ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2025, том ?, с. ??–??

МАТЕМАТИКА

УДК ???.??

РАЗРАБОТКА СИСТЕМЫ ГОЛОСОВОГО УПРАВЛЕНИЯ МАНИПУЛЯТОРОМ ДЛЯ ВЗАИМОДЕЙСТВИЯ С ОКРУЖАЮЩЕЙ СРЕДОЙ

© **2025 г.** **А. Л. Валиуллин**[1,*], **Г. А. Стойко**[1,**]

Представлено академиком ?. ?. ???
Поступило ??.??.2025
После доработки ??.??.2025
Принято к публикации ??.??.2025

Была разработана система голосового управления коллаборативным манипулятором KUKA LBR iiwa 14 с использованием компьютерного зрения для задачи обнаружения и подачи ручного инструмента.

Ключевые слова и фразы: коллаборативный манипулятор, компьютерное зрение, система управления манипулятором, голосовое управление

DOI: ??.???/???

ВВЕДЕНИЕ

Развитие человеко-машинного взаимодействия в последние годы приобретает всё большее значение в контексте автоматизации производственных и сервисных процессов. Особенно актуальной становится задача создания интуитивных и эффективных интерфейсов управления, позволяющих оператору взаимодействовать с робототехническими системами без физического контакта. Одним из наиболее естественных и универсальных способов такого взаимодействия является голосовое управление, которое в сочетании с современными методами компьютерного зрения способно обеспечить полуавтоматическую или полностью автоматическую координацию действий манипулятора в реальной среде.

На этом фоне всё большую популярность приобретают модели типа VLA (Vision-Language-Action), которые сразу из коробки способны интерпретировать команды на естественном языке, распознавать объекты в визуальной сцене и автоматически выполнять действия с ними. Такие модели демонстрируют впечатляющие результаты в симулированных и реальных средах, однако требуют значительных вычислительных ресурсов, специализированного оборудования и сложной настройки, что делает их труднопримени мыми в компактных, прикладных и индустриальных системах, где ключевыми являются скорость отклика, простота и надёжность внедрения [1, 2].

В рамках данной работы мы предлагаем более лёгкий, модульный подход к мультимодальному управлению манипулятором, не опирающийся на тяжёлые VLA-модели. Разработанная система управления реализована для коллаборативного манипулятора KUKA LBR iiwa 14 и предназначена для выполнения прикладной задачи — обнаружения и подачи ручного инструмента. Для распознавания объектов используется обычная USB-камера в связке с предобученной моделью YOLOv8 (You Look Only Once) на специализированном под нашу задачу датасете, в связке с моделью сегментации CLIP. Голосовой интерфейс построен на базе классических акустических признаков (MFCC/PLP) и реализован в виде системы с фиксированным набором команд. Взаимодействие между всеми модулями осуществляется с использованием UDP-протокола. Такое решение позволяет обеспечить простоту интеграции между программными модулями, минимальную задержку при передаче команд и независимость подсистем (зрения, распознавания речи и управления движением), что критически важно для обеспечения плавности и надёжности работы в реальном времени.

Таким образом, работа направлена на создание эффективной и доступной системы мультимодального управления, ориентированной на реальное взаимодействие с окружающей средой в условиях ограниченных вычислительных ресурсов.

[1] *Университет Иннополис; Лаборатория Робототехники, Иннополис, Россия*

[*] *E-mail: alik.valiullin2002@yandex.ru*

[**] *E-mail: stojko.g@yandex.ru*

ОБЗОР ПОДХОДОВ И МЕСТО VLA В КОНТЕКСТЕ МУЛЬТИМОДАЛЬНОГО УПРАВЛЕНИЯ

Мультимодальные системы, сочетающие голос, зрение и действия, являются активной областью исследований в робототехнике. Традиционно задачи голосового управления роботами решались с помощью классических методов обработки речи, таких как извлечение признаков (MFCC, PLP) и последующая классификация с помощью VAE (Variational Autoencoder) или прочих нейросетей [1] Такие решения доказали свою эффективность в приложениях с фиксированным набором команд и ограниченными условиями работы.

В последние годы появились масштабные архитектуры типа VLA, которые используют глубоко обученные модели для комплексной интерпретации речевых команд, визуального восприятия сцены и генерации соответствующих действий. Уже в 2019 году в обзоре [2] отмечалось, что большинство существующих решений страдает от узкой специализации, высокой чувствительности к сценарию и отсутствия универсальности при работе с произвольными объектами. Также подчёркивалось, что практически все полуавтономные или автономные визуальные системы опираются на сложные модели и требуют мощной вычислительной базы, включая отдельные графические процессоры, что резко ограничивает их применение вне лабораторных условий.

Спустя шесть лет, несмотря на бурное развитие крупных мультимодальных языковых моделей, эти фундаментальные ограничения остались актуальны. Так, в одном из самых полных обзоров 2025 года — статье [3] — подчёркивается, что большинство современных VLA-архитектур демонстрируют отличные результаты в моделировании поведения, но крайне редко используются в реальных робототехнических системах из-за высокой задержки вывода (output latency), значительных требований к вычислительным ресурсам и сложности развертывания.

В частности, современные VLA-модели работают на частотах генерации 3–5 Гц, что в разы ниже необходимой частоты обновления управляющего сигнала для безопасного и плавного взаимодействия с физическим миром, где требуются как минимум 50–100 Гц [3]. Для сравнения, простые системы, основанные на классических методах обработки речи и визуального распознавания, обеспечивают отклик с задержкой менее 100 мс — что критично для взаимодействия в реальном времени. Дополнительно отмечается, что VLA-системы обычно требуют дорогих GPU-кластеров и сложных механизмов распределённого вывода, что практически исключает их применение в мобильной, встроенной или локальной робототехнике.

Для наглядности ниже представлена таблица, демонстрирующая основные различия между тяжеловесной VLA-архитектурой и нашей системой:

Таблица 1. Сравнение VLA-моделей и предложенной системы

Параметр	**VLA-модель (например, SayCan / RT-2)**	**Предложенная система**
Архитектура	Vision + LLM + Action Planner	YOLO + CLIP + классификатор речи
Размер модели	5 - 20 млрд параметров (LLM)+0.5 - 2 млрд (Vision)	3.2 млн параметров YOLO 8n + 63 млн параметров CLIP + 6 млн (аудио-классификатор)
Вычислительные ресурсы	Высокие ($>$ 16 ГБ VRAM, $>$ 16 - 32 ГБ RAM)	Низкие ($<$ 12 Гб VRAM, $<$ 8 ГБ RAM)
Средняя задержка реакции	Высокая (200–1000 мс)	Низкая ($<$100 мс)
Частота обновления	3–5 Гц	$>$50 Гц
Применимость в реальном времени	Ограничена	Полноценная
Связь между модулями	Внутренние трансформеры, multi-head attention	UDP-пакеты между независимыми модулями, модульная архитектура
Надёжность при сбоях и перегрузках	Низкая (зависит от целостности всех компонентов)	Высокая (каждый модуль может перезапускаться отдельно)
Интерфейс управления	Естественный язык, но требует интерпретации	Ограниченный набор команд, но стабильный

Таким образом, несмотря на очевидные успехи VLA-моделей в моделировании общего поведения агента, их применимость в задачах, требующих низкой задержки, компактности, отказоустойчивости

и способности к работе в реальном времени, остаётся крайне ограниченной. Наш подход, напротив, исходит из инженерной прагматики и направлен на достижение высокой функциональности с опорой на минимально необходимую архитектуру и эффективную коммуникацию между модулями. Это делает систему пригодной для развертывания в реальных прикладных сценариях, включая промышленную среду и лабораторную робототехнику.

ПОСТАНОВКА ЗАДАЧИ

Современные робототехнические системы, ориентированные на взаимодействие с человеком, требуют от интерфейсов управления высокой интуитивности, адаптивности и возможности функционирования в режиме реального времени. Особенно остро такие требования проявляются в задачах, связанных с коллаборативными манипуляторами, которые работают в непосредственной близости с человеком и нередко — по его запросу или команде.

Контекст. Коллаборативный манипулятор KUKA LBR iiwa 14, обладает высокой точностью и безопасной конструкцией, что делает его идеальным для применения в гибких системах автоматизации, не требующих ограждений. Однако само по себе это устройство не наделено высокоуровневым «пониманием» среды и не обладает встроенными средствами восприятия и интерпретации команд.

Одной из практических задач, возникающих в процессе обслуживания оператора, является автоматическое распознавание конкретного инструмента на рабочем столе (например, отвёртки, гаечного ключа, плоскогубцев, молотка) и его последующая передача человеку. Желательно, чтобы такая задача могла выполняться по голосовой команде, без необходимости использования ручных интерфейсов, кнопок или панелей управления. Это особенно важно в контексте:

- производственных процессов, где у оператора заняты руки;
- медико-биологических лабораторий, где требуется стерильность;
- образовательных или демонстрационных целей (интуитивный интерфейс).

Требования к системе. Таким образом, требуется разработать систему, которая удовлетворяет следующим критериям:

1. Голосовое управление
   Пользователь должен иметь возможность задавать команды (например, «дай мне отвёртку») с помощью речи. Система должна распознать ключевые команды из ограниченного словаря и преобразовать их в формализованные действия.
2. Автоматическое зрительное восприятие
   Система должна идентифицировать нужный объект в поле зрения (рабочий стол) с помощью компьютерного зрения, без использования маркеров или QR-кодов. Распознавание должно быть устойчиво к изменениям освещения, фону и положению объекта.
3. Контроль манипулятора
   Манипулятор KUKA LBR iiwa 14 должен выполнить захват соответствующего объекта и подать его оператору, соблюдая ограничения по траектории и безопасности.
4. Работа в реальном времени
   Вся система (включая обработку изображения, распознавание речи, принятие решения и управление роботом) должна функционировать с минимальной задержкой и быть пригодной для работы в условиях реального времени.
5. Модульность и лёгкость развёртывания
   Система должна быть построена из независимых модулей, работающих по лёгкому сетевому протоколу, что упростит отладку, масштабирование и внедрение в другие проекты.
6. Работа на ограниченных ресурсах
   Система не должна полагаться на облачные вычисления или требовать тяжёлых моделей VLA-класса. Все компоненты должны быть работоспособны на доступном оборудовании.

Ограничения и предположения. В процессе работы предполагается:

- использовать статичный набор голосовых команд, не более 10–15 (например, "дай мне отвёртку");
- использовать модель YOLO, обученную под поставленную задачу для распознавания объектов;
- использовать модель сегментации CLIP и PCA для определения места захвата объекта;
- ограничить рабочую зону пространством стола с инструментами;
- обеспечить предварительное согласование координат камеры и манипулятора;

Цель работы. Таким образом, основная цель данной работы — разработать и продемонстрировать прототип системы голосового управления коллаборативным манипулятором KUKA LBR iiwa 14,

использующей методы компьютерного зрения и речевого интерфейса, пригодной для задач подачи ручного инструмента в условиях реального времени и ограниченных вычислительных ресурсов.

ОБЗОР РАЗРАБОТАННОЙ СИСТЕМЫ

Для распознавания голосовых команд в системе используется модель Vosk — лёгкая и автономная система, не требующая подключения к облачным сервисам. В частности, применена модель vosk-model-small-ru-0.22, специально оптимизированная для русского языка и способная работать в реальном времени. Аудиопоток с микрофона обрабатывается в непрерывном режиме в отдельном потоке, после чего анализируется текстовая расшифровка. Происходит лемматизация, из которой извлекается ключевое слово (например, «молоток»), и на её основе формируется структура запроса, содержащая имя целевого объекта. Эта структура затем передаётся в модуль визуальной обработки.

В качестве средства визуального восприятия сцены использовалась обычная USB-камера, закреплённая на эндефекторе манипулятора. Такое размещение позволяет получать изображение непосредственно из рабочей зоны захвата. Камера передаёт видео в реальном времени на вычислительный узел, где происходит последующая обработка — детекция, сегментация и определение координат точки захвата.

Для задачи детекции ручного инструмента была использована модель YOLOv8n — компактный и эффективный вариант в семействе YOLO. Обучение проводилось на кастомном датасете Mechanical tools-10000 с платформы Roboflow, включающем 9302 изображения 5 классов инструментов (отвёртка, гаечный ключ, плоскогубцы, молоток, дрель), собранных в разнообразных ракурсах и освещении. Разметка включала прямоугольные bounding boxes. Модель обучалась 50 эпох с использованием transfer learning на предобученных весах от ultralytics, что позволило добиться хорошей сходимости и устойчивости к переобучению даже при относительно небольшом объёме датасета. Финальная модель обеспечивает точность (mAP@0.5) в среднем 91% при минимальном размере и возможности запуска в реальном времени.

Для уточнения формы объекта используется семантическая сегментация с помощью модели CLIP от CIDAS. После получения точной маски применяется PCA, поскольку этот метод эффективно определяет главные оси симметрии объекта. В случае вытянутых объектов (например, отвёртки или гаечного ключа) основная компонента соответствует направлению инструмента, а проекция центра масс на вторую компоненту обеспечивает устойчивую, симметричную точку захвата. Это решение оказалось простым, эффективным и лишённым необходимости в обучении, что идеально для real-time применения.

В процессе работы системы после успешной детекции инструмента (YOLOv8n) и уточняющей сегментации (CLIP + PCA) формируется точка захвата — координата в пространстве задачи (task space), представленная в виде (x, y). В текущей реализации координата высоты захвата z вычислялась вручную, угол ориентации эндеффектора θ является фиксированным. Эти координаты после преобразований являются мировыми и отмасштабированными относительно системам координат камеры и эндеффектора.

Полученные координаты передаются по протоколу UDP в отдельную программу-интерпретатор, запущенную на управляющем хосте. Эта программа решает задачу обратной кинематики (inverse kinematics) для манипулятора KUKA LBR iiwa 14. На основании полученной целевой точки и текущего состояния манипулятора она вычисляет набор углов сочленений $q = [q_1, q_2, ..., q_7]$, которые обеспечивают достижение требуемого положения и ориентации захвата.

После вычисления значений q, эти данные отправляются напрямую на контроллер манипулятора по интерфейсу FRI (Fast Robot Interface), который обеспечивает низкоуровневую передачу управляющих параметров с высокой частотой и минимальной задержкой. Таким образом, вся цепочка — от голосовой команды до физического движения — разделена на независимые модули, соединённые через UDP и FRI, что повышает гибкость системы и облегчает отладку, расширение и интеграцию.

В текущей итерации системы в качестве исполнительного устройства использовался захват с сильным ниодимовым магнитом, закреплённом на конце эндеффектора манипулятора. Для повышения надёжности взаимодействия в процессе захвата, на сами инструменты также были дополнительно прикреплены небольшие неодимовые магниты. Такая схема позволила реализовать простую и надёжную механику захвата без необходимости в сложных пальцевых или пневматических захватах.

Для наглядности разработанная система голосового управления коллаборативным манипулятором KUKA LBR iiwa 14 представлена на рисунке 1.

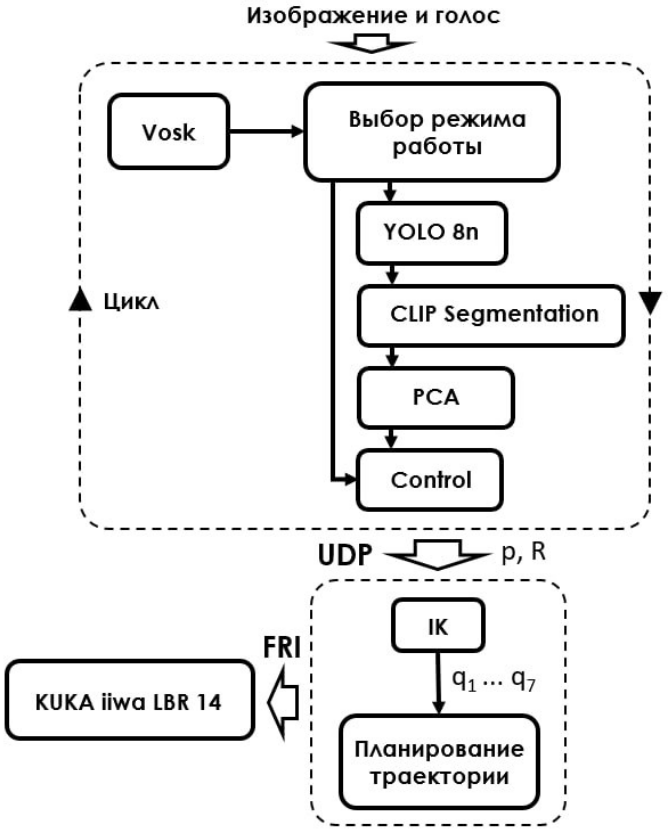

Рис. 1. Архитектура системы голосового управления манипулятором

ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

Для оценки работоспособности разработанной системы были проведены экспериментальные испытания, охватывающие все ключевые компоненты — от голосового ввода до физического захвата инструмента манипулятором. Целью экспериментов было подтверждение корректности и устойчивости выполнения полной цепочки действий в условиях, приближённых к реальному рабочему сценарию.

Эксперименты проводились в лабораторной обстановке с искусственным освещением и использованием рабочего стола с размещёнными на нём ручными инструментами. В составе тестового набора присутствовали 4 объекта: отвёртка, гаечный ключ, плоскогубцы и молоток. Все инструменты, кроме гаечного ключа, были оснащены небольшими неодимовыми магнитами для совместимости с магнитным захватом. Камера была закреплена на эндеффекторе манипулятора и направлена вертикально вниз.

Рабочая зона манипулятора в виде стола с инструментами и сам манипулятор KUKA iiwa LBR 14 изображены на рисунке 2.

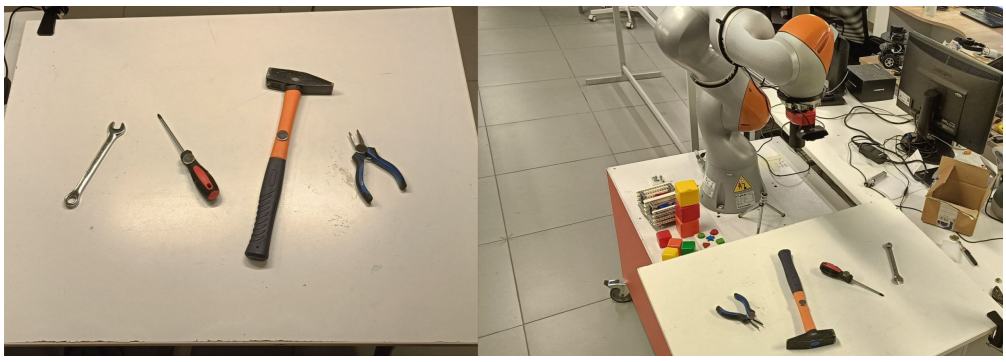

Рис. 2. Рабочая зона манипулятора в виде стола с инструментами и сам манипулятор

В ходе проведённых экспериментов система показала стабильную работу на всех этапах обработки команды — от голосового ввода до физического взаимодействия манипулятора с объектом. Распознавание голосовых команд с использованием модели Vosk происходило корректно в большинстве случаев. YOLOv8n уверенно детектировала ручные инструменты на изображениях с USB-камеры, закреплённой на эндеффекторе, демонстрируя высокую точность локализации при различных ориентациях и положениях объектов на столе.

Модель CLIP хорошо справлялась с уточнением формы нужного инструмента. Применение PCA для вычисления точки захвата давало в большинстве случаев правильную точку захвата инструмента и обеспечивало успешное притяжение инструмента с помощью магнитного захвата.

Средняя задержка между моментом подачи голосовой команды и началом движения манипулятора составила около 1.1 секунды, где распознавание команды через vosk происходило примерно за 1000 мс, а задержка передачи управляющих воздействий на манипулятор составила менее 100 мс.

Результат работы системы распознавания выбранного инструмента показан на рисунке 3. В репозитории данного проекта представлен код и результаты проделанной работы.

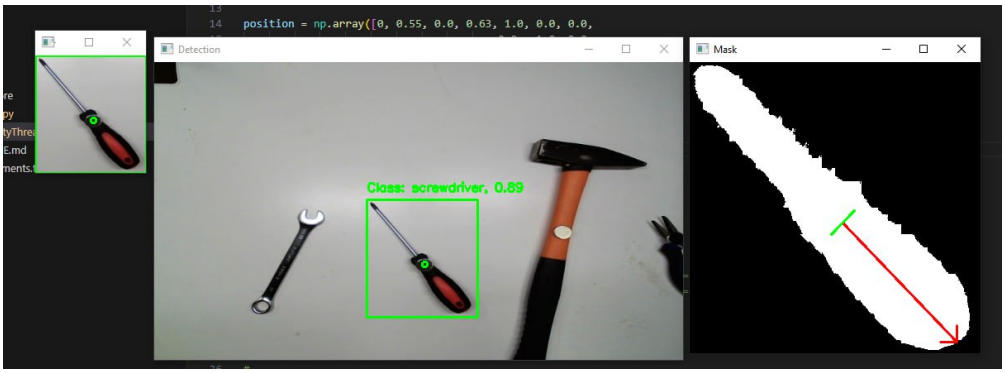

Рис. 3. Результат работы системы распознавания выбранного инструмента

В целом, система показала себя как надёжный мультимодальный интерфейс управления манипулятором, способный уверенно выполнять команды пользователя в реальном времени. Несмотря на удовлетворительные результаты, есть много направлений работы для последующего улучшения системы управления.

ВЫВОДЫ И ДАЛЬНЕЙШИЕ УЛУЧШЕНИЯ

В данной работе была разработана, реализована и экспериментально проверена система голосового управления коллаборативным манипулятором KUKA LBR iiwa 14 с использованием компьютерного зрения и семантической сегментации. Система ориентирована на выполнение прикладной задачи — захвата и подачи ручного инструмента по голосовой команде в условиях реального времени. При этом особое внимание уделялось отказу от ресурсоёмких VLA-моделей в пользу более лёгких и модульных решений, пригодных для развёртывания на ограниченных вычислительных ресурсах.

Основными результатами работы стали:

- Успешная интеграция речевого интерфейса с помощью модели vosk-model-small-ru-0.22, обеспечивающей локальное и стабильное распознавание коротких голосовых команд;
- Надёжная детекция инструментов с использованием YOLOv8n, обученной на кастомном датасете Mechanical tools-10000 из Roboflow;
- Уточнение формы объекта с помощью CLIP;
- Определение точки захвата с помощью PCA, без необходимости использования обучаемых моделей или датчиков обратной связи;
- Реализация полной цепочки управления от визуального восприятия до кинематического управления манипулятором через FRI с использованием обмена по UDP;

Система продемонстрировала, что даже с использованием лёгких моделей и дешёвого оборудования возможно построить мультимодальный интерфейс, пригодный для прикладных задач в производстве, лабораториях и образовательной среде.

Дальнейшие направления развития.Несмотря на полученные положительные результаты, разработанная система может быть расширена и улучшена по нескольким направлениям:

1. Расширение набора объектов

Дальнейшее обучение YOLO на большем количестве инструментов и условиях (разные фоны, материалы) повысит универсальность и надёжность системы в произвольной обстановке.

2. Замена магнитного захвата на активный
   Несмотря на простоту магнитного захвата, в будущем целесообразно внедрение универсального захвата. В этом случае необходимо будет учитывать ещё и ориентацию захватываемого объекта.
3. Улучшение алгоритма определения точки захвата объекта
   В текущей версии системы точка захвата рассчитывается автоматически с помощью PCA. Такой подход оказался достаточно эффективным для симметричных и сбалансированных объектов (например, отвёрток или ключей), однако в случае с инструментами со смещённым центром массы — например, молотком — он может приводить к неоптимальному выбору точки. В перспективе можно внедрить небольшой модуль, обученный на размеченных изображениях с указанием предпочтительных точек захвата для каждого инструмента.
4. Общее улучшение плавности управления
   В текущей реализации движение осуществляется по заранее рассчитанной конечной точке с помощью команды из модуля обратной кинематики, переданной на KUKA LBR iiwa 14 через FRI. Хотя такой подход обеспечивает точность позиционирования, перемещение манипулятора может выглядеть резким или излишне прямолинейным. Вместо одномоментной команды «перейти в точку», целесообразно использовать траектории.

Таким образом, разработанная система служит доказательством того, что эффективное мультимодальное управление робототехническим манипулятором может быть реализовано без привлечения ресурсоёмких нейросетевых архитектур, с акцентом на инженерную простоту, устойчивость и надёжность. В дальнейшем она может стать основой для более универсальных автономных ассистентов, работающих в производстве, медицине и повседневной среде.

СПИСОК ЛИТЕРАТУРЫ

[1] *Amal Punchihewa, Zuriawati Mohd Arshad* Voice Command Interpretation for Robot Control, 2011 // School of Engineering & Advanced Technology Massey University, Palmerston North, New Zeland / IBM Systems & Technology Group IBM Malaysia Sdn. Bhd., Selangor, Malaysia.

[2] *Stefan Hein Bengtson, Thomas Bak, Lotte N. S. Andreasen Struijk, Thomas Baltzer Moeslund* A review of computer vision for semi-autonomous control of assistive robotic manipulators (ARMs), 2019 // Visual Analysis of People (VAP) Laboratory, Department of Architecture, Design, and Media Technology, Aalborg University, Aalborg, Denmark / Automation and Control, Department of Electronic Systems, Aalborg University, Aalborg, Denmark / Department of Health Science and Technology, Aalborg University, Aalborg, Denmark.

[3] *Ranjan Sapkota, Yang Cao, Konstantinos I. Roumeliotis, Manoj Karkee* Vision-Language-Action Models: Concepts, Progress, Applications and Challenges, 2025 // Cornell University, Biological & Environmental Engineering, Ithaca, New York, USA / The Hong Kong University of Science and Technology, Department of Computer Science and Engineering, Hong Kong / University of the Peloponnese, Department of Informatics and Telecommunications, Greece.

DEVELOPMENT OF A VOICE CONTROL SYSTEM FOR MANIPULATOR FOR INTERACTION WITH THE ENVIRONMENT

A. L. Valiullin[a,*], G. A. Stoyko[a,]**

[a]Innopolis University, Robotics Laboratory,
Innopolis, Russian Federation

Presented by Academician of the ??? ?. ?. ???

We have developed a voice control system for the collaborative manipulator KUKA LBR iiwa 14 using computer vision for the task of detecting and giving hand tools.

Keywords: collaborative manipulator, computer vision, control system for the manipulator, voice control

REFERENCES

[1] *Amal Punchihewa, Zuriawati Mohd Arshad* Voice Command Interpretation for Robot Control, 2011 // School of Engineering & Advanced Technology Massey University, Palmerston North, New Zeland / IBM Systems & Technology Group IBM Malaysia Sdn. Bhd., Selangor, Malaysia.

[2] *Stefan Hein Bengtson, Thomas Bak, Lotte N. S. Andreasen Struijk, Thomas Baltzer Moeslund* A review of computer vision for semi-autonomous control of assistive robotic manipulators (ARMs), 2019 // Visual Analysis of People (VAP) Laboratory, Department of Architecture, Design, and Media Technology, Aalborg University, Aalborg, Denmark / Automation and Control, Department of Electronic Systems, Aalborg University, Aalborg, Denmark / Department of Health Science and Technology, Aalborg University, Aalborg, Denmark.

[3] *Ranjan Sapkota, Yang Cao, Konstantinos I. Roumeliotis, Manoj Karkee* Vision-Language-Action Models: Concepts, Progress, Applications and Challenges, 2025 // Cornell University, Biological & Environmental Engineering, Ithaca, New York, USA / The Hong Kong University of Science and Technology, Department of Computer Science and Engineering, Hong Kong / University of the Peloponnese, Department of Informatics and Telecommunications, Greece.

DEVELOPMENT OF A VOICE CONTROL SYSTEM FOR MANIPULATOR FOR INTERACTION WITH THE ENVIRONMENT

A. L. Valiullin[a,*], **G. A. Stoyko**[a,**]

[a]Innopolis University, Robotics Laboratory,
Innopolis, Russian Federation

Presented by Academician of the ??? ?. ?. ???