$_M$	*AP*$_L$
Two-stage methods							
Faster R-CNN+++ [4]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FIN [5]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [6]	Inception-ResNet-v2 [14]	34.7	55,5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [7]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [8]	DarkNet-19 [8]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [9-10]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD523 [10]	ResNet-101-SSD DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [11]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [11]	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608x608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9
COCO for YO:Ov3							

Были рассмотрены двухуровневые (Faster R-CNN+++ [4], Faster R-CNN w FPN [5], Faster R-CNN by G-RMI [6], Faster R-CNN w TDM [7] и одноуровневые модели сверточных нейронных сетей (YOLOv2 [8], SSD513 [9-10], DSSD513 [10], RetinaNet [11], YOLOv3 [12]). Результаты сравнения работы этих моделей показаны на датасете COCO (табл. 1). В качестве метрики для измерения точности использовалась Mean Average Precision –

средняя точность детекции всех объектов [13], которая вычисляется как среднее значение детекций для всех классов объектов. Для каждого класса вычисляется точность предсказания (Average Precision – AP). Данная метрика связана с подсчетом коэффициента Жаккара (intersection over union), где для найденного объекта подсчитывается площадь совпадающей ограничивающей рамки.

Из результатов исследования видно, что максимальная точность показана моделью RetinaNet.

Был подготовлен датасет из 1000 кадров из видеофайлов, предоставленных РЖД, из которых 500 изображений с человеком и 500 изображений без человека (в основном, эти изображения содержат железную дорогу, рельсы и поезд). На данном датасете были исследованы вышеупомянутые модели. В результате данного исследования была выбрана модель RetinaNet для определения людей на кадре. В табл. 2 показаны результаты метрик для отобранной модели при решении классификационной задачи нахождения людей.

Среднее время обработки моделью RetinaNet одного кадра 0,076 с, что также в среднем быстрее чем другие модели.

В датасете COCO существует категория «person» (человек), поэтому предобученная модель подходит для задачи нахождения человека в кадре. Модель была реализована с помощью библиотек Keras и Tensorflow (python3.6).

Результатом данного этапа является json-файлы для каждого кадра видео, в которых имеется информация о наличии или отсутствии людей в кадре и координаты ограничивающей их рамки при их присутствии.

Табл. 2. Результаты предсказания модели RetinaNet на датасете из 1000 кадров. Класс 0 – изображение не содержит в себе человека. Класс 1 – изображение содержит в себе человека. Всего было отобрано 1000 изображений, 500 из них – с человеком, 500 – без. Support – количество людей в каждом классе. Precision, recall, f1-score — точность, полнота, f1-мера для каждого из классов.
Tab. 2. Results of the RetinaNet model predicting on a dataset of 1000 frames. Class 0 – the image does not contain a person. Class 1 – the image contains a person. In total, 1000 images were selected, 500 of them with a person, 500 without. Support – the number of people in each class.

	Precision	Recall	F1-score	Support
0	0.96	0.99	098	500
1	0.99	0.96	0.98	500
accuracy				1000
macro avg	0.98	0.98	0.98	1000
weighted avg	0.98	098	0.98	1000

3. Определение в видеокадре ключевых точек людей

Следующим этапом решения задачи является определение координат рук, локтей и плеч сотрудников.

Эта задача относится к следующему классу задач машинного зрения: определение ключевых параметрических точек человека и оценка позы человека по изображению (Pose estimation).

В рамках исследования были рассмотрены следующие модели, обученные на COCO датасете: PAFs [15], OpenVino [16], CPN [17], Simple [18]. Также был исследован метод постобработки PoseFix [19] на всех моделях.

Модели исследовались на датасете из 500 кадров, содержащих изображения людей, на которых вручную были размечены ключевые точки (всего размечено 6 точек для каждого человека, по две точки на кисти, локти, плечи). Модели сравнивались по двум параметрам:

точность и время обработки. Точность оценивалась по метрике Object Keypoint Similarity (OKS) [2].

Рис 2. Сравнение моделей Pose Estimation по метрике OKS (средняя точность похожести точек) и времени обработки кадра. Модель tf_simplehuman_pose_256x192 выбрана в качестве оптимальной по детекции и по времени по сравнению с другими моделями
Fig 2. Comparison of Pose Estimation models by the OKS metric (average accuracy of similarity of points) and frame processing time. The tf_simplehuman_pose_256x192 model was selected as optimal in detection and in time compared to other models

Результаты сравнения моделей представлены на рис. 2, названия моделей показаны как tf_pose_*, openvino_*, tf_cpn_pose* и tf_simplehumanpose_* с различными входными данными соответственно (цифры после названия означают размер исходного изображения). По результатам сравнения была выбрана tf_simplehuman_pose_256_192, которая показала время обработки одного кадра 0,11с и среднюю метрику OKS = 0,82.

Модель была реализована с помощью библиотеки Tensorflow. На вход в данную модель подается кадр видео и координаты сотрудника, полученные на предыдущем этапе.

Результатом данного алгоритма является json-файлы, в которых хранятся координаты 6 ключевых точек (или меньше, если алгоритм не нашел всех точек) – плеча, локтя и кисти (левой и правой руки соответственно) для каждого кадра видео, на которых есть человек.

4. Распознавание в видеокадре инструментов и определение координат их ограничивающей рамки (Object Detection)

Следующим этапом решения задачи является определение координат инструментов. Данная задача относится к проблеме распознавания объектов (object detection). Для её решения выбрана модель RetinaNet (она же была использована на первом этапе). Так как в датасете COCO нет категорий инструментов, используемых в РЖД, для дообучения нейросети была необходима подготовка датасета с изображениями, на которых были размечены инструменты (пример разметки инструментов показан на рис. 3).

Рис. 3. Исходный кадр видео с отмеченными инструментами. Каждый инструмент в кадре помечен соответствующими координатами ограничивающей его рамки
Fig. 3. The original frame of the video with the marked tools. Each tool in the frame is marked with the corresponding coordinates of the bounding box

Первоначально из четырёх видео были получены кадры, в которых содержатся инструменты (часть кадров показана на рис. 4). Видео были отобраны таким образом, что человек может различить инструмент (хорошее освещение, и съемка не далее, чем с расстояния 7 метров).

Рис. 4. Датасет №1, собранный из 4-ех видео. Примеры некоторых кадров
Fig. 4. Dataset number 1, assembled from 4 ex video. Examples of some frames

Был создан первый тестовый датасет для обучения: в датасете №1 содержатся все инструменты каждого кадра без аугментации, всего 25198 изображений молотков, 21805 изображений инструмента зубило, 133663 изображений инструмента ключ, 48551 изображений инструмента плоскогубцы. В качестве тестовых данных было использовано одно из видео «Rakurs4», которое содержит 10450 изображений инструмента зубило, 9461 изображений инструмента молоток, 10179 изображений инструмента плоскогубцы, 16798 изображений инструмента ключ.

Была обучена нейросеть RetinaNet [8] в течение 20 эпох (batch size=4). Нейросеть реализована с помощью библиотеки Tensorflow.

В результате дообучения нейросети были получены результаты точности детекции, представленные в табл. 3.

Табл. 3. Средняя точность детекции инструментов на тестовой выборке, созданной из видео «Rakurs4», нейросетью, обученной на датасете 1
Tab. 3. The average accuracy of detection of instruments on a test sample created from the video "Rakurs4", a neural network trained on dataset 1

	Средняя точность детекции объектов (инструментов)				
	Зубило	Молоток	Плоскогубцы	Ключ	Все инструменты
Датасэт 1	0.9788	0.1245	0.8034	0.5412	0.6120

5. Эксперименты с аугментацией

Было произведено несколько экспериментов для улучшения детекции инструментов нейросетью. Исходные данные (картинки с объектами инструментов: зубило, плоскогубцы, молоток, ключ, полученные из датасета №1) были увеличены с помощью различных аугментаций, и полученные изображения использовались в качестве тренировочных данных для обучения нейросети RetinaNet [9]. Исследователи доказали, что увеличение изображений в тренировочном датасете с помощью различных аугментаций приводит к увеличению качества предсказания нейросети [20-21], проводимые эксперименты также показали улучшение детекции объектов на тестовых данных.

Датасет №1

Описан в предыдущем разделе (раздел 4).

Датасет №2

Было увеличено количество объектов с поворотами: все исходные изображения (датасет №1) были повернуты произвольно на угол от -30 до 30 градусов. 20 эпох обучения.

Датасет №3

К исходному датасету (датасету №1) было применено масштабирование в непропорциональном соотношении (случайно от 0 до 30% от исходной длины и ширины изображения).

Датасет №4

Исходные изображения (датасет №1) были отражены по горизонтали.

Датасет №5

Объединены датасеты №1, №2, №3 и №4 и создан датасет №5.

Датасеты №6, №7, №8

Изображения из датасетов №2, №3 и №4 были перемешаны в случайном порядке

Датасет №9

Решено применить масштабирования в большем интервале совместно со смещениями (в пределах 20% по горизонтали и по вертикали),

Датасет №10

В десятом эксперименте к исходным данным (датасет №1) были применены повороты (случайное значение от -30 до 30 градусов), масштабирование (пропорционально по длине и ширине в случайном значении от 0.2 до 1.2) и смещение (случайное значение в пределах 20% относительно длины и ширины).

Датасет №11

Так как количество элементов в каждом классе инструмента неуравновешенно (несбалансированные данные), была выдвинута гипотеза, что применение аугментации для выравнивания количества классов может привести к улучшению качества детекции. Была удалена часть инструментов (ключ) и добавлен за счет аугментаций инструмент

молоток. После балансировки количество изображений для тренировки нейросети увеличилось и пропорции классов выровнялись (рис. 5).

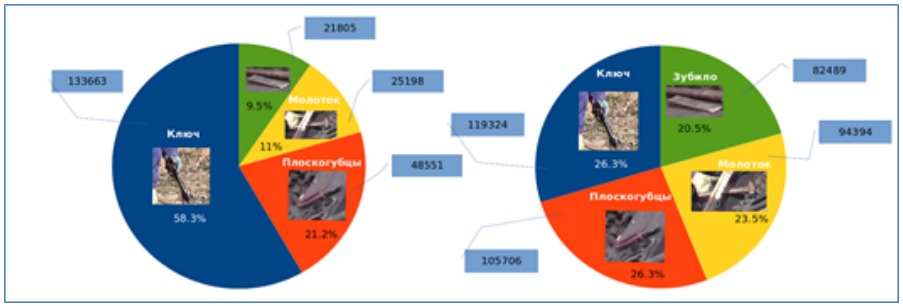


Рис. 5. Первоначальная пропорция количества объектов в датасете №1 (слева) и аугментированного датасета №11 (справа)
Fig. 5. The initial proportion of the number of objects in dataset number 1 (left) and augmented dataset number 11 (right)

Датасет №12

Было увеличено количество ключей (за счет добавления тех изображений, которые были удалены в датасете №11).

Датасет №13

Был составлен датасет №13 из 10 размеченных видео с аугментацией (датасет №1 – часть этих 10 видео).

Датасет №14

Для улучшения точности детекции нейросети было решено использовать дополнительные изображения из открытых источников. Был найден датасет KTH Handool Dataset [22-23], который представляет собой изображения для трех инструментов – молоток, плоскогубцы, отвертка – на разном фоне, под различным светом и сделанными под разными положениями камеры. Всего в датасете KTH Handtool 4500 изображений, по 1500 изображений на каждый объект. Для каждого изображения имеется xml-файл, где указано название объекта и краевые точки прямоугольника, в котором находится объект в изображении.

Датасет №14 состоит из изображений, полученных из датасета №13 и датасета KTH Handool Dataset. Так как некоторые инструменты из датасета KTH Handtool похожи на инструменты, которые детектируются в видео РЖД, была выдвинута гипотеза, что при добавлении дополнительных данных будет улучшение детекции инструментов.

На рис. 6 приводятся изображения плоскогубцев, используемых в компании РЖД и представленных в данных KTH Handtool.

На основе результата тренировки нейросети сделан вывод, что добавление дополнительных изображений для тренировки нейросети может привести к улучшению детекции, если инструменты добавляемого датасета схожи с инструментами РЖД.

Рис. 6. Плоскогубцы. Сверху – инструменты из датасета KTH Tool, внизу – инструменты, полученные из видео РЖД. Можно увидеть сходство инструментов
Fig. 6. Pliers. At the top are the tools from the KTH Tool dataset, at the bottom are the tools obtained from the Russian Railways video. You can see the similarity of tools

Все результаты приведены в табл. 4. По результатам проведенных экспериментов можно сделать следующие выводы. Использование аугментации при обучении нейросети RetinaNet улучшает детекцию объектов, но необходимо также делать балансировку объектов в каждом классе. Самыми эффективными преобразованиями являются повороты в случайном порядке от 0 до 360 градусов и масштабирование в интервале от 20% до 120% от исходного размера.

Табл. 4. Средняя точность детекции инструментов нейросетью на тестовой выборке, полученной из видео «Rakurs4»
Table 4. The average accuracy of detection of instruments by a neural network in a test sample obtained from the video «Rakurs4»

	Средняя точность детекции объектов				
	Зубило	**Молоток**	**Плоскогубцы**	**Ключ**	**Все инструменты**
Датасет №1	0.9788	0.1245	0.8034	0.5412	0.6120
Датасет №2	0.9765	0.0918	0.8880	0.6169	0.6433
Датасет №3	0.9725	0.1679	0.9461	0.6140	0.6751
Датасет №4	0.9163	0.3747	0.9296	0.5659	0.6966
Датасет №5	0.9062	0.1106	0.9554	0.5944	0.6417
Датасет №6	0.9978	0.3590	0.9475	0.6951	0.7498
Датасет №7	0.9961	0.4379	0.9314	0.7626	0.7820
Датасет №8	0.9532	0.0333	0.6913	0.5339	0.5529
Датасет №9	0.9987	0.4714	0.7768	0.7509	0.7495
Датасет №10	0.9992	0.4644	0.9471	0.6028	0.7534
Датасет №11	0.9981	0.4899	0.9621	0.4471	0.7200
Датасет №12, 24 эпохи	0.9963	0.4525	0.9491	0.5983	0.7490
Датасет №12, 40 эпох	0.9861	0.4380	0.9594	0.6252	0.7522
Датасет №13	0.9184	0.0618	0.8011	0.7642	0.6364
Датасет №14	0.9414	0.0618	0.9612	0.8280	0.6998

Можно заметить, что датасет №13 и датасет №14 дают низкие показатели детектирования для инструмента молоток, это связано с несбалансированностью данных датасетов. Датасет KTH Handtool Dataset имеет потенциал для улучшения алгоритма детекции, так как детекция плоскогубцев улучшилась на 16% при добавлении данного датасета в обучающую выборку (датасеты №13 и №14 для плоскогубцев в табл. 4).

Результатом данного этапа является json-файл, в котором хранится список распознанных инструментов с их координатами для каждого кадра видео.

6. Определение инструментов, которые находятся в руках у человека

Последним этапом в рамках данной задачи является определение инструментов, которые находятся в руках человека. Для этого используются все объединенные результаты предыдущих алгоритмов для каждого кадра видео. Сложность обусловлена тем, что на изображении находятся несколько различных инструментов, могут присутствовать несколько людей, и при расчёте расстояний до объектов необходимо учитывать их масштаб и возможные комбинации инструмент-человек. Кроме того, нельзя не учитывать тот факт, что с некоторых ракурсов инструменты могут не определяться, так как могут быть перекрыты другими объектами (рельсами, самим человеком и т.д.).

Метод заключается в построении модели, которая предскажет, находится ли данный инструмент в руках у человека или нет. Каждый инструмент проверяется для каждого человека; таким образом, алгоритм работает даже в случае, когда в кадре присутствует большое количество инструментов и людей, что говорит о гибкости данного подхода.

Данный этап является интеграционным, так как производится расчёт на основе результатов предсказаний моделей на предыдущих этапах. Предсказания на данном этапе происходят на основе подсчета расстояний между точками:

– координаты точки кисти;
– координаты центра инструментов.

Однако для обучения модели этих данных недостаточно, так как видео имеет различный масштаб и, соответственно, расстояние на различных видео между точками для находящегося в руке инструмента будет сильно отличаться. Для решения данной проблемы были созданы дополнительные признаки, такие как отношение размера инструмента к расстоянию между точками (от плеча до локтя, от локтя до кисти) каждой руки, а также квадраты и логарифмы этих значений.

При обучении модели важную роль играет баланс классов. Сгенерированные тренировочные данные с видео показаны в табл. 5.

Табл. 5. Сгенерированные тренировочные данные для обучения классификационной модели детектирования инструмента в руке человека (класс 1) или не в руке (класс 0)
Tab. 5. Generated data for training the classification model of detecting an instrument in a person's hand (class 1) or not in a hand (class 0)

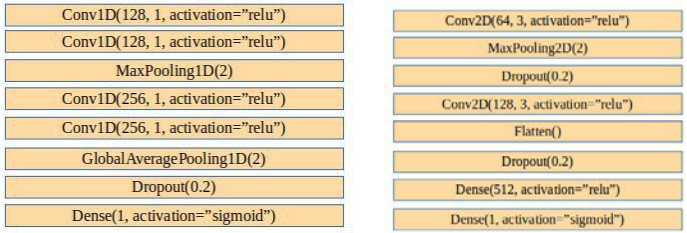
Инструмент	В руке (1)	Не в руке (0)
Плоскогубцы	2892	24221
Молоток	862	10752
Гаечный ключ	24568	40621

Исходя из таблицы 5, можно сделать вывод о том, что нужна балансировка классов. Сделать ее можно двумя способами:

– удалить лишние нули и привести количество всех классов к одному;

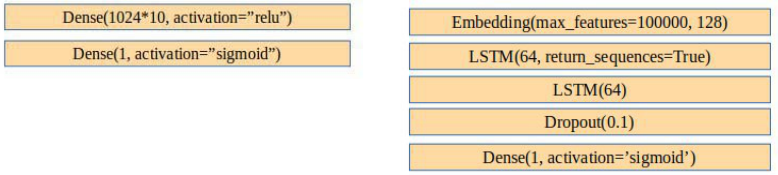
– случайным образом продублировать значения классов, примеров которых недостаточно, до определенного количества и удалить лишние примеры классов, число значений которых велико.

В процессе исследования были оценены различные типы нейронных сетей (рис. 7 и 8): полносвязные, свёрточные (с 1D и 2D свёртками), сети с кратковременной памятью и LSTM (Long short-term memory). Кроме того, дополнительно был протестирован подход с дообучением для сети ResNet50 [24]. Результат предсказания получается в результате сравнения ответа модели с порогом (если значение ниже порога, значит инструмент в руках, иначе – не в руках), который подбирается эмпирически.



a)1D свертка b) Нейросеть с 2D свертками

Рис. 7. Архитектура моделей с 1D свертками (a) и 2D свертками (b)
Fig. 7. Architecture of models with 1D convolution (a) and 2D convolution (b)



a) Полносвязная нейросеть b) Нейросеть с LSTM слоями

Рис. 8. Архитектура модели полносвязной нейросети (a) и нейросети с LSTM слоями (b)
Fig. 8. Architecture of a fully connected neural network model (a) and a network with LSTM layers (b)

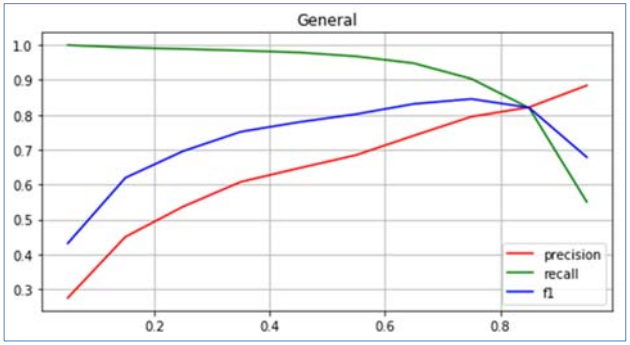


Рис. 9. Предсказания сети с 1D свёртками на тестовых данных. Показаны зависимости точности (precision), чувствительности (recall) и их гармоническое среднее (f1) от порога
Fig. 9. Predictions of a network with 1D convolutions on test data. Shown are the dependences of accuracy (sensitivity), sensitivity (recall) and their harmonic mean (f1) on the threshold

В результате сравнения, наилучшим образом себя показала модель с 1D свёртками [25], архитектура которой представлена на рис. 7; на рис. 9 представлены показаны ее точность, чувствительность и среднее гармоническое в зависимости от порога. Основные параметры для тренировки сети: оптимизатор – Adadelta [26], размер батча (выборки) – 500, количество эпох – от 5 до 10.

Для всех типов нейронных сетей, кроме полносвязной, данные группировались таким образом, что получалась одна таблица (фрейм) для предсказания, в которой каждая строчка – это информация об инструменте в определенный момент времени (каждый момент времени представлен кадром видео, в 1 секунде 45 кадров видео), таким образом, учитывается временная составляющая.

В табл. 6 представлены результаты при наивысшем f1 – score, при пороге равном 0.7, выше которого считается, что объект найден.

Табл. 6. Общие результаты лучшей модели с 1D свёртками
Tab. 6. General results of the best model with 1D convolution

Класс	**Точность**	**Полнота**	**f1-score**
0	0.97	0.94	0.96
1	0.79	0.90	0.85

Табл. 7. Агрегирующая таблица результатов для классов 1
Tab. 7. Aggregate results table for grades 1

Модель	**Точность**	**Полнота**	**f1-score**
Нейросеть с 1D свертками [27]	0.79	0.90	0.85
Нейросеть с 2D свертками [27]	0.66	0.95	0.78
Полносвязная нейросеть [28]	0.25	0.81	0.39
Нейросеть с LSTM слоями [29]	0.25	0.81	0.39

Результатом данного этапа и всей задачи в целом является json-файлы – список сотрудников с их координатами и для каждого сотрудника список инструментов, которые находятся у них в руках для каждого кадра видео (если модель задетектировала инструмент) с координатами.

7. Заключение

Описанные алгоритмы показали высокую точность детекции инструментов в руках человека, благодаря чему могут быть использованы для дальнейшей разработки. Задача детектирования действий человека по видео является не новой для науки, но методы, предложенные в данной статье, опираются на новейшие разработки в области компьютерного зрения, с применением методов, не использованных ранее. Данное исследование может быть использовано для оптимизации труда (определение норм времени по видео), а также в качестве рекомендательных показаний сотрудникам РЖД о том, сколько времени необходимо на выполнение операции, и оптимизации рабочего труда по видео.

Предполагается вести дальнейшие исследования по следующим направлениям.

1. Детектирование людей и их действий по видео:
   1.1. настройка и постобработка RetinaNet.
2. Определение ключевых точек:
   2.1. модель 3D pose estimation;
3. Определение инструментов:
   3.1. дообучение моделей с разными ракурсами;
   3.2. исследование различных аугментаций;
   3.3. детектирование новых инструментов.
4. Определение человека с инструментом:
   4.1. исследование различных атрибутов для классификационной модели;
   4.2. исследование моделей, работающих с последовательностями кадров (RNN, LSTM) для детектирования действий человека по видео.
5. Исследование моделей Action Detection.
6. Оптимизация рабочего труда:
   6.1. отбор видео, в которых сотрудник затрачивает минимальное количество времени на выполнение операции;
   6.2. дальнейшее использование данного видео в качестве рекомендации другим сотрудникам для улучшения производительности.

Список литературы / References

[1]. Методические рекомендации по изучению затрат рабочего времени в структурных подразделениях ОАО «РЖД». Утверждены распоряжением ОАО «РЖД» от 10 апреля 2018 / Guidelines for the study of the costs of working time in structural divisions of JSC Russian Railways. Approved by the order of Russian Railways on April 10, 2018

[2]. Keypoint evaluation metrics used by COCO. Available at: http://cocodataset.org/#keypoints-eval, accessed 05.01.2020.

[3]. Andrea Gaetano Tramontano. Deep Learning Networks for Real-time Object Detection on Mobile Devices. Master's Degree Thesis, University of Padova, Italy, 2018/2019.

[4]. C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. arXiv:1701.06659, 2017.

[5]. J. Huang, V. Rathod et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3296-3297.

[6]. D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi. Iqa: Visual question answering in interactive environments. arXiv:1712.03316, 2017.

[7]. O. Russakovsky, L.-J. Li, and L. Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2121–2131.

[8]. J. Parham, J. Crall, C. Stewart, T. Berger-Wolf, and D. Rubenstein. Animal population censusing at scale with citizen science and photographic identification. In Proc. of the AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good, 2017, pp. 37-44.

[9]. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. ´ Focal loss for dense object detection. arXiv:1708.02002, 2017.

[10]. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, vol. 88, no. 2, 2010, pp. 303– 338.

[11]. I. Krasin, T. Duerig et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification, 2017. Available at: https://github.com/openimages, accessed 05.01.2020.

[12]. Joseph Redmon, Ali Farhadi: YOLOv3: An Incremental Improvement. arXiv:1804.02767, 2018

[13]. Jonathan Hui. mAP (mean Average Precision) for Object Detection. Available at: https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173, accessed 05.01.2020.

[14]. M. Scott. Smart camera gimbal bot scanlime:027, Dec 2017. 4

[15]. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part affinity fields. CVPR, 2017.
[16]. D. Osokin, Real-time 2d multi-person pose estimation on CPU: Lightweight OpenPose arXiv:1811.12004
[17]. Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103-7112.
[18]. Xiao, Bin, Haiping Wu, and Yichen Wei. Simple Baselines for Human Pose Estimation and Tracking. Lecture Notes in Computer Science, vol. 11210, 2018, pp. 472-487.
[19]. G. Moon, J.Y. Chang and K.M. Lee. PoseFix: Model-Agnostic General Human Pose Refinement Network. In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7765-7773.
[20]. Arun Gandhi. Data Augmentation. How to use Deep Learning when you have Limited Data – Part 2. Available at: https://nanonets.com/blog/data-augmentation-how-to-use-deep-learning-when-you-have-limited-data-part-2/, accessed 05.01.2020.
[21]. Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, vol. 60, no. 6, 2017, pp. 84-90.
[22]. Massimiliano Mancini, Hakan Karaoguz, Elisa Ricci, Patric Jensfelt, Barbara Caputo. Kitting in the Wild through Online Domain Adaptation. arXiv:1807.01028, 2018.
[23]. Hakan Karaoguz, Patric Jensfelt. Fusing Saliency Maps with Region Proposals for Unsupervised Object Localization, arXiv:1804.03905, 2018.
[24]. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv:1506.01497, 2015.
[25]. Kiranyaz S., Ince T. & Gabbouj M. Real-Time Patient-Specific ECG Classification by 1D Convolutional Neural Networks. IEEE Transactions on Biomedical Engineering, vol. 63, issue 3, 2016, pp.664–675.
[26]. M.D. Zeiler. ADADELTA: an adaptive learning rate method. arXiv:1212.5701, 2012.
[27]. Zha, Xuefan. (2018). A Comparison of 1-D and 2-D Deep Convolutional Neural Networks in ECG Classification. arXiv:1810.07088, 2018.
[28]. G. Huang, Z. Liu, and K.Q. Weinberger. Densely connected convolutional networks. arXiv:1608.06993, 2017..
[29]. A. Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. arXiv:1808.03314, 2018.

Информация об авторах / Information about authors

Сергей Евгеньевич ШТЕХИН – старший специалист по анализу данных. Сфера научных интересов: компьютерное зрение.

Sergey Evgenievich SHTEKHIN – Senior Data Analyst. Research interests: computer vision.

Юстина Алексеевна ИВАНОВА – специалист по анализу данных. Сфера научных интересов: компьютерное зрение, временные ряды, рекомендательные системы.

Justina Alekseevna IVANOVA – Data Analyst. Research interests: computer vision, time series, recommendation systems.

Денис Константинович КАРАЧЕВ – специалист по анализу данных. Сфера научных интересов: компьютерное зрение, беспилотный транспорт.

Denis Konstantinovich KARACHEV – Data Analyst. Research interests: computer vision, unmanned vehicles.