УДК 004.891.3

ИНСТРУМЕНТ ДЛЯ РАСПОЗНАВАНИЯ ЯЗЫКА ЖЕСТОВ ИЗ ВИДЕОПОТОКА В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

И. И. Закирьянов[1] [0009-0009-3448-4844], И. З. Хаялеева[2] [0009-0007-5837-7010], И. А. Валишин[3] [0009-0006-6891-031X], Е. Д. Курито[4] [0009-0000-6214-135X], А. Н. Фасхутдинов[5] [0009-0001-2766-4048]

[1–5] *Институт информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, ул. Кремлевская, 35, г. Казань, 420008*

[1]zakiryanov.iskander@mail.ru, [2]izidakh@yandex.ru, [3]iskander1998@list.ru, [4]ekurito@gmail.com, [5]azatazat835@mail.ru

Аннотация

Разработан инструмент, распознающий из видеопотока слова или отдельные буквы в режиме реального времени. Рассмотрены возможности и перспективы его применения в современном обществе. Приведены результаты экспериментов по проверке работоспособности этого инструмента на примере английских слов и латинских букв.

Ключевые слова*: распознавание жестов, нейронные сети, компьютерное зрение, YOLO*

ВВЕДЕНИЕ

В современном мире, где коммуникация играет ключевую роль во множестве сфер, разработка инструментов для распознавания языка жестов становится все более актуальной и необходимой. Это подтверждается статистикой по инвалидности до 2022 года в России, согласно которой можно сделать вывод, что в России на сегодняшний день проживает около 16 тысяч человек, страдающих болезнями уха и сосцевидного отростка [1].

Основная цель проведенного исследования – разработать инструмент распознавания жестов из видеопотока в режиме реального времени. Для достижения этой цели были поставлены следующие задачи:

1. Провести анализ существующих решений;
2. Выбрать наиболее подходящую модель;

3. Создать или найти наборы данных;
4. Провести тестирование в реальном времени.

ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

В настоящее время существует ряд инструментов, частично решающих задачу, поставленную выше. В ходе исследования были обнаружены и изучены инструменты, позволяющие распознавать язык жестов.

Одним из самых популярных таких инструментов является MediaPipe [2], позволяющий распознавать как статические, так и динамические жесты руки. Для более подробного анализа была также рассмотрена статья [3], в которой авторы подробно рассматривают структуру руки, ее ключевые точки и методы распознавания статических и динамических жестов на основе этих точек. Названная работа демонстрирует успешные возможности распознавания различных жестов и движений. Основными достоинствами подхода, основанного на MediaPipe, являются его простота, быстродействие и точность распознавания. Однако есть и ограничения: ограниченный набор распознаваемых жестов и необходимость модификации кода для распознавания новых жестов.

Отметим, что распознавание жестов является важной задачей в робототехнике, потому что уже появилась потребность разрабатывать новые и более естественные подходы к взаимодействию человека и машины. В упомянутой статье предложена модель для распознавания жестов рук в режиме реального времени. Эта модель принимает в качестве входных данных электромиографические (ЭМГ) сигналы, измеряемые на предплечье с использованием коммерческого датчика Myo Armband. Использованы также автоэнкодер для автоматического извлечения признаков и искусственная нейронная сеть с прямой связью для классификации жестов. Названная модель может распознавать те же 5 жестов, что и система распознавания вышеупомянутого датчика, достигая средней точности распознавания 85,08% ± 14,92% при среднем времени отклика 3 ± 1 мс. Предлагаемая нами модель является общей – это подразумевает, что она может распознавать жесты любого пользователя, даже если его данные не включены в обучающий набор данных. Однако существенным недостатком является то, что для определения

жестов в этой модели на вход необходимы электромиографические сигналы, измеряемые датчиком, что не всегда удобно, а также количество жестов является ограниченным [4].

АРХИТЕКТУРА РАЗРАБАТЫВАЕМОГО ИНСТРУМЕНТА

Для решения задачи распознавания языка жестов было принято решение использовать модель YOLO [5], так как она является одной из самых популярных для задачи распознавания объектов в режиме реального времени [6]. Для определения оптимальной версии модели были протестированы различные версии (5-я и 8-я). Данные версии были выбраны, поскольку в версии 5 произошли значительные изменения в архитектуре моделей [7], а версия 8 представляет собой наиболее актуальное обновление.

Dataset

Обучение модели YOLO требует наличие данных в формате «изображение–координаты детектируемого объекта на изображении». Для этого был использован предварительно размеченный набор данных [8]. Количественный объем датасета составляет 1728 изображений. Размерность тестовой и обучающей выборки для подачи входных данных в модель составляет 416x416 пикселей. Определить гендерную и возрастную принадлежность объектов на изображениях в наборе данных не представляется возможным.

Для повышения эффективности обучения и обобщающей способности модели данные подвергались случайной аугментации, встроенной в функцию обучения YOLO. Встроенная аугментация в YOLO — это процесс, при котором изображения подвергаются различным трансформациям и модификациям прямо в процессе обучения модели. В YOLO аугментация применяется к обучающему набору изображений на лету, прежде чем эти изображения будут переданы модели для обучения.

Для обучения модели применены разнообразные техники аугментации, включая изменение яркости (рис. 1 b), отражения (рис. 1 c), случайные повороты (рис. 1 d), масштабирование (рис. 1 e) и добавление перспективы (рис. 1 f). Примеры этих трансформаций представлены на рисунке 1 на примере из обучающей выборки. Эти трансформации выполняются случайным образом для каждого

изображения перед каждой эпохой обучения, что создает разнообразие данных для обучения.

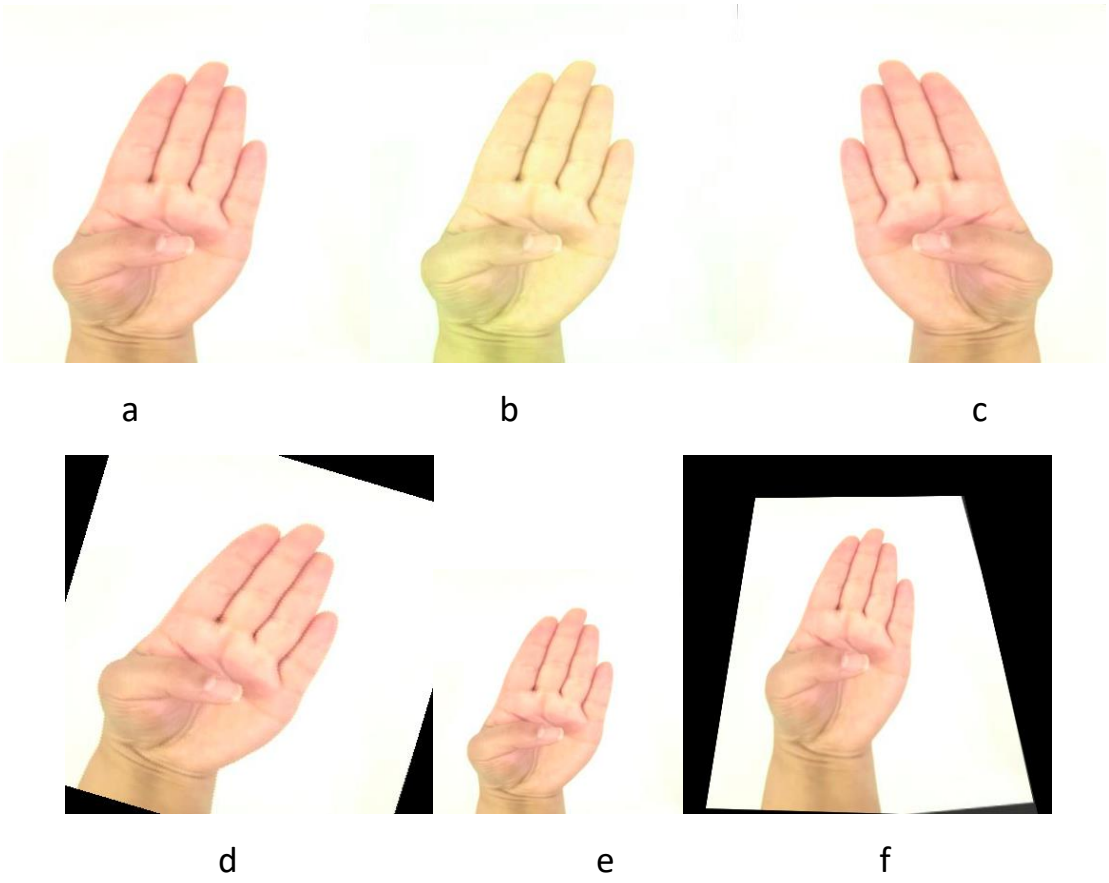

Рис. 1. Виды аугментации. a – исходное изображение, b – изменения яркости, c – отзеркаливание, d – поворот, e – масштабирование, f – добавление перспективы

РЕЗУЛЬТАТЫ

Сравнения моделей YOLO

Для определения качества работы обученной модели были проведены эксперименты, описанные ниже. Во всех экспериментах количество эпох обучения оставалось постоянным для возможности сравнения результатов. Было выбрано 100 эпох, так как большее количество неизменно приводило к переобучению. В качестве метрик для сравнения были выбраны precision, recall на обучающей выборке и mAP для тестовой выборки.

1. Выбор наилучшей версии

Были протестированы модели YOLOv.5 и YOLOv.8. Результаты обучения показаны на рис. 2.

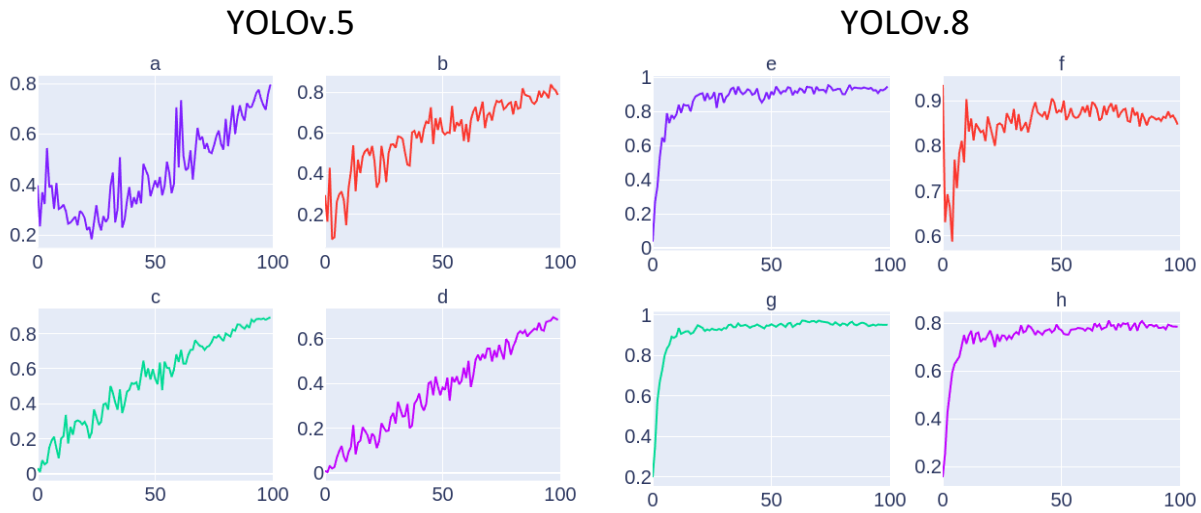

Рис. 2. a – presicion (YOLO v.5 тренировочный датасет), b - recall (YOLO v.5 тестировочный датасет), c – mAP (YOLO v.5 тренировочный датасет), d – mAP (YOLO v.5 тестировочный датасет), e – presicion (YOLO v.8 тренировочный датасет), f - recall (YOLO v.8 тестировочный датасет), g – mAP (YOLO v.8 тренировочный датасет), h – mAP (YOLO v.8 тестировочный датасет)

Изучив результаты обучения, можно увидеть, что все метрики при переходе на версию 8 улучшились. Обнаружено, что при переходе на более новую версию метрики достигают плато уже на 20–30 эпохах, в то время как в версии 5 100 эпох оказалось недостаточно для достижения плато. Кроме того, во всех метриках отмечается увеличение абсолютных значений, примерно на 10%. В дальнейших тестах использовалась YOLOv.8.

2. Влияние аугментации

Было протестировано обучение без аугментации (б.а.) и с ее использованием (с а.). Результаты обучения модели приведены на рис. 3.

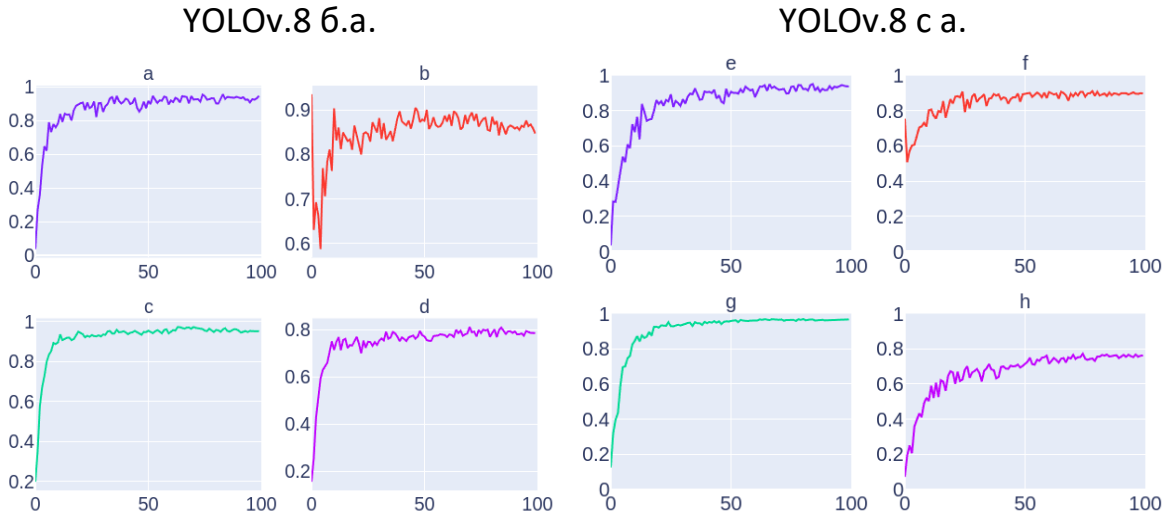

Рис. 3. a – presicion (б.а. тренировочный датасет), b – recall (б.а., тренировочный датасет), c – mAP (б.а., тестировочный датасет), d – mAP (б.а., тестировочный датасет), e – presicion (с а., тренировочный датасет), f- recall (с а., тренировочный датасет), g – mAP (с а. тестировочный датасет), h – mAP (с а. тестировочный датасет)

Проанализировав результаты, можно сделать следующие выводы. На данных без аугментации модель переобучается на тренировочных данных, что показывает стабильность метрики precision (рис. 3(a)) и уменьшение метрики recall (рис. 3(b)), тогда как на данных с аугментацией метрики стабильно увеличиваются. Похожее поведение можно наблюдать на тестовом наборе, когда на данных без аугментации модель достигает плато на 20–30 эпохах, а на данных с аугментацией виден постепенный рост.

Эксперименты

Тестирование разработанного инструмента проводилось в режиме реального времени с задержкой ≈10 мс. В экспериментах приняли участие 4 добровольца. Также было разработано программное обеспечение для запуска YOLO и демонстрации вероятности распознанной буквы в режиме реального времени в видеопотоке. Результат распознавания обозначен в виде прямоугольника с распознанной буквой и вероятностью классификации.

В ходе испытания были поставлены следующие задачи: распознание одного или несколько символов одновременно, распознание символов нескольких рук, распознание символов и сложение слова из символов.

В эксперименте участвовали четыре добровольца: трое мужчин (возраст: 25 лет – доброволец №1, 23 года – добровольцы №2 и №4) и одна женщина (возраст: 22 года – доброволец №3).

1. Тест нескольких рук вместе, показывающих одну букву

На первом этапе тестирования был опробован одновременный показ одной и той же буквы всеми добровольцами. В ходе эксперимента между руками участников соблюдалась дистанция для исключения ошибок. Также руки располагались таким образом, чтобы не перекрывать лицо участника, во избежание ошибки распознавания буквы из-за сливания с фоном. Примеры приведены на рисунках 4, 5.

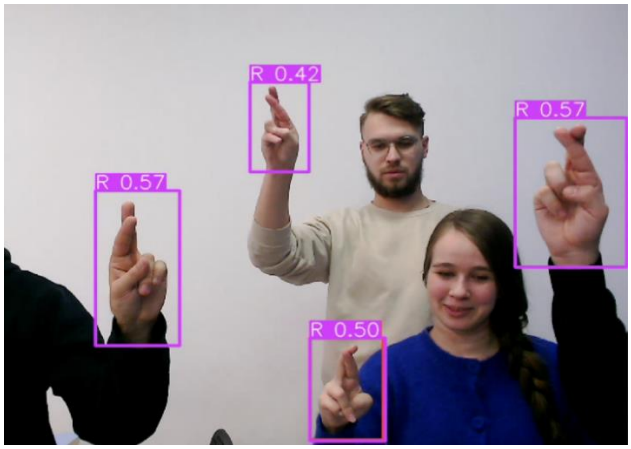

Рис. 4. Распознавание латинской буквы “R”

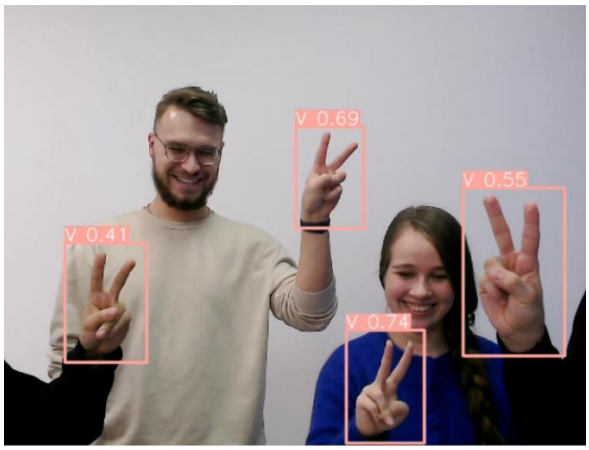

Рис. 5. Распознавание латинской буквы “V”

Результаты эксперимента сведены в таблицу 1. В ней каждой строке соответствует одна буква латинского алфавита, а каждому столбцу – объект эксперимента. Значения в последнем столбце и последней строке соответствуют среднему арифметическому по всем объектам и буквам. Жирным выделены максимальная и минимальная средняя вероятности. В случаях, когда инструмент не смог распознать показываемую букву, выставлена вероятность, равная 0.

Таблица 1. Результаты точности эксперимента с одновременным показом букв

Буква/ Доброволец	**1**	**2**	**3**	**4**	**Среднее по буквам**
A	0,28	0,39	0,56	0,48	0,44
B	0,78	0,58	0,46	0,89	0,68
C	0,77	0,26	0,29	0,68	0,50
D	0,44	0,57	0,43	0,35	0,45
E	0,00	0,54	0,27	0,60	**0,35**
F	0,67	0,58	0,26	0,71	0,56
G	0,44	0,47	0,64	0,90	0,61
H	0,88	0,39	0,37	0,91	0,64
I	0,31	0,80	0,55	0,86	0,63
J	0,57	0,27	0,32	0,76	0,48
K	0,62	0,65	0,66	0,93	0,72
L	0,90	0,59	0,28	0,33	0,53
M	0,81	0,46	0,00	0,55	0,46
N	0,44	0,70	0,38	0,00	0,38
O	0,34	0,54	0,28	0,91	0,52
P	0,88	0,49	0,56	0,29	0,56
Q	0,92	0,77	0,60	0,82	**0,78**
R	0,57	0,42	0,50	0,57	0,52
S	0,52	0,53	0,30	0,56	0,48
T	0,71	0,00	0,56	0,79	0,52
U	0,49	0,29	0,70	0,37	0,46

V	0,41	0,69	0,74	0,55	0,60
W	0,83	0,72	0,92	0,94	0,85
X	0,46	0,61	0,81	0,83	0,68
Y	0,62	0,55	0,52	0,36	0,51
Z	0,49	0,70	0,66	0,44	0,57
Средний результат	0,58	0,52	**0,49**	**0,63**	

На основе данных этой таблицы можно сделать следующие выводы: (1) инструмент справляется со своей задачей и распознает правильную букву в большинстве случаев; (2) но вероятность распознавания может варьироваться от 0,00 до 0,94 в зависимости от «правильности» расположения рук, расстояния до камеры, а также гендера; (3) принято считать, что влияние вышеперечисленных факторов основано на данных обучающей выборки, и для улучшения результатов необходимо дополнение анализируемых данных.

Влияние гендера можно увидеть, анализируя среднее значение результатов по добровольцам. Для добровольца-женщины он оказался наименьшим. Однако при тестировании букв “A”, “U” и “V” этот доброволец имеет лучшие результаты (рис. 6). Можно предположить, что это связано с большей гибкостью рук у добровольца-женщины по сравнению с другими добровольцами-мужчинами.

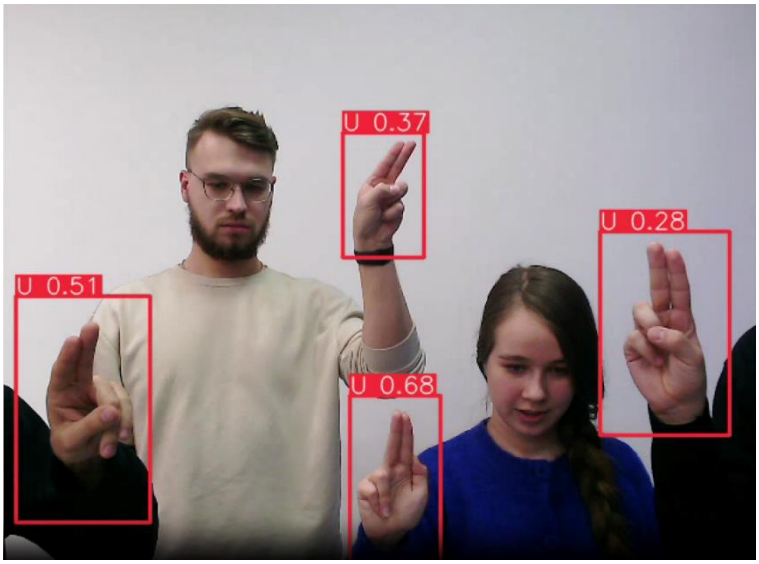

Рис. 6. Распознавание латинской буквы “U”

2. **Тест «несколько рук вместе, показывающих разные буквы»**

Следующим этапом тестирования стало проведение эксперимента с использованием разных букв с целью собрать полноценное слово. Для испытания были выбраны слова “love”, “word”, “boys”, “CUDA”, так как они содержат в себе буквы, результаты распознавания которых имеют сильные различия (см. табл. 1) и имеют длину, равную количеству добровольцев. К тому же технические свойства камеры, использованной для экспериментов, не позволяли поместить в кадр больше букв, соблюдая условия, необходимые для корректного распознавания.

В ходе проведения эксперимента с использованием инструмента распознавания языка жестов в реальном времени были достигнуты желаемые результаты. Испытатели смогли успешно составить руками выбранные слова (см. рис. 7–10).

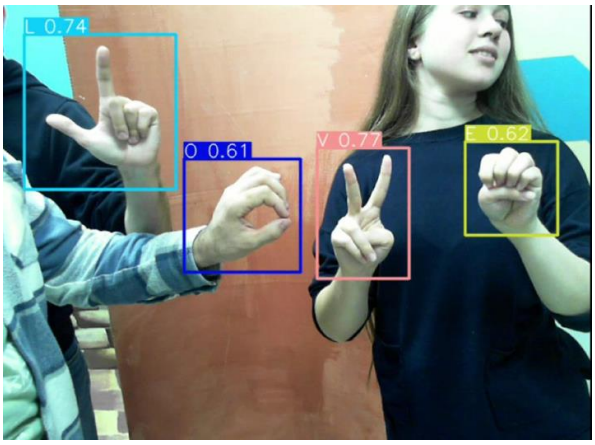

Рис. 7. Демонстрация слова “LOVE” в модели распознавания в реальном времени

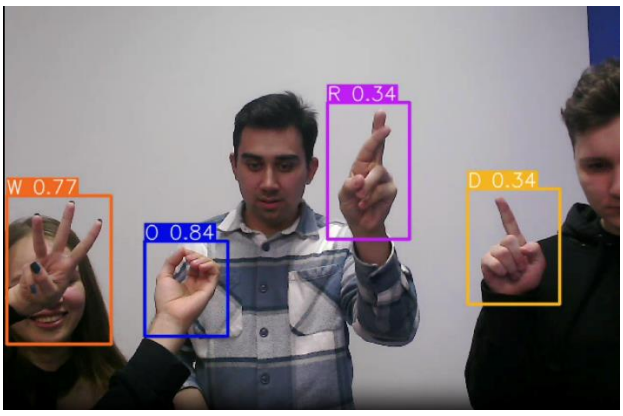

Рис. 8. Демонстрация слова “WORD” в модели распознавания в реальном времени

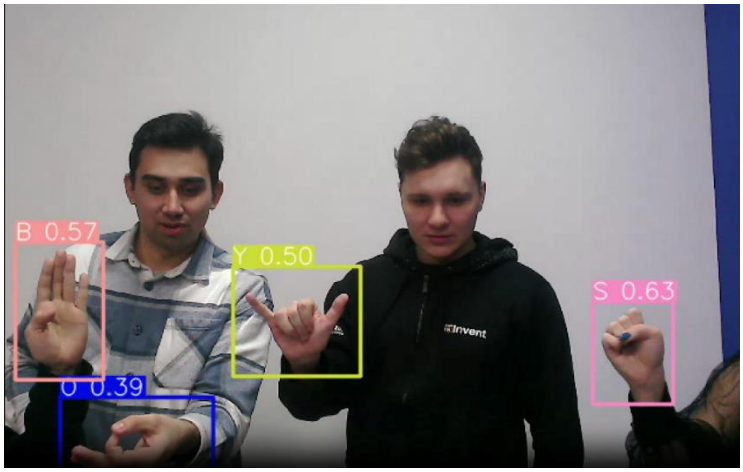

Рис. 9. Демонстрация слова “BOYS” в модели распознавания в реальном времени

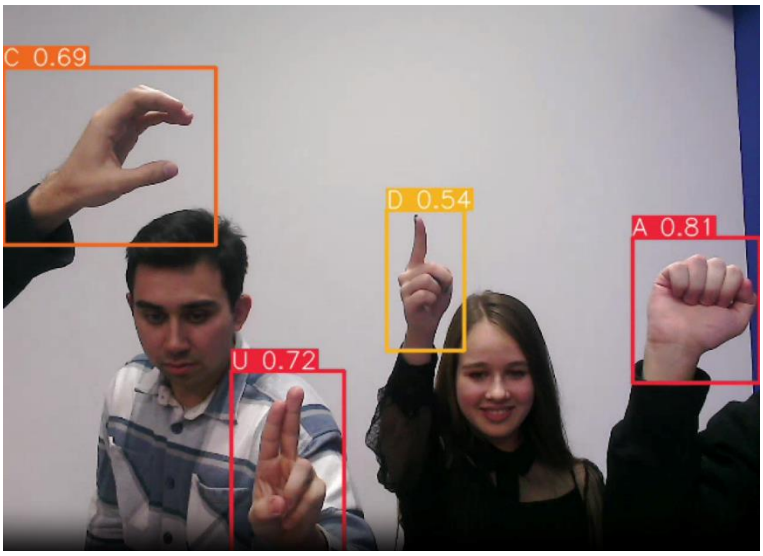

Рис. 10. Демонстрация слова “CUDA” в модели распознавания в реальном времени

В ходе данного эксперимента средние результаты добровольцев №1 и №3 улучшились, добровольца №2 – ухудшились, добровольца №4 – не изменились. Самую высокую вероятность распознавания имела буква “A”, самую худшую – буквы “D” и “R”. Результаты проведенного эксперимента приведены в таблице 2.

Таблица 2. Результаты точности эксперимента с показом слова

Буква/Доброволец	**1**	**2**	**3**	**4**	**Среднее по буквам**
A				0,81	0,81
B	0,57				0,57
C	0,69		0,54		0,62
D				0,34	0,34
E			0,62		0,62
L	0,74				0,74
O	0,84	0,50			0,67
R		0,34			0,34
S			0,63		0,63
U		0,72			0,72
V			0,77		0,77
W			0,77		0,77
Y				0,50	0,50
Средний результат	0,71	0,52	0,66	0,55	

По результатам двух экспериментов самой распознаваемой оказалась буква “W”, самой трудной для распознавания – буква “D”. Среди добровольцев самыми распознаваемыми в среднем оказались жесты добровольца №1, самые худшие результаты – у добровольца №2. Результаты сравнения приведены в таблицах 3 и 4.

Таблица 3. Результаты точностей распознавания букв по двум типам экспериментов

Буква/Эксперимент	**1**	**2**	**Среднее по экспери-ментам**
A	0,44	0,81	0,63
B	0,68	0,57	0,63
C	0,50	0,62	0,56
D	0,45	0,34	0,40
E	0,35	0,62	0,49
L	0,53	0,74	0,64
O	0,52	0,67	0,60
R	0,52	0,34	0,43
S	0,48	0,63	0,56
U	0,46	0,72	0,59
V	0,60	0,77	0,69
W	0,85	0,77	0,81
Y	0,52	0,50	0,51
Средний результат	0,54	0,62	

Таблица 4. Результаты точностей распознавания букв в ходе двух экспериментов по добровольцам

Доброволец/Экспери-мент	**1**	**2**	**Среднее по экспери-ментам**
1	0,58	0,71	0,65
2	0,52	0,52	0,52
3	0,49	0,66	0,58
4	0,63	0,55	0,59
Средний результат	0,56	0,61	0,59

ВЫВОДЫ

В ходе экспериментов выявлено, что на результат распознавания могут влиять такие факторы, как:

- «правильность» расположения рук;
- ракурс;
- расстояние до камеры.

Также для корректного распознавания символов является обязательным обеспечение контрастности между цветом руки и фона. При несоблюдении этого правила буквы могут распознаваться неправильно, как, например, на рис. 11. В этом случае вместо буквы “O” была распознана буква “F”.

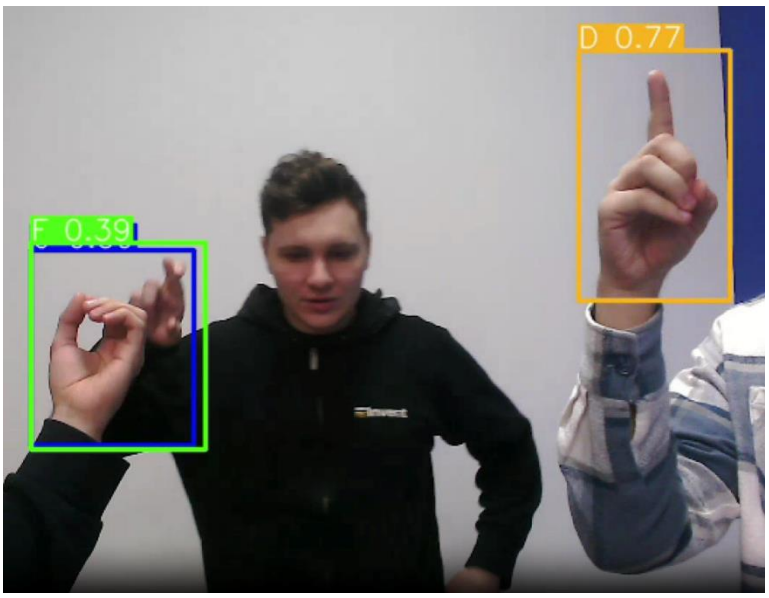

Рис. 11. Неправильное распознавание буквы из-за сливания руки с фоном

ЗАКЛЮЧЕНИЕ

Разработан инструмент для распознавания жестов в режиме реального времени. Результаты его тестирования оказались удовлетворительными и выявили такие проблемы, как неправильный ракурс, помехи из-за слияния с фоном и некорректное расстояние от камеры до руки.

Подчеркнем, что разработка инструмента для распознавания жестов в режиме реального времени является важным шагом в развитии технологий взаимодействия человека с компьютером. Этот инструмент будет иметь значительный потенциал развития и применения в различных сферах, включая виртуальную и

дополненную реальности, медицину, игровую индустрию и другие. Он может повысить удобство и эффективность взаимодействия, а также создать новые возможности для людей с нарушениями слуха. По нашему мнению, будущее данной технологии весьма перспективно, и ее использование будет только расширяться и совершенствоваться в ближайшее время.

Благодарности

Авторы выражают благодарность Максиму Олеговичу Таланову, доценту кафедры интеллектуальной робототехники Института ИТИС КФУ, за помощь в исследовании и рекомендации по проведению экспериментов.

СПИСОК ЛИТЕРАТУРЫ

1. Статистика по инвалидности в Российской Федерации. URL: https://rosstat.gov.ru/folder/13964.

2. *Лугарези К.* MediaPipe: фреймворк для построения конвейеров восприятия // Распределенные, параллельные и кластерные вычисления. 2019.

3. *Поляк М.Д., Кузьмин А.Д.* Распознавание жестов в видеопотоке // Сборник докладов Третьей Международной научной конференции. Санкт-Петербург, 2023.

4. *Чанг Е.А., Бенальказар М.Е.* Модель распознавания жестов рук в реальном времени с использованием методов глубокого обучения и сигналов ЭМГ. 27-я Европейская конференция по обработке сигналов. Ла-Корунья, Испания, 2019.

5. Документация YOLOv8. URL: https://docs.ultralytics.com

6. Чжоу Ч. и др. Обнаружение объектов за 20 лет: Обзор // Компьютерное зрение и распознавание образов. 2019.

7. *Сари И. и др.* Сравнение производительности архитектур YOLOv5 и YOLOv8 при обнаружении человека с помощью аэрофотоснимков // Ultima Computing Jurnal Sistem Komputer. 2023.

8. Набор данных букв американского языка жестов. URL: https://public.roboflow.com/object-detection/american-sign-language-letters/1

TOOL FOR REAL-TIME SIGN LANGUAGE RECOGNITION FROM A VIDEO STREAM

I. I. Zakiryanov[1] [0009-0009-3448-4844], I. Z. Khayaleeva[2] [0009-0007-5837-7010],
I. A. Valishin[3] [0009-0006-6891-031X], E. D. Kurito[4] [0009-0000-6214-135X],
A. N. Faskhutdinov[5] [0009-0001-2766-4048]

[1–5] *Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, ul. Kremlyovskaya, 35, Kazan, 420008*

[1]zakiryanov.iskander@mail.ru, [2]izidakh@yandex.ru, [3]iskander1998@list.ru, [4] ekurito@gmail.com, [5]azatazat835@mail.ru

Abstract

A tool has been developed that recognizes words or individual letters from a video stream in real time. The possibilities and prospects for its application in modern society are considered. The results of experiments to test the performance of this tool using the example of English words and Latin letters are presented.

Keywords*: gesture recognition, neural networks, computer vision, YOLO.*

REFERENCES

1. Disability statistics in the Russian Federation. URL: https://rosstat.gov.ru/folder/13964.

2. *Lugaresi K.* MediaPipe: A Framework for Building Perception Pipelines // Distributed, Parallel, and Cluster Computing. 2019.

3. *Polyak M.D., Kuzmin A.D*. Gesture recognition in a video stream. Collection of reports of the Third International Scientific Conference. St. Petersburg, 2023.

4. *Chung E.A., Benalcázar M.E.* Real-Time Hand Gesture Recognition Model Using Deep Learning Techniques and EMG Signals. 27th European Signal Processing Conference (EUSIPCO). A Coruna, Spain, 2019.

5. Ultralytics YOLOv8 Docs. URL: https://docs.ultralytics.com

6. *Zou Z.* Object Detection in 20 Years: A Survey. IEEE, 2019.

7. *Sary I et al.* Performance Comparison of YOLOv5 and YOLOv8 Architectures in Human Detection using Aerial Images. Ultima Computing Jurnal Sistem Komputer. 2023.

8. American Sign Language Letters Dataset.
URL: https://public.roboflow.com/object-detection/american-sign-language-letters/1

СВЕДЕНИЯ ОБ АВТОРАХ

ЗАКИРЬЯНОВ Искандер Илгизарович – студент магистратуры Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Iskander Ilgizarovich ZAKIRYANOV – Master's student at the Institute of Information Technology and Intelligent Systems, Kazan Federal University.

email: zakiryanov.iskander@mail.ru

ORCID: 0009-0009-3448-4844

ХАЯЛЕЕВА Изида Зуфаровна – студентка магистратуры Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Izida Zufarovna KHAYALEEVA – Master's student at the Institute of Information Technology and Intelligent Systems, Kazan Federal University.

email: izidakh@yandex.ru

ORCID: 0009-0007-5837-7010

ВАЛИШИН Искандер Айратович – студент магистратуры Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Iskander Airatovich VALISHIN – Master's student at the Institute of Information Technology and Intelligent Systems, Kazan Federal University.

email: iskander1998@list.ru

ORCID: 0009-0006-6891-031X

КУРИТО Егор Дмитриевич – студент магистратуры Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Egor Dmitrievich KURITO – Master's student at the Institute of Information Technology and Intelligent Systems, Kazan Federal University.

email: ekurito@gmail.com

ORCID: 0009-0000-6214-135X

ФАСХУТДИНОВ Азат Нафисович – студент магистратуры Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Azat Nafisovich FASKHUTDINOV – Master's student at the Institute of Information Technology and Intelligent Systems, Kazan Federal University.

email: azatazat835@mail.ru

ORCID: 0009-0001-2766-4048

Материал поступил в редакцию 17 декабря 2023 года