|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

|  |  |
| --- | --- |
| ФАКУЛЬТЕТ | ИУ «Информатика, искусственный интеллект и системы управления» |

|  |  |
| --- | --- |
| КАФЕДРА | ИУ-1 «Системы автоматического управления» |

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ***

***НА ТЕМУ:***

|  |
| --- |
|  |
| ***Система распознавания жестов*** |
| ***на видеопоследовательности*** |
| ***для реализации интерфейсов*** |
| ***человеко-машинного взаимодействия*** |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | ИУ1-81Б |  | 28/05/2022 |  | Д.И. Юдаков |
|  | (Группа) |  | (Подпись, дата) |  | (И.О. Фамилия) |

|  |  |  |  |
| --- | --- | --- | --- |
| Руководитель | 28/05/2022 |  | К.В. Парфентьев |
|  | (Подпись, дата) |  | (И.О. Фамилия) |

|  |  |  |  |
| --- | --- | --- | --- |
| Нормконтролер | 28/05/2022 |  | Т.Ю. Цибизова |
|  | (Подпись, дата) |  | (И.О. Фамилия) |

*2022 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Заведующий кафедрой | | | | | | | ИУ-1 | | |
|  | | | | | | | (Индекс) | | |
|  |  | | | |  | К.А. Неусыпин | | | |
|  | | | | | | (И.О. Фамилия) | | | |
|  | « | 11 | » | февраля | | | 20 | 22 | г. |

**ЗАДАНИЕ**

**на выполнение выпускной квалификационной работы бакалавра**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Студент группы | | | ИУ1-81Б | | | | |  | | | |
|  | | | | | | | | | | | |
| Юдаков Дмитрий Игоревич | | | | | | | | | | | |
| (фамилия, имя, отчество) | | | | | | | | | | | |
|  | | | | | | | | | | | |
| Тема квалификационной работы | | | | | | | | Система распознавания жестов на | | | |
| видеопоследовательности для реализации интерфейсов человеко-машинного | | | | | | | | | | | |
| взаимодействия | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
| Источник тематики (НИР кафедры, заказ организаций и т.п.) | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
| Тема квалификационной работы утверждена распоряжением по факультету | | | | | | | | | | | |
|  | № |  | | от | « |  | » |  | 20 |  | г. |
|  | | | | | | | | | | | |
| ***Часть 1.*** |  | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |
|  | | | | | | | | | | | |

|  |  |
| --- | --- |
| ***Часть 2.*** |  |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
| ***Часть 3.*** |  |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |
|  | |

***Оформление квалификационной работы:***

|  |  |  |
| --- | --- | --- |
| Расчетно-пояснительная записка на |  | листах формата А4. |
| Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.) | | |
| Презентация на слайдах | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |
|  | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Дата выдачи задания | « | 11 | » | февраля | 20 | 22 | г. |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| В соответствии с учебным планом выпускную квалификационную работу выполнить в | | | | | | | |
| полном объеме в срок до | « | 28 | » | мая | 20 | 22 | г. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Руководитель ВКР** |  | 11/02/2022 |  | К.В. Парфентьев |
|  |  | (Подпись, дата) |  | (И.О. Фамилия) |
|  |  |  |  |  |
| **Студент** |  | 11/02/2022 |  | Д.И. Юдаков |
|  |  | (Подпись, дата) |  | (И.О. Фамилия) |

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **ФАКУЛЬТЕТ** | ИУ | |  |  | | **КАФЕДРА** | ИУ-1 | |  |  | | **ГРУППА** |  | |  | УТВЕРЖДАЮ   |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Заведующий кафедрой | | | | | | | ИУ-1 | | | |  | | | | | | | (Индекс) | | | |  |  | | | |  | К.А. Неусыпин | | | | |  | | | | | | (И.О. Фамилия) | | | | |  | « | 11 | » | февраля | | | 20 | 22 | г. | |

**КАЛЕНДАРНЫЙ ПЛАН**

**выполнения выпускной квалификационной работы**

|  |  |  |
| --- | --- | --- |
| студента: | Юдакова Дмитрия Игоревича | |
|  | (фамилия, имя, отчество) | |
| Тема квалификационной работы | | Система распознавания жестов на |
| видеопоследовательности для реализации интерфейсов человеко-машинного | | |
| взаимодействия | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Наименование этапов выпускной квалификационной работы** | **Сроки выполнения этапов** | | **Отметка о выполнении** | |
| **план** | **факт** | **Должность** | **ФИО, подпись** |
|  | Задание на выполнение работы. Формулирование проблемы, цели и задач работы | |  | | --- | | 11/02/2022 | | дата | | |  | | --- | | 11/02/2022 | | дата | | Руководитель ВКР |  |
|  | 1 часть | |  | | --- | | 26/03/2022 | | дата | | |  | | --- | | 26/03/2022 | | дата | | Руководитель ВКР |  |
|  | Утверждение окончательных формулировок решаемой проблемы, цели работы и перечня задач | |  | | --- | | 26/03/2022 | | дата | | |  | | --- | | 26/03/2022 | | дата | | Заведующий кафедрой |  |
|  | 2 часть | |  | | --- | | 27/04/2022 | | дата | | |  | | --- | | 27/04/2022 | | дата | | Руководитель ВКР |  |
|  | 3 часть | |  | | --- | | 10/05/2022 | | дата | | |  | | --- | | 10/05/2022 | | дата | | Руководитель ВКР |  |
|  | 1-я редакция работы | |  | | --- | | 15/05/2022 | | дата | | |  | | --- | | 15/05/2022 | | дата | | Руководитель ВКР |  |
|  | Подготовка доклада и презентации | |  | | --- | | 27/05/2022 | | дата | | |  | | --- | | 27/05/2022 | | дата | | Руководитель ВКР |  |
|  | Заключение руководителя | |  | | --- | | 28/05/2022 | | дата | | |  | | --- | | 28/05/2022 | | дата | | Руководитель ВКР |  |
|  | Допуск работы к защите на ГЭК (нормоконтроль) | |  | | --- | | 28/05/2022 | | дата | | |  | | --- | | 28/05/2022 | | дата | | Нормоконтролер |  |
|  | Внешняя рецензия | |  | | --- | | 28/05/2022 | | дата | | |  | | --- | | 28/05/2022 | | дата | |  |  |
|  | Защита работы на ГЭК | |  | | --- | |  | | дата | | |  | | --- | |  | | дата | |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Студент:* | 11/02/2022 |  | *Руководитель работы:* | 11/02/2022 |
|  | (подпись, дата) |  |  | (подпись, дата) |

АННОТАЦИЯ

????????????????????????????????????????

СОДЕРЖАНИЕ

[СОДЕРЖАНИЕ 1](#_Toc105019386)

[ВВЕДЕНИЕ 3](#_Toc105019387)

[1. ПОДХОДЫ РАСПОЗНАВАНИЯ ЖЕСТОВ 5](#_Toc105019388)

[1.1. Распознавание жестов для человеко-машинного взаимодействия 6](#_Toc105019389)

[1.2. Сенсорные технологии 8](#_Toc105019390)

[1.2.1. Методы на основе получения изображений 8](#_Toc105019391)

[1.2.2. Методы, основанные не на изображениях 9](#_Toc105019392)

[1.2.3. Сравнение сенсорных технологий 10](#_Toc105019393)

[1.3. Идентификация жестов 11](#_Toc105019394)

[1.3.1. Визуальные характеристики 12](#_Toc105019395)

[1.3.2. Алгоритмы обучения 13](#_Toc105019396)

[1.3.3. Модель человека 14](#_Toc105019397)

[1.3.4. Выводы по подходам идентификации жестов 15](#_Toc105019398)

[1.4. Отслеживание жестов 16](#_Toc105019399)

[1.4.1. Отслеживание единичной гипотезы 16](#_Toc105019400)

[1.4.2. Расширенные методы отслеживания 17](#_Toc105019401)

[1.4.3. Сравнение различных подходов отслеживания жестов 18](#_Toc105019402)

[1.5. Классификация жестов 19](#_Toc105019403)

[1.6. Примеры систем распознавания жестов 23](#_Toc105019404)

[2. НЕЙРОСЕТЕВЫЕ ТЕХНОЛОГИИ В ДЕТЕКТИРОВАНИИ И ИДЕНТИФИКАЦИИ КИСТИ ЧЕЛОВЕКА 26](#_Toc105019405)

[2.1. Нейронные сети и их виды 27](#_Toc105019406)

[2.1.1. Свёрточная нейронная сеть 32](#_Toc105019407)

[2.1.2. Рекуррентные нейронные сети 33](#_Toc105019408)

[2.2. Детекторы, использующие ограничивающие прямоугольники 38](#_Toc105019409)

[2.2.1. YOLO 39](#_Toc105019410)

[2.2.3. SSD 40](#_Toc105019411)

[2.3. Применение SSD в задаче детектирования кисти 52](#_Toc105019412)

[3. РАЗРАБОТКА ИНТЕРФЕЙСА ДЛЯ УПРАВЛЕНИЯ ЖЕСТАМИ 54](#_Toc105019413)

[3.1. Методика управления жестами 54](#_Toc105019414)

[3.2. Требования к реализации 55](#_Toc105019415)

[3.3. Подготовка и обучение модели распознавания статических жестов 58](#_Toc105019416)

[3.4. Расширение модели распознавания статических жестов 62](#_Toc105019417)

[3.5. Подготовка и обучение модели распознавания динамических жестов 66](#_Toc105019418)

[3.6. Оценка работы интерфейса 71](#_Toc105019419)

[ЗАКЛЮЧЕНИЕ 73](#_Toc105019420)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 75](#_Toc105019421)

ВВЕДЕНИЕ

Вычислительная способность компьютера стала необходимой в жизни современного человека – без неё довольно трудно жить и работать в нынешних реалиях, а именно его умение быстро помочь в решении самых разнообразных задач, начиная от проведения инженерных расчётов до оптимизации управления предприятием, высоко оценивается человечеством. Камнем преткновения остаётся взаимодействие системы «человек-машина» -- оно должно быть максимально упрощено, чтобы люди и машины коммуницировали на уровне естественных средств общения.

Развитие компьютерных технологий получило прирост за последнюю половину века и стремительно увеличивается и по сей день. Они уже способны обрабатывать и анализировать информацию подобно человеку: распознавать текст, изображения, анализировать звуки и мелодии, произносить осмысленные предложения, распознавать голосовые команды и реагировать на прикосновения пальцев. Но одной из самых приоритетных и наиболее трудных задач в области информационных технологий и интеллектуальных систем является задача распознавания жестов.

Актуальность рассматриваемой тематики обусловлена возможностью применения предлагаемого подхода для управления объектами без тактильного контакта и голосовой идентификации команд, а также простотой с точки зрения конечного пользователя.

В данной работе идёт речь о детектировании кисти руки, извлечения её конфигурации с помощью ключевых точек и распознавание жеста для генерации в соответствии с ним сигнала выхода. В ходе работы исследуются различные методы и техники детектирования кисти человека и распознавания жеста, в частности – методы с использованием средств глубокого обучения. Важным требованием для реализации является минимизация времени задержки на обработку каждого кадра (скорость работы) и точность идентификации жеста.

Поставленной **целью** является разработка метода детектирования, извлечения конфигурации и классификации жеста с помощью языка высокого уровня Python.

Для достижения цели необходимо решить ряд задач:

1. Сделать обзор существующих подходов распознавания жестов, провести их сравнение, выделить особенности.
2. Сделать обзор теоретического материала по детектированию кисти руки человека и её последующей идентификации.
3. Разработать и реализовать методику распознавания жестов.
4. Провести анализ разработанной методики, выявить недостатки и описать пути их устранения.

1. ПОДХОДЫ РАСПОЗНАВАНИЯ ЖЕСТОВ

В последнее время концепция взаимодействия между людьми и роботизированными машинами вызвала множество научных интересов. Вместо того, чтобы машины заменили людей на рабочих местах, сотрудничество человека с ними – это направление, позволяющее сохранить рабочие места, организовав им возможность работать вместе. Кооперация людей и роботов может освободить людей от тяжёлых задач, если будут созданы эффективные каналы связи между ними. Хотя такие каналы связи по-прежнему ограничены, распознавание жестов эффективно применяется в качестве интерфейса между людьми и компьютерами в течение длительного периода времени.

Роботизированные системы уже стали ключевыми компонентами в различных отраслях промышленности. В последнее время концепция Human-Robot Collaboration (HRC) значительно подогрела исследовательские интересы. Литературные примеры предполагали, что работники-люди обладают несравненными навыками решения проблем и сенсорно-двигательными способностями, но имеют ограниченную силу и точность [35, 23]. Однако роботизированные системы имеют высокую стойкость к усталости, скорость, повторяемость и лучшую производительность, но они ограничены в гибкости. HRC может освободить работников-людей от тяжелых задач и устанавливать каналы связи между людьми и роботами для повышения общей эффективности.

В идеале связка HRC должна работать так же, как и человеческая совместная команда. Тем не менее, традиционно методы разделения времени или пространственного разделения применяются в системах HRC, что снижает производительность как для рабочих, так и для роботов [35]. Чтобы создать эффективную команду HRC, в качестве примеров можно проанализировать группы сотрудничества человека с человеком. В совместной работе и сотрудничестве людей существует две теории: теория совместных намерений и теория обучения [14, 63, 13, 11]. Чтобы применить теории в команде HRC, существует три опытных подхода, которые приносят пользу команде HRC:

* Все члены команды в команде HRC должны иметь один и тот же план исполнения;
* Все члены команды в команде HRC должны знать контекст среды совместной работы;
* Команда HRC должна иметь структурированные способы общения.

В данной работе основное внимание уделяется именно третьему пункту, то есть структурированным способам коммуникации.

Жесты – один из способов общения. Мимика, жесты рук и позы тела являются эффективными каналами связи в сотрудничестве между людьми [23, 8]. Жесты могут быть разделены на три типа [42]:

* Жесты тела (действия или движения всего тела);
* Жесты рук и кистей рук (позы рук, жесты рук);
* Головные и лицевые жесты (кивания или вращения головы, подмигивания, губы);

Распознавание жеста относится к математической интерпретации человеческих движений вычислительным устройством. Чтобы взаимодействовать с человеком, машины должны правильно понимать человеческие жесты и действовать в соответствии с жестом в достаточной степени точности. В среде HRC должен быть доступен естественный способ коммуникации между роботами и людьми.

1.1. Распознавание жестов для человеко-машинного взаимодействия

Чтобы распознать жесты в контексте человеко-машинного взаимодействия, полезно исследовать общую и упрощённую модель обработки информации. Как показано на рис. 1, Parasuraman представил [45] обобщённую модель обработки информации, состоящую из четырёх этапов.

|  |
| --- |
|  |
| Рис. 1. Четырёхступенчатая модель обработки информации [45]. |

На основе общей модели (рис. 1), предлагается конкретная модель распознавания жестов в человеко-машинном взаимодействии. Как показано на рис. 2, существует пять основных частей, связанных с распознаванием жестов для HRC: сбор данных с датчиков, идентификация жестов, отслеживание жестов, классификация жестов и отображение жестов.

|  |
| --- |
|  |
| Рис. 2. Модель процесса распознавания жестов при взаимодействии человека и машины. |

Поясним каждый из этапов:

* Сбор данных с датчиков: жестовые данные фиксируются датчиками;
* Идентификация жеста: в каждом фрейме жест отображается из необработанных данных;
* Отслеживание жеста: локализованный жест отслеживается во время движения. Для статических жестов отслеживание жестов не требуется;
* Классификация жестов: отслеживание движения жестов классифицируется в соответствии с заранее определёнными типами жестов;
* Отображение жеста: результат распознавания жестов преобразуется в управляющие команды.

1.2. Сенсорные технологии

Перед процессом распознавания жестов рабочие данные должны быть собраны датчиками. Существуют две основные категории сбора данных: основанные на изображениях и не основанные на изображениях. Рассмотрим их подробнее.

1.2.1. Методы на основе получения изображений

Технологии часто вдохновляются природой: человек использует глаза, что распознавать жесты, поэтому для машин разумно использовать камеры, что «видеть» жесты. Подходы, основанные на работе с изображениями, составляют четыре категории:

* *Маркер*. В маркерном подходе датчиком является оптическая камера. В большинстве решений на основе маркеров пользователям необходимо носить видимые маркеры.
* *Камера*. В начале 90-х годов исследователи начали анализировать жесты при помощь одиночной камеры [57, 58]. Недостатком однокамерного подхода является ограничение угла обзора, что влияет на надёжность системы [30]. Однако в недавних исследованиях применялся однокамерный подход к высокоскоростному распознаванию жестов [34]. Система использует датчик скорости и специально разработанный процессор визуальных вычислений для достижения высокоскоростного распознавания жестов.
* *Стереокамера*. Чтобы добиться надёжного распознавания жестов, исследователи предложили стереоподход к созданию 3D-зрения. Здесь определяются стереокамерных подходы как приложения, которые используют две оптические камеры для создания информации о глубине 3D. Многие стереоподходы к камерам соответствовали аналогичному процессу [17, 41]. Хотя системы стереокамер улучшили устойчивость во внешней среде, они по-прежнему имеют таким проблемы, как сложность вычислений и трудности с калибровкой [64].
* *Сенсор глубины*. В последнее время технологии глубокого зондирования быстро развиваются. Датчики глубины обладают рядом преимуществ по сравнению с традиционными стереокамерами. Например, можно предотвратить проблемы настройки калибровки и условий освещения [59]. Кроме того, выходной информацией датчика глубины является информация о глубине 3D, по сравнению с информацией о цвете, эта информация упрощает проблему идентификации жестов [42].

1.2.2. Методы, основанные не на изображениях

В распознавании жестов в течение долгого времени доминировали датчики на основе изображений. Недавние разработки в микроэлектромеханических системах (МЭМС) и сенсорных технологиях значительно улучшили технологии распознавания жестов, основанные не на изображениях. Рассмотрим основные из них:

* *Перчатки*. Жестовые интерфейсы на основе перчаток также используются для распознавания жестов. Обычно методы на основе перчаток требуют проводного подключения акселерометров и гироскопов. Однако громоздкая перчатка с проводами может вызывать проблемы в человеко-машинном взаимодействии, и данные подходы имеют сложности в процедурах калибровки и настройки.
* *Браслет (носимая электроника)*. Другая бесконтактная технология использует сенсоры на браслетах. Сенсоры установлены на браслете или аналогичных носимых устройствах. Сенсоры на браслетах позволяют использовать беспроводные технологии и датчики электромиограммы, что позволяет избежать подключения кабелей. Сенсоры должны контактировать с запястьем, при этом руки и пальцы пользователя могут быть свободны.
* *Бесконтактные устройства*. В третьем типе технологий, не связанных с изображениями, используются датчики, не предназначенные для ношения. Бесконтактные датчики могут обнаруживать жесты без контакта с человеческим телом. Google представил Project Soli, системы радиолокационного контроля и распознавания жестов на радиочастотном спектре (радар) [22]. Устройство способно распознавать разные жесты рук на небольшом расстоянии. В течение многих лет MIT является ведущим новатором в области распознавания жестов. Технология электрического поля была впервые разработана в MIT [56]. Недавно Adib [3, 4, 2] из MIT представил систему WiTrack и RF-Capture, которая отслеживает движение пользователя по радиочастотным сигналам, отражённым от человеческого тела. Система способна захватывать человеческие жесты даже из другой комнаты через стену с точностью до 20 см. Таким образом, технологии, не пригодные для ношения, являются перспективными и быстрорастущими сенсорными технологиями для распознавания жестов.

1.2.3. Сравнение сенсорных технологий

Таблица 1 содержит сравнение различных сенсорных технологий, показаны преимущества и недостатки различных подходов. Понятно, что ни один датчик не подходит для всех приложений. На основе вышеуказанных методов можно выделить два вида сенсорных технологий:

* Сенсоры для использования внутри помещений: датчики глубины являются наиболее перспективными технологиями на основе изображений, они обладают преимуществами простоты калибровки, установки и скорости обработки данных.
* Бесконтактные сенсоры являются наиболее перспективной технологией среди подходов, основанных на не изображении. Они не требуют прямого контакта с пользователями.

|  |  |  |
| --- | --- | --- |
| Таблица 1. | | |
| Достоинства и недостатки различных сенсорных технологий. | | |
|  | Достоинства | Недостатки |
| Маркеры | Низкая вычислительная нагрузка | Маркеры на теле пользователя |
| Камера | Простота установки | Низкий уровень надёжности |
| Стереокамера | Надёжность | Сложность вычислений,  трудности с калибровкой |
| Камера глубины | Высокая частота кадров | Разрешение зависит от мощности света и отражения,  невозможно использовать на расстоянии более 2 метров. |
| Перчатка | Скорость отклика,  точность трекинга | Громоздкое устройство с  проводами |
| Браслет (носимая электроника) | Скорость отклика,  область действия | Браслет должен контактировать с человеческим телом |
| Бесконтактные устройства | Не требует контакта с телом | Низкое разрешение, технология недостаточно зрелая |

1.3. Идентификация жестов

Идентификация жеста – первый шаг в процессе распознавания жестов после получения необработанных данных, полученных с датчиков. Идентификация жеста означает обнаружение жестовой информации и сегментацию соответствующей жестовой информации из необработанных данных. Популярные технологии для решения проблемы идентификации жестов основаны на визуальных особенностях, алгоритмах обучения и человеческих моделях.

1.3.1. Визуальные характеристики

Человеческие руки и тело обладают уникальными визуальными особенностями. В распознавании жестов на основе изображений жесты состоят из фрагментов человеческих рук и/или тела. Поэтому использование таких визуальных признаков в идентификации жестов вполне обоснованно.

* *Цвет.* Цвет – это простая визуальная функция для идентификации жестов из фоновой информации. Однако на системы распознавания жестов на основе цветов сильно влияют освещение и тени в сложной среде человеко-машинного взаимодействия. Ещё одна распространённая проблема в обнаружении цвета кожи заключается в том, что цвет кожи человечка сильно различается среди человеческих рас. Из-за вышеперечисленных проблем, в современных подходах, цвет кожи рассматривается только как один из многих параметров при идентификации жестов.
* *Локальные признаки.* В распознавании жестов на основе изображения условия освещения сильно влияют на качество идентификации жестов. Поэтому многие исследователи используют метод локальных признаков, которые не чувствителен к условиям освещения. Локальный подход к объектам – это детализированный подход на основе текстур. Он раскладывает изображение на более мелкие области, которые не соответствуют частям тела. Одной из наиболее важных локальных функций является преобразование признаков инвариантных объектов (SIFT) [40]. Метод SIFT является вращательным, трансляционным, масштабируемым и частично осветляющим инвариантом. Существует несколько подобных методов локальных признаков, например, SURF и ORB, предложенные в более поздние годы [9, 50]. Как правильно, подходы к локальным особенностям также рассматриваются только как один из множества параметров при идентификации жестов. Несколько методов идентификации, таких как методы формы и контура, методы движения и методы обучения, основаны на локальных признаках.
* *Форма и контур.* Другой интуитивный и простой способ идентификации жестов – использование уникальной формы и контуры человеческого тела в среде человеко-машинного взаимодействия. Существенный вклад в определение формы и соответствия был внесён Belongie и другими в [10]. Они ввели метод дескриптора контекста формы. Они ввели метод дескриптора контекста формы. Дескриптор контекста формы используется для обнаружения похожих фигур на разных изображениях. Разработка датчика глубины даёт возможность точно измерять форму поверхности. 3D-модели, созданные на основе технологий, позволяют очень детально представлять форму человеческого тела [5].
* *Движение.* В конкретной среде человеко-машинного взаимодействия человек является единственным движущимся объектов в массиве необработанных данных. Поэтому движение является полезной функцией для обнаружения человеческих жестов. Оптических поток является ключевой технологией идентификации жестов на основе движения. Оптический поток не нуждается в вычитании фона, что является преимуществом по сравнению с подходами на основе формы и контура. Дарал и Турау [61] представили знаменитый метод гистограмм ориентированных градиентов (HOG). Дескрипторы HOG делят изображения кадры изображения на блоки, для каждого из которых вычисляется гистограмма. Среди подходов, основанных не на изображении, распознавание жестов на основе движения является популярным методом. Пороговые значения и фильтрация обычно применяются к необработанным данным с датчика для идентификации человеческих жестов.

1.3.2. Алгоритмы обучения

Недавняя тенденция идентификации жестов заключается в применении алгоритмов обучения, особенно для обнаружения статических жестов, которые могут быть представлены в одном кадре. Методы визуальной функции основаны на различных визуальных особенностях, в то время как алгоритмы обучения используют алгоритмы машинного обучения для определения жестов из необработанных данных датчика. Хотя некоторые алгоритмы основаны на методах визуальных признаков, удаление фона изображения не всегда необходимо для корректной работы алгоритмов. В системах распознавания жестов широко применяются алгоритмы обучения, такие как метод опорных векторов (SVM, Support-Vector Machine), искусственные нейронные сети (ANN, Artificial Neural Network) и случайные решения (RDF, Random Decision Forest).

1.3.3. Модель человека

В отличие от вышеупомянутых подходов, модельных подход использует явную модель человеческого тела для восстановления позы человеческого тела. Поэтому модельный подход также называется генеративным подходом. Поскольку идентификация жестов на основе модели человеческого тела даёт преимущество упрощения процесса классификации жестов, подход с использованием модели человека стал популярным решением при работе с датчиками глубины. Современные исследования идентификации жестов на основе модели человека можно разделить на два типа: идентификация модели рук и идентификация модели скелета тела.

* *Модель руки*. Существует три основных подхода к моделированию ручных моделей: модель формы, трёхмерная модель и модель ручного скелета. Подход, основанные на форме, соответствует предварительно построенной формы руки из наблюдений. Подход 3D-модель интерпретирует проблему обнаружения руки как проблему оптимизации, которая минимизирует различия между моделью 3D-руки и наблюдаемой рукой [44].
* *Модель телесного скелета*. Чтобы определить жест тела, подробная человеческая модель бесполезна. Однако, в задаче идентификации жестов тела, модель именно скелета тела является популярным подходом. Модель скелета тела представляет собой упрощённую модель человеческого тела, которая сохраняет информацию о суставах тела. Большая часть литературы по алгоритмом для обработки скелета тела основана на информации о глубине, собранной с датчиков глубины.

1.3.4. Выводы по подходам идентификации жестов

В таблице 2 представлены наиболее подходящие способы идентификации жестов для каждого популярного датчика.

|  |  |
| --- | --- |
| Таблица 2. | |
| Методы идентификации жестов для разных жестов. | |
| Датчик | Метод идентификации жестов |
| Системы с одной камерой | В системах с одной камерой могут быть реализованы методы визуальных признаков и алгоритмы обучения. Для обеспечения надежной работы следует применять алгоритмы обучения. Чтобы добиться более быстрой обработки изображений, следует применять метод визуальных признаков. |
| Датчик глубины | Поскольку метод человеческой модели использует и упрощает информацию из облака точек, метод человеческой модели является лучшим вариантом для систем с датчиками глубины. |
| Браслет (сенсоры на базе браслета) | Никакие визуальные методы не могут применяться для систем на базе браслетов. Обычно собранные данные требуют базовой фильтрации для идентификации жестов, а алгоритмы обучения будут реализованы на более поздней стадии классификации жестов. |
| Бесконтактные датчики | Бесконтактные датчики также принимают данные сигналов вместо изображений. Из-за того, что радиочастотные сигналы содержат шумы, современные системы фильтрации и обработки должны быть реализованы в бесконтактных системах. |

Из-за характера человеко-машинного взаимодействия работники-люди являются наиболее важными его членами. Несмотря на понимание жестов человеческого тела, подход моделирования человеческого тела также будет контролировать человеческие движения, что обеспечивает безопасную среду для совместной работы роботов и людей. Как упоминалось ранее, модель скелета тела упрощает телесную модель человека, а информация о суставах тела легко извлекается. Более того, скелетные подходы могут упростить последующий процесс классификации жестов. Поэтому в настоящее время модели с использованием скелетного подхода является подходящим решением для распознавания жестов в системах человеко-машинного взаимодействия.

1.4. Отслеживание жестов

При распознавании жестов понятие отслеживания используется по-разному в разной литературе. В данной работе понятие отслеживания определяется как процесс поиска временных соответствий между кадрами. В частности, производится фокусировка на проблеме отслеживания жестов, которая связывает идентифицированный жест в предыдущих кадрах с текущим фреймом. Для статических жестов, то есть жестов, которые могут быть представлены одним кадром, отслеживание жестов не требуется.

1.4.1. Отслеживание единичной гипотезы

Отслеживание единичной гипотезы относится к оценке наилучшего соответствия при согласовании с минимальной ошибкой. Поэтому при использовании единичной гипотезы отслеживания жест представлен только одной гипотезой. Большинство усовершенствованных алгоритмов отслеживания основаны на технологиях единичной гипотезы.

* *Средний сдвиг*. Отслеживание методом среднего сдвига – это базовая технология отслеживания. Метод выполняет сопоставление с гистограммами цвета RGB [15]. Для каждого нового кадра средний сдвигающий трекер сравнивает расстояние Бхаттачария между гистограммами целевого окна нового кадра и кадрами старого кадра. Полное математическое объяснение этого подхода можно найти в [15].
* *Фильтр Калмана*. Фильтр Калмана (KF) представляет собой рекурсивный алгоритм реального времени, используемый для оптимальной оценки лежащих в основе состояний ряда шумовых и неточных результатов измерений, наблюдаемых с течением времени. Полный математический вывод KF представлен в [33]. В настоящее время KF активно развивается и применяется в различных областях, таких как аэрокосмическая промышленность, робототехника и экономика.
* *Расширенные фильтры Калмана*. KF предполагает, что вектор состояния является линейной моделью. Расширенный фильтр Калмана (EKF, Extended Kalman Filter) является функциональным алгоритмом отслеживания, даже если модель нелинейна [25]. Другим алгоритмом, который решает ту же проблему под другим углом, является сигма-точечный фильтр Калмана (UKF, Unscented Kalman Filter) [65]. UKF решает проблему, применяя детерминированный подход с взвешенной выборкой. Распределение состояний представлено с использованием минимального набора выбранных точек выборки.

1.4.2. Расширенные методы отслеживания

За последнее время появилось много продвинутых методов отслеживания. Некоторые из этих передовых методов использовали часть алгоритмов отслеживания, упомянутых выше. Другие методы улучшили отслеживаемость с помощью алгоритмов обнаружения или обучения.

* *Расширенное отслеживание модели*. Для долгосрочных проблем отслеживания многие алгоритмы отслеживания терпят неудачу, поскольку цель поддерживает фиксированные модели. Расширенное отслеживание модели сохраняет целевое поведение или внешний вид из последних кадров изображения, поэтому для целевой оценки зарезервировано больше целевой информации.
* *Отслеживание путём обнаружения*. Ещё один вид алгоритмов отслеживания строится на базе алгоритмов обучения идентификации жестов. Для этих алгоритмов отслеживания в кадрах изображений применяется классификатор или детектор, чтобы идентифицировать жест из справочной информации [36]. Одним из характерных подходов является трекер отслеживания, обучения и обнаружения. Этот подход объединяет результаты детектора объекта с устройством оптического отслеживания потока. Ещё одна типичная технология отслеживания по обнаружения – применение множественного обучения экземпляров [7]. Алгоритм обучения может повысить надёжность трекера и уменьшить количество параметров.

1.4.3. Сравнение различных подходов отслеживания жестов

Smeulders и другие [55] представил результат теста различных алгоритмов отслеживания жестов. Итоговый результат является оценкой нормализованного F-балла. F-балл даёт представление о среднем охвате ограничивающего блока объекта и истинной ограничивающей рамки, поэтому чем выше F-оценка, тем лучше качество отслеживания. На рис. 3 представлены результаты испытаний на разных видеофрагментах.

|  |
| --- |
|  |
| Рис. 3. Результат теста алгоритмов отслеживания на разных видеофрагментах [55]. |

Kalman Appearance Tracker (трекер появления Калмана) и Mean Shift Tracker (трекер среднего сдвига) относятся к трекеру единичной гипотезы. Отслеживание с помощью Sampling Tracker (трекер отбора) и Incremental Visual Tracker (инкрементный визуальный трекер) относится к расширенной модели отслеживания. Трекеры отслеживания (Tracking, Learning and Detection Tracker) и трекер обнаружения нескольких экземпляров (Multiple Instance Learning Tracker) относятся к методам отслеживания с помощью обнаружения. Легко заметить, что трекеры одиночной гипотезы менее производительны, чем остальные. Таблица 3 содержит резюме различных подходов к отслеживанию жестов.

|  |  |
| --- | --- |
| Таблица 3. | |
| Описание основных подходов к отслеживанию. | |
| Подход | Описание |
| Единичная гипотеза | Быстрый и простой алгоритм. Подходит для отслеживания одного жеста в контролируемой среде. |
| Множественные  гипотезы | Возможность отслеживания нескольких целей одновременно. Подходит для отслеживания нескольких жестов в контролируемой среде. |
| Расширенная модель трекинга | Целевая история сохраняется и доступна для целевой оценки. Подходит для долговременного отслеживания жестов. |
| Отслеживание путём обнаружения | Алгоритм обучения повышает надежность и снижает уровень шума. Этот комбинированный подход имеет предпочтительную производительность в тестах. Подходит для отслеживания жестов в сложной среде. |

1.5. Классификация жестов

Классификация жестов – это последний и самый важный шаг в распознавании жестов. Большинство человеческих жестов – это динамические жесты. Один динамический жест всегда состоит из нескольких кадров. Чтобы классифицировать динамические жесты, классификация жестов должна выполняться после или вместе с отслеживанием жестов. Рассмотрим основные методы классификации жестов.

* *Метод K-ближайших соседей (K-Nearest Neighbours)*. Является фундаментальным и базовым алгоритмом классификации жестов, который классифицирует входные данные в соответствии с ближайшими примерами обучения [46].
* *Скрытая Марковская модель.* Является комбинацией ненаблюдаемой цепи Маркова и стохастического процесса.
* *Метод опорных векторов.* Метод опорных векторов является дискриминационным классификатором, определяемым разделительной гиперплоскостью [**Ошибка! Источник ссылки не найден.**, 53]. Границы решения классификации определяются путём максимизации расстояния от границы. Оптимальная разделительная гиперплоскость максимизирует запас обучающих данных. Примеры обучения, наиболее близкие к оптимальной гиперплоскости, называются вспомогательными векторами. Общей проблемой для данного метода является то, что число опорных векторов растёт линейно с размером обучающего набора.
* *Метод ансамбля.* Основное предположение ансамблевого метода состоит в том, что ансамбли более точны, чем слабые отдельные классификаторы. Одним из известных методов ансамбля является бустинг [52, 51]. Алгоритм бустинга начинается с нескольких слабых классификаторов. Слабые классификаторы применяются многократно. В обучающей итерации часть обучающих образцов используется в качестве входных данных. После итерации обучения создается новая граница классификации. После всех итераций алгоритм бустинга объединяет эти границы и сливается в одну конечную границу предсказания. Одним из известных ансамблевых методов является алгоритм Адабуст (Adaboost). Существенным преимуществом алгоритма Адабуст является то, что Адабуст не нуждается в большом количестве обучающих данных.
* *Динамическое деформирование временных рядов (Dynamic Time Warping, DTW).* Является оптимальным алгоритмом выравнивания для двух последовательностей. DTW генерирует кумулятивную матрицу расстояний, которая искажает последовательности нелинейным способом. Первоначально DTW использовался для распознавания речи, но в последнее время активно используется в распознавании жестов [12, 6]
* *Искусственные нейронные сети (ANN).* Являются семейством моделей обработки информации, основанных на биологических нейронных сетях. ANN состоит из множества взаимосвязанных обрабатывающих нейронов, которые работают параллельно. Каждый нейрон получает входные данные, обрабатывает их и даёт выходные данные. ANN может использоваться для оценки функций, зависящих от большого количества входных данных.
* *Глубокое обучение (Deep Learning).* Является быстро развивающейся отраслью машинного обучения. Глубокое обучение позволяет компьютеру моделировать данные с абстракциями высокого уровня, используя множественную нейронную сеть на уровне обработки. Более того, в отличие от традиционных алгоритмов обучения, глубокое обучение не требует ручной подготовки данных, что позволяет использовать преимущества экспоненциально увеличивающихся объёмов доступных данных и вычислительных мощностей.

В таблице 4 представлены преимущества и недостатки подходов к классификации жестов.

|  |  |  |
| --- | --- | --- |
| Таблица 4. | | |
| Преимущества и недостатки подходов к классификации жестов. | | |
| Подход | Преимущества | Недостатки |
| Метод K-ближайших  соседей | Простота | K-параметр следует выбирать осторожно |
| Скрытая  Марковская модель | Гибкость обучения и проверки, прозрачность модели | Необходимо отрегулировать множество свободных параметров |
| Метод опорных векторов | Могут применяться различные функции ядра | Число опорных векторов растёт линейно с размеров обучающего набора |
| Метод  ансамбля | Не требуется большого количество данных обучения | Легко переобучить, чувствительность к шумам и выбросам |
| Динамическое деформирование временных рядов | Надёжное нелинейное выравнивание между образцами | Сложность времени и по объёму данных |
| Искусственные нейронные сети | Может обнаруживать сложные нелинейные зависимости между переменными | Принцип «чёрного ящика», не может использоваться при наличии небольшого набора данных для обучения |
| Глубокое  обучение | Не нуждается в хорошей подготовке признаков, превосходит другие методы машинного обучения | Требуется большое количество обучающих данных и значительные вычислительные мощности |

Одной из тенденций является подход глубокого обучения. Основными ограничениями глубокого обучения являются ограниченные вычислительные мощности. Однако экспоненциально возрастающая вычислительная мощность может легко решить эту проблему. Количество приложений классификации жестов, основанных на глубоком обучении, быстро растет. Другая тенденция заключается в объединении различных алгоритмов классификации. Каждый алгоритм классификации имеет свои преимущества и недостатки. Чтобы использовать это, различные классификаторы могут быть объединены для достижения лучшей производительности. Также было замечено, что важно координировать алгоритмы классификации жестов с помощью алгоритмов идентификации и трекинга жестов.

1.6. Примеры систем распознавания жестов

Одним из примеров распознавания жеста при помощи вычисления расстояние между точками является «Система распознавания жестов в реальном времени с цветными перчатками» (Real-Time Hand-Tracking with a Color Glove), которую создали Роберт Ванг (Robert Y. Wang) и Йован Попович (Jovan Popovic) из лаборатории компьютерных наук и искусственного интеллекта Массачусетского технологического института [66].

Распознавание жестов возможно при надевании цветных перчаток, по которым программа распознаёт жесты рук при помощи веб-камеры по наличию цветных пикселей на изображении. После обнаружения этих пикселей происходит вычисление расстояния между ними. Вычислив это расстояние, программа выдаёт информацию о распознанном жесте. Вид цветных перчаток изображён на рисунке 4.

|  |
| --- |
|  |
| Рис. 4. Перчатки, при помощи которых программа распознаёт жесты. |

Этот способ распознавания очень интересен в своей реализации, но он имеет огромный недостаток: необходимость надевания специальных перчаток для того, чтобы программа смогла выделить цветные пиксели. Без них программа неработоспособна.

Следующим примером является распознавание жестов при помощи сигналов с тензодатчиков. Представителем этой группы являются перчатки GRASP – говорящие перчатки для глухонемых (Glove-base Recognition of Auslan using Simple Processing) [29].

Их принцип работы следующий: к компьютеру подключены две перчатки при помощи кабелей. Программа обрабатывает сигналы тензодатчиков и, в зависимости от информации, полученной с этих датчиков, выдаёт информацию о распознанном жесте. Внешний вид устройства приведён на рисунке 5.

|  |
| --- |
|  |
| Рис. 5. Перчатка, при помощи которой программа распознаёт жесты. |

Данный способ реализации так же имеет недостаток в необходимости надевания специальной перчатки, без которой программа не сможет распознать жест.

Еще одним примером реализации распознавания жестов является программа [1], преобразующая жесты русской ручной азбуки в текстовый вид, работает непосредственно с веб-камерой, при помощи которой определяется положение руки в кадре. После этого строится «скелет» человеческой руки, по которому программа распознаёт жест и переводит его в букву русской азбуки. Вид работы программы приведён на рисунке

6.

|  |
| --- |
| page8image2919337056 |
| Рис. 6. Автоматическое преобразование жестов русской ручной азбуки в текстовый вид. |

Эта программа является наиболее удобной в использовании, так как не требует наличия различных устройств, надетых на руку, но работает только с полутоновыми изображениями, то есть изображениями в оттенках серого. Такие изображения очень светочувствительны, что является большим недостатком этой программы.

2. НЕЙРОСЕТЕВЫЕ ТЕХНОЛОГИИ В ДЕТЕКТИРОВАНИИ И ИДЕНТИФИКАЦИИ КИСТИ ЧЕЛОВЕКА

Человеко-машинное взаимодействие (Human-computer interaction - HCI) – это междисциплинарное научное направление, изучающее взаимодействие между людьми и машинами. Предметом HCI является изучения, планирование и разработка методов взаимодействия человека с машиной, где в роли машины может выступать персональный компьютер, компьютерная система больших масштабов, система управления процессами и т.д. [16]. Под взаимодействием понимается любая коммуникация между человеком и машиной. Одним из методов HCI, получившим широкое распространение в последние годы, является взаимодействие, основанное на жестах человека [**Ошибка! Источник ссылки не найден.**].

Распознавание жестов на основе ключевых точек предоставляет не только гибкий способ классификации жеста, поскольку в этом случае жест определяется лишь последовательностью координат этих точек, но и дополнительными возможностями, таким как привязка расположения какого-либо объекта к расположению конкретной ключевой точки в кадре (относительном или абсолютном).

Использование одного из класса методов искусственного интеллекта – машинного обучения – для решения самых разнообразных задач приобретает всё более широкий спектр, особенно это характерно для тех задач, найти алгоритмическое решение которых либо не представляется возможным, либо оно неразумно и неэффективно. Машинное обучение характеризуется не прямым решением задачи, а обучением за счёт применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятности, теории графов и различные техники работы с данными в цифровой форме.

Подмножеством машинного обучения является глубокое обучение, являющееся совокупностью методов машинного обучения, основанных на обучении представлениям, а не специализированных алгоритмах под конкретные задачи, то есть на основе искусственных нейронных сетей (ИНС). Процесс обучения называется *глубоким*, так как структура ИНС состоит из нескольких входных, выходных и скрытых слоёв. Каждый слой содержит части, преобразующие входные данные в сведения, которые следующие слой может использовать для определённой задачи прогнозирования. Рассмотрим нейронные сети и их виды подробнее.

2.1. Нейронные сети и их виды

*Нейронная сеть* (или искусственная нейронная сеть, ИНС) – это математическая модель, а также её программная или аппаратная реализация, построенная по принципу организации и функционирования *биологических нейронных сетей* – сетей нервных клеток живого организма. Если же говорить проще, то нейронная сеть – это последовательность *нейронов*, соединённых между собой синапсами. На рис. 7а изображена схема простой нейросети, состоящей из трёх слоёв: зелёным обозначены входные нейроны, голубым – скрытые нейроны (скрытый слой) и жёлтым – выходной нейрон.

Математически, нейрон представляют как некоторую нелинейную функцию от единственного аргумента – линейной комбинации всех входных сигналов. Данную функцию называют *функцией активации* или *передаточной функцией*. На рис. 7б изображена схема нейрона, цифрой 1 обозначены нейроны, выходные сигналы которых поступают на вход данного, цифрой 2 – сумматор входных сигналов, цифрой 3 – вычислитель передаточной функции и цифрой 4 – нейроны, на входы которых подаётся выходной сигнал данного.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис. 7. Схема простой нейронной сети (а), искусственного нейрона (б). | |

Описать нейрон уравнением можно так:

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | (1) |
| где: | | | |
|  | | | |
| и |  | соответственно сигналы на входах и веса входов | |
|  |  | индуцированное локальное поле, | |
|  |  | передаточная функция. | |

Возможные значения сигналов на входах нейрона обычно считают заданными в отрезке . Дополнительный вход и соответствующий ему вес используются для инициализации нейрона. Под инициализацией понимается смещение активационной функции нейрона по горизонтальной оси, то есть формирование порога чувствительности нейрона. Иногда к выходу нейрона специально добавляют некую случайную величину, называемую сдвигом.

Передаточная функция нейрона определяет зависимость сигнала на выходе нейрона от взвешенной суммы сигналов на его входах. В большинстве случаев она является монотонно возрастающей, а область значений лежит в или . Нейрон полностью характеризуется своей передаточной функцией. Рассмотрим основные типы передаточных функций.

*Линейная передаточная функция*. Задаётся формулой:

Иногда возможен сдвиг функции по обеим осям. Одним из преимуществ является дифференцируемость на всей числовой оси.

*Пороговая передаточная функция*. Иногда называют *функцией Хевисайда*. Представляет собой резкий скачок на некоторый уровень . Задаётся формулой:

Здесь – сдвиг функции активации относительно горизонтальной оси. Недостаток – не дифференцируема на всей оси абсцисс.

*Сигмоидальная передаточная функция*. Является самой часто используемой функцией. Существует несколько видов её математической записи. *Логистическая функция*:

где t – параметр функции, определяющий её крутизну. Важным достоинством этой функции является простота её производной:

Особенностью нейронов с такой передаточной характеристикой является их усиление сильных сигналов существенно меньше, чем слабых. Это предотвращает насыщение от больших сигналов.

Другой математической записью сигмоидальной функции является *гиперболический тангенс*:

Она отличается от логистической функции тем, что её область значений лежит в интервале

*Радиально-базисная функция передачи*. Она принимает в качестве аргумента расстояние между входным вектором и некоторым наперёд заданным центром активационной функции. Значение этой функции тем выше, чем ближе входной вектор к центру. В качестве такой функции можно взять *функцию Гаусса*:

|  |  |  |
| --- | --- | --- |
|  | | |
| где: | | |
|  |  | расстояние между центром и вектором входных сигналов , |
|  |  | определяет скорость спада функции при удалении вектора от центра и называется *шириной окна*, |
|  |  | определяет сдвиг активационной функции по оси абсцисс. |

В качестве расстояния между векторами могут быть использованы различные метрики, обычно используется евклидово расстояние.

Также в качестве передаточных функций могут использоваться: *экспонента*, *тригонометрический синус*, *модуль, квадратичная*.

Каждая нейронная сеть включает в себя первый входной слой нейронов. Этот слой не выполняет каких-либо преобразований и вычислений, его задача состоит в том, чтобы принимать и распределять входные сигналы по остальным нейронам. Таким образом, этот слой единственный, являющийся общим для всех типов нейросетей, а критерием для деления является уже дальнейшая структура:

* + - 1. *Однослойная структура нейронной сети*. Представляет собой структуру взаимодействия нейронов, в которой сигналы со входного слоя сразу направляются на выходной слой, который не только преобразует сигнал, но и сразу же выдаёт ответ. Как уже было сказано, первый входной слой только принимает и распределяет сигналы, а нужные вычисления происходят уже во втором слое. Входные нейроны являются объединёнными с основным слоем с помощью синапсов с разными весами, обеспечивающими качество связей.
      2. *Многослойная нейронная сеть*. Данная сеть помимо выходного и входного слоёв имеет ещё несколько скрытых промежуточных слоёв. Число этих слоёв зависит от степени сложности нейронной сети. Соответствующие решения обладают большими возможностями, если сравнивать с однослойными, ведь в процессе обработки данных каждый промежуточный слой – это промежуточный этап, на котором осуществляется обработка и распределение информации.

Кроме количества слоёв, нейронные сети можно классифицировать по направлению распределения информации по синапсам между нейронами:

* + - 1. *Нейросети прямого распространения (однонаправленные)*. В этой структуре сигнал перемещается строго по направлению от входного слоя к выходному. Движение сигнала в обратном направлении не осуществляется и в принципе невозможно. Сегодня разработки этого плана распространены широко и успешно решают задачи распознавания образов, прогнозирования и кластеризации.
      2. *Рекуррентные нейронные сети (с обратными связями)*. Для данных типов сетей характерно движение сигнала и в прямом, и в обратном направлении. В итоге результат выхода способен возвращаться на вход. Выход нейрона определяется весовыми характеристиками и входными сигналами и дополняется предыдущими выходами, снова вернувшимися на вход. Этим нейросетям присуща функция кратковременной памяти, на основании чего сигналы восстанавливаются и дополняются во время их обработки.
      3. *Радиально-базисные функции*.
      4. *Самоорганизующиеся карты*.

Основными критериями выбора типов среди огромного многообразия искусственных нейронных сетей являются способность эффективно работать с изображениями и с последовательностями. Сверточная нейронная сеть удовлетворяет первому критерию, а рекуррентная – второму. Рассмотрим их поподробнее.

2.1.1. Свёрточная нейронная сеть

Свёрточная нейронная сеть (CNN – Convolutional Neural Network) – специальная архитектура нейронных сетей, предложенная Яном Лекуном в 1988 году и нацеленная на эффективное распознавание образов. Как следует из названия, вместо обычных слоёв в ней используется операция *свёртки*, суть которой состоит в том, что каждый фрагмент изображения умножается на матрицу (ядро) свёртки по элементно, а результат суммируется и записывается в аналогичную позицию выходного изображения.

Работа свёрточной нейронной сети обычно интерпретируется как переход от конкретных особенностей изображения к более абстрактным деталям, и далее к ещё более абстрактным деталям вплоть до выделения понятий высокого уровня. При этом сеть самонастраивается и вырабатывает сама необходимую иерархию абстрактных признаков (последовательности карт признаков), фильтруя маловажные детали и выделяя существенное.

Подобная интерпретация носит скорее метафорический или иллюстративный характер. Фактически «признаки», вырабатываемые сложной сетью, малопонятны и трудны для интерпретации настолько, что на практике суть этих признаков даже не пытаются понять, тем более «подправлять», а вместо этого для улучшения результатов распознавания меняют структуру и архитектуру сети. Так, игнорирование системой каких-то существенных явлений может говорить о том, что либо не хватает данных для обучения, либо структура сети обладает недостатками, и система не может выработать эффективных признаков для данных явлений.

Оператор свёртки с ядром , получая на вход изображение , на выходе формирует матрицу :

Для многоканального изображения (или просто трёхмерной матрицы) используется трёхмерное ядро свёртки :

Так же обычно в свёрточную нейронную сеть включаются оператор пулинга (MaxPool), для которого определяются два параметра: размер ядра () и сдвиг (:

Пример свёрточной сети изображён на рис. 8.

|  |
| --- |
|  |
| Рис. 8. Пример свёрточной сети. |

Вначале применяются разных ядер свёртки размерности , получая из матрицы трёхмерную матрицу . Затем используется оператор пулинга , который уменьшает матрицу в два раза, сводя её к размеру . Ещё разных ядер свёртки той же размерности и слой пулинга «сжимают» матрицу до размера , получая величин, которые подаются на выходной слой с нейронами.

2.1.2. Рекуррентные нейронные сети

Рекуррентные нейронные сети (RNN – Recurrent Neural Network) – вид нейронных сетей, где связи между элементами образуют направленную последовательность. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки. В отличие от многослойных перцептронов, рекуррентные сети могут использовать свою внутреннюю память для обработки последовательностей *произвольной длины*, поэтому сети RNN применимы в таких задачах, где нечто целостное разбито на части, например: распознавание рукописного текста, распознавание речи или распознавание жестов.

Любая рекуррентная нейронная сеть имеет форму цепочки повторяющихся модулей нейронной сети. В обычной RNN структура одного такого модуля очень проста, например, он может представлять собой один слой с функцией активации (гиперболический тангенс) (рис. 9).

|  |
| --- |
|  |
| Рис. 9. Структура обычной RNN. |

Одна из привлекательных идей RNN состоит в том, что они потенциально имеют возможность связывать предыдущую информацию с текущей задачей, так, например, знания о предыдущем кадре видео могут помочь в понимании текущего кадра. Однако предоставление такой возможности зависит от разрыва между актуальной информацией и точкой её применения. По мере роста этого расстояния, RNN теряют способность связывать информацию. Однако модификация RNN LSTM не имеет такой проблемы.

Долгая краткосрочная память (Long Short-Term Memory – LSTM) – особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучения долговременным зависимостям. Данный вид был разработан в 1997 году специально для решения проблемы долговременной зависимости, то есть они способны запоминать информацию на долгие периоды времени [28].

Структура LSTM по аналогии с RNN также напоминает цепочку, но её модули выглядят иначе (рис. 10): модули имеют четыре особым образом взаимодействующих между собой слоя вместо одного.

|  |
| --- |
|  |
| Рис. 10. Структура LSTM. |

Ключевой компонент LSTM – это состояние ячейки (cell state) – горизонтальная линия, проходящая по верхней части схемы (рис. 11).

|  |
| --- |
|  |
| Рис. 11. Состояние ячейки в LSTM. |

Состояние ячейки напоминает ленту на конвейере, проходя напрямую через всю цепочку, оно участвует лишь в нескольких линейных преобразованиях, то есть информация «течёт» по ней, не подвергаясь изменениям. Тем не менее, LSTM может удалять информацию из состояния ячейки, посредством регулирования структурами, называемыми фильтрами (gates). Фильтры позволяют пропускать информацию на основании некоторых условий и состоят из слоя сигмоидальной нейронной сети и операции поточечного умножения. Сигмоидальный слой имеет диапазон выходных значений, лежащий в отрезке , который обозначает, какую долю каждого блока информации следует пропустить дальше по сети. Ноль в данном случае означает «не пропускать ничего», единица – «пропустить всё». В LSTM три таких фильтра, позволяющих защищать и контролировать состояние ячейки.

Первый шаг в LSTM – определить, какую информацию можно выбросить из состояния ячейки. Это решение принимает сигмоидальный слой, называемый «слоем фильтра забывания» (forget gate layer) (рис. 12а). В зависимости от и этот слой на основе выхода сигмоидального слоя принимает решение о сохранении каждого числа из состояния ячейки . Следующий шаг – решить, какая новая информация будет храниться в состоянии ячейки. Этот этап состоит из двух частей. Сначала сигмоидальный слой под названием «слой входного фильтра» (input layer gate) (рис. 12б) определяет, какие значения следует обновить. Затем -слой строит вектор новых значений-кандидатов , которые можно добавить в состояние ячейки. И наконец старой состояние ячейки заменяется на новое состояние , умножая его на и прибавляя . Это новые значения-кандидаты, умноженные на – на сколько необходимо обновить каждое из значений состояния. Последним этапом работы LSTM сети является отбор информации, получаемой на выходе. Выходные данные будут основаны на применении сигмоидального слоя и текущем состоянии ячейки, пройденном через -слой.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
|  |  |
| в) | г) |
| Рис. 12. Фильтры в LSTM: «слой фильтра забывания» (a), «слой входного фильтра» (б), замена старого состояния на новое (в), получение информации на выход (г). | |

Таким образом, хорошая приспособленность к обучению на задачах классификации и относительная невосприимчивость к длительности временных разрывов даёт LSTM преимущество по отношению к альтернативным рекуррентным нейронным сетям.

2.2. Детекторы, использующие ограничивающие прямоугольники

Когда нейронные сети в компьютерном зрении не применялись так широко, для детектирования объектов на изображении использовался достаточно простой подход. Тренировался бинарный классификатор, который мог определить есть ли на картинке объект определённого типа или нет. После того как классификатор был натренирован, по изображению «проходило» скользящее окно (sliding window) и для каждого прямоугольника, обрамлённого этим окном, применялся этот бинарный классификатор. Поскольку на одном изображении объекты могут быть и маленькими, и большими, то необходимо было либо запускать окно разных размеров, либо из исходного изображения строить пирамиду разных масштабов и запускать окно по каждому слою пирамиды. Основная задача была найти такие *особенности (features)*, которые бы позволили натренировать хороший бинарный классификатор.

С внедрением свёрточных нейронных сетей произошёл скачок в поиске хороших особенностей. Например, уже в 2013 появилась статья [54], про сеть Overfeat, где подход со скользящим окном обобщался на свёрточную нейронную сеть, применяемую в качестве классификатора.

Затем появилась линейка R-CNN, Fast R-CNN, Faster R-CNN в статьях [21], [20], [49], и расширения этого подхода. Здесь, вместо скользящего окна, используется предварительный детектор для поиска объектов, а затем сеть определяет класс объекта и уточняет его прямоугольник. Точность у данных детекторов довольно высока, но они всё ещё довольно медленные.

Наконец, третий подход появился в статье [48] – YOLO (You Only Look Once) и потом продолжился в [37] – SSD (Single Shot Detector). Для ускорения процесса детектирования в этих работах предлагается не искать объекты, пробегая по изображению скользящим окном или каким-то вариантом поискового алгоритма, а выбрать некоторое фиксированное количество прямоугольников и для каждого проверить наличие в нём объекта, и, если объект найден, то определить какого он класса и уточнить ограничивающую его рамку (*bounding box*). Например, в [48] на изображение накладывается решётка, в каждой клетке которой проверяется наличие центра какого-нибудь объекта. Рассмотрим данный подход подробнее.

2.2.1. YOLO

Авторы [48] предложили вместо того, чтобы искать объекты с последующей их проверкой классификатором, разбить изображение равномерной сеткой и для каждой ячейки определять есть ли в ней центр какого-то объекта, какого он класса и его точный прямоугольник. То есть для каждой ячейки определять:

1. Набор из пятёрок вида: , где относительные координаты центра объекта внутри ячейки, размеры объекта, уверенность сети, что объект есть и его прямоугольник задан правильно;
2. вероятностей (по количеству классов объектов), что объект с центром в данной ячейке принадлежит соответствующему классу.

Итак, получая на вход изображение, на выходе сеть должна выдавать тензор размера . В общем виде, схема сети представлена на рис. 13.

|  |
| --- |
|  |
| Рис. 13. Схема сети YOLO. |

Авторы предложили:

1. Подавать на вход изображения размером ;
2. Разбивать их при помощи решётки с ;
3. Для каждой ячейки выдавать предсказанных прямоугольника;
4. Так как в качестве набора данных для обучения берётся Pascal VOC, в котором размечено 20 классов, то .

Таким образом, на выходе сети будет тензор . В качестве базовой сети для выделения особенностей авторы берут GoogLeNet с некоторыми упрощениями и в двух вариантах: свёрточных слоя или для особо быстрого детектирования.

2.2.3. SSD

YOLO показал неплохие результаты: скорость выросла, а качество упало не очень сильно. Поэтому авторы статьи [37] решили улучшить этот подход, повысив качество, при этом не снижая скорости.

Было предложено, так же, как и в YOLO, не искать объекты на фотографии, а задать фиксированный набор прямоугольников и проверять наличие объекта в каждом из них. Но, поскольку в YOLO претендентов было слишком мало, и к тому же особенности брались с последнего свёрточного слоя сети, а значит мелкие объекты могли в такой ситуации потеряться, то было решено брать особенности с разных свёрточных слоёв, что привело к увеличению набора претендентов. Для того, чтобы не заставлять сеть тренироваться, отдавая для каждого объекта ровно одно обнаружение, дополнительно после детектирования применяется NMS (Non-Maximum Suppression). На рисунке 14 изображено сравнение моделей SSD и YOLO. В SSD модель в конце базовой сети добавляются слоёв, предсказывающих смещения к *анкерам* (default box) различных масштабов и соотношений сторон и связанные с ними вероятности. SSD с размером входного изображения значительно превосходит аналог YOLO по точности и по скорости.

|  |
| --- |
|  |
| Рис. 14. Сравнение моделей детекторов SSD и YOLO. |

Для работы SSD необходимо только входное изображение и рамки истинности для каждого объекта во время обучения (рис. 15а). С помощью свертки оценивается небольшой набор (например, ) полей по умолчанию с разными соотношениями сторон в каждом месте на нескольких картах признаков с разными масштабами (например, (рис. 15б) и (рис. 15в)). Для каждой рамки по умолчанию прогнозируется как смещение формы, так и достоверность для всех категорий объектов . Во время обучения сначала сопоставляются эти рамки по умолчанию с рамками истинности. Например, совпали две рамки по умолчанию с котом и одно с собакой на рис. 15а, которые рассматриваются как положительные, а остальные – как отрицательные. Потеря модели представляет собой взвешенную сумму между потерей локализации (Smooth L1 [21]) и потерей достоверности (например, Softmax).

|  |  |  |
| --- | --- | --- |
|  |  |  |
| а) | б) | в) |
| Рис. 15. Пример работы SSD детектора. | | |

Как уже было сказано, подход SSD основан на свёрточной сети с прямой связью, которая создаёт набор ограничивающих рамок (прямоугольников) фиксированного размера и оценивает наличие экземпляров класса объектов в этих рамках с последующим подавлением немаксимумов для получения результатов окончательных обнаружений. Модель SSD изображена на рис. 16.

|  |
| --- |
|  |
| Рис. 16. Подробное описание модели SSD. |

Первые слои нейросети основаны на стандартной архитектуре VGG-16, используемой для высококачественной классификации изображений (*базовая сеть)*. Затем в сеть добавляется вспомогательная структура для создания обнаружений со следующими ключевыми особенностями:

1. *Мультимасштабные карты ключевых точек для детектирования*.

В конец усечённой базовой сети добавляются свёрточные слои ключевых точек. Эти слои постепенно уменьшаются в размере и позволяют прогнозировать обнаружение в нескольких масштабах. Свёрточная модель для прогнозирования обнаружений различается для каждого слоя объектов (Overfeat [54] и YOLO [48]), которые работают с картой признаков в одном масштабе;

1. *Свёрточные «предсказатели» для детектирования.*

Каждый добавленный слой признаков (или, возможно, существующий слой признаков из базовой сети) может создавать фиксированный набор прогнозов обнаружения с использованием набора свёрточных фильтров. Для слоя признаков размера с каналами базовым элементом для прогнозирования параметров потенциального обнаружения является небольшое ядро , которое даёт либо оценку для категории, либо смещение формы относительно координат анкера. В каждом из положений, где ядро применяется, оно выдает выходное значение. Выходные значения смещения прямоугольника измеряются относительно положения анкера относительно каждого местоположения карты признаков (архитектура YOLO использует промежуточный полносвязный слой вместо свёрточного фильтра для этого шага).

1. *Анкеры (*или *default boxes).*

Набор анкеров связывается с каждой ячейкой карты признаков для нескольких карт признаков в начале нейросети. Анкеры разбивают карту признаков свёрткой так, что положение каждой ячейки относительно соответствующей ячейки фиксировано. В каждой ячейке карты признаков прогнозируется смещение относительно форм анкеров в ячейке, а также оценки для каждого класса, которые указывают на присутствие экземпляра класса в каждой из этих рамок. В частности, для каждой рамки из в заданном месте вычисляются оценки класса и 4 смещения относительно исходной формы анкера. В результате получается фильтров, которые применяются вокруг каждого местоположения на карте признаков, которые дают выходов для карты признаков размера .

Ключевое отличие обучения SSD от обучения типичного детектора, который использует предложения регионов (то есть предлагается какой-либо регион изображения и классифицируется) заключается в эталонной информации, которая должна быть назначена конкретным выходам в фиксированном наборе выходов детектора. Некоторая вариация этого также требуется для обучения YOLO [48] и для этапа предложения региона Faster R-CNN [**Ошибка! Источник ссылки не найден.**] и Multibox [18]. Как только это назначение определено, функция потерь и обратное распространение применяются сквозным образом. Обучение также включает в себя выбор набора анкеров и масштабов по умолчанию для обнаружения, а также стратегии аугментации и получения сложных «неверных» тренировочных данных (hard negative mining).

*Стратегия сопоставления*. Во время обучения необходимо определять, какие анкеры соответствуют верному обнаружению и тренировать сеть соответствующим образом. Каждый верный прямоугольник, сопоставляются с каждым анкером, различающимся по местоположению, соотношению сторон и масштабу и с наилучшим коэффициентом Жаккара (как в MultiBox [18]). Затем, в отличие от MultiBox, в SSD сопоставляются анкеры c истинными прямоугольниками с коэффициентом Жаккара выше порогового значения (. Это упрощает задачу обучения, позволяя сети предсказывать высокие значения для нескольких перекрывающихся прямоугольников, а не требовать от неё выбора только одного с максимальным перекрытием.

|  |  |  |
| --- | --- | --- |
|  | | |
| где: | | |
|  |  | число совпавших анкеров. Если , то потери |
|  |  | весовой коэффициент, устанавливающийся равным с помощью перекрёстной проверки (кросс-валидации или cross-validation). |
|  |  |  |

*Цель обучения*. Цель обучения SSD получена из цели Multibox [18], но расширена до обработки нескольких категорий объектов. Пусть индикатор для сопоставления -того анкера с -м истинным прямоугольником категории . Из приведённой выше стратегии сопоставления следует, что можно иметь . Общая целевая функция потерь представляет собой взвешенную сумму потерь локализации (loc) и потерь достоверности (conf):

* Потери локализации – это потери Smooth L1 [19] между предсказанным прямоугольником () и истинным прямоугольником (). Подобно Faster R-CNN [**Ошибка! Источник ссылки не найден.**] происходит регрессия к смещениям для центра ограничивающего блока по умолчанию (), а также для его ширины () и высоты ():
* Потери достоверности – это потери Softmax по достоверности нескольких классов ():

|  |  |  |
| --- | --- | --- |
|  | | |
| где: | | |
|  |  | . |

*Выбор масштабов и соотношений сторон для анкеров*. Для обработки различных масштабов объектов некоторые методы [54, 26] предполагают обработку изображения разных размеров и последующее объединение результатов. Однако, используя карты признаков из нескольких разных слоёв в одной сети для прогнозирования, можно имитировать тот же эффект, а также использовать общие параметры для всех масштабов объектов. Предыдущие работы [24, 20] показали, что использование карт признаков из нижних слоёв может улучшить качество семантической сегментации, поскольку нижние слои захватывают более мелкие детали входных объектов. Точно так же в [32] показано, что добавление главного контекста, объединённого из карты признаков, может помочь сгладить результаты сегментации. Руководствуясь этими методами, в SSD для обнаружения используются как нижняя, так и верхняя карта признаков.

Известно, что карты признаков с разных уровней внутри сети имеют разные (эмпирические) размеры рецептивного поля [68]. К счастью, в структуре SSD анкеры не обязательно должны соответствовать фактическим рецептивным полям каждого слоя. Мозаика анкеров разрабатывается таким образом, чтобы определённые карты признаков научились реагировать на определённые масштабы объектов. Предположим, нужно использовать карт признаков для предсказания. Масштаб блоков по умолчанию для каждой карты признаков рассчитывается как:

где , означающие, что самый нижний слой имеет масштаб , а самый верхний слой и все слои меду ними расположены через равные промежутки. Для анкеров устанавливаются различные пропорции и обозначаются как . Можно вычислить ширину () и высоту для каждого анкера. Для соотношения сторон также добавляется анкер, масштаб которого и в результате получается анкеров для каждого местоположения карты признаков. Центр каждого анкера устанавливается в , где это размер -ого квадрата карты признаков, .

На практике можно также спроектировать распределение анкеров, чтобы наилучшим образом соответствовать конкретному набору данных. Вопрос о том, как спроектировать оптимальную мозаику, также остаётся открытым.

Комбинируя прогнозы для всех анкеров с разными масштабами и соотношениями сторон из всех местоположений на многих картах признаков, получается разнообразный набор прогнозов, охватывающий различные размеры и формы входных объектов. Например, на рис. 15 собака сопоставляется с блоком по умолчанию на карте признаков , но не с каким-либо блоком по умолчанию на карте признаков . Это связано с тем, что эти блоки имеют разные масштабы и не соответствуют блоку собаки, и поэтому во время тренировки считаются негативами.

*Получение сложных «неверных» тренировочных данных (hard negative mining).* После этапа сопоставления большинство анкеров являются неверными, особенно когда количество возможных анкеров велико. Это вносит значительный дисбаланс между положительными и отрицательными обучающими примерами. Вместо того, чтобы использовать все отрицательные примеры, они сортируются по их наибольшей потере достоверности для каждого анкера и выбираются так, чтобы соотношение между отрицательными и положительными примерами было не более чем . Было обнаружено [37], что это приводит к более быстрой оптимизации и более стабильному обучению.

*Аугментация данных.* Чтобы сделать модель более устойчивой к различным размерам и формам входных объектов, каждое обучающее изображение случайным образом изменяется по одному из следующих правил:

* + Используется исходное изображение без изменений;
  + Из исходного изображения вырезается фрагмент таким образом, чтобы минимальный коэффициент Жаккара с объектами составлял или ;
  + Из исходного изображения вырезается случайный фрагмент.

Размер каждого фрагмента выборки лежит в отрезке от исходного размера изображения, а соотношение сторон находится в диапазоне от до . Перекрывающаяся часть истинного прямоугольника сохраняется, если его центр находится в фрагменте выборки. После вышеупомянутого шага размер каждого фрагмента изменяется до фиксированного и фрагмент горизонтально отражается с вероятностью , в дополнение к применению некоторых фотометрических искажений, подобных описанным в [26].

Как уже было отмечено, существует два класса методов детектирования объектов на изображениях: один основан на скользящих окнах, а другой – на классификации предложений регионов. До появления свёрточных нейронных сетей уровень развития этих двух подходов – модели DPM (Deformable Part Model [19]) и выборочного поиска [62] – имели сопоставимую производительность. Однако после резкого улучшения, внесённого R-CNN (Region-Based Convolutional Neural Network) [21], который сочетает в себе выборочный поиск предложений регионов и пост-классификацию на основе свёрточной сети, методы обнаружения объектов на основе предложений регионов стали преобладающими. Сравнение подходов представлено в таблице 5.

|  |  |  |
| --- | --- | --- |
| Таблица 5. | | |
| Сравнение подходов на тесте Pascal VOC2007. | | |
| Метод | mAP, % | FPS |
| Faster R-CNN (VGG16) |  |  |
| Fast YOLO |  |  |
| YOLO (VGG16) |  |  |
| SSD300 |  |  |
| SSD512 |  |  |

SSD300 – единственный метод, способный работать в реальном времени и набравший более 70% mAP (mean Average Precision – средняя точность предсказания), являющийся компромиссом между скоростью и качеством работы.

Первоначальный подход R-CNN был улучшен различными способами. Первый набор подходов улучшает качество и скорость пост-классификации, поскольку требует классификации тысяч кадров изображений, что дорого и затратно по времени. SPPnet [26] значительно ускоряет исходный подход R-CNN за счёт ввода слоя объединения слоя объединения пространственных пирамид, который более устойчив к размеры и масштабу региона и позволяет слоям классификации переиспользовать признаки, вычисленные по картам признаков, сгенерированным для нескольких разрешений изображения. Fast R-CNN [21] расширяет возможности SPPnet, так что он может точно настраивать все уровни от начала до конца, минимизируя потери как для достоверности, так и для регрессии ограничивающей рамки, которая впервые была представлена в MultiBox [18] для изучения объектности.

Второй набор подходов улучшает качество генерации предложений, используя глубокие нейронные сети. В самых последних работах, таких как MultiBox [18], предложения региона выборочного поиска, которые основаны на низкоуровневых признаках изображения, заменены предложениями, сгенерированными непосредственно из отдельной глубокой нейронной сети. Это дополнительно повышает точность обнаружения, но приводит к несколько сложной настройке, требующей обучения двух нейронных сетей с зависимостью между ними. Faster R-CNN [**Ошибка! Источник ссылки не найден.**] заменяет предложения выборочного поиска предложениями, полученными из сети региональных предложений (RPN), и вводит метод интеграции RPN с Fast R-CNN путем чередования точной настройки общих свёрточных слоёв и слоёв прогнозирования для этих двух сетей. Таким образом, предложения регионов используются для объединения признаков среднего уровня, а последний этап классификации обходится дешевле. SSD очень похож на сеть региональных предложений (RPN) в Faster R-CNN, поскольку также используется фиксированный набор блоков (по умолчанию) для прогнозирования, аналогично полям привязки в RPN. Но вместо того, чтобы использовать их для объединения признаков и оценки другого классификатора, оценка для каждой категории объектов получается одновременно в каждом блоке. Таким образом, этот подход позволяет избежать осложнений, связанных с объединением RPN с Fast R-CNN, и его легче обучать, быстрее и проще интегрировать в другие задачи.

Другой набор методов, непосредственно связанных с подходом SSD, полностью пропускает шаг предложения и напрямую прогнозирует ограничивающие рамки и достоверность для нескольких категорий. OverFeat [54], глубокая версия метода скользящего окна, предсказывает ограничивающую рамку непосредственно из каждого местоположения самой верхней карты признаков после того, как известны достоверности базовых категорий объектов. YOLO [39] использует всю самую верхнюю карту признаков для прогнозирования как достоверности для нескольких категорий, так и ограничивающих рамок (которые являются общими для этих категорий). Метод SSD попадает в эту категорию, поскольку в нём нет шага предложения, но используются блоки по умолчанию. Однако этот подход является более гибким, чем существующие методы, поскольку существует возможность использовать блоки по умолчанию с разными соотношениями сторон для каждого местоположения признака из нескольких карт признаков в разных масштабах. Если будет использоваться только один блок по умолчанию для каждого местоположения из самой верхней карты признаков, то SSD будет иметь архитектуру, аналогичную OverFeat [54]; если используется вся самая верхняя карта признаков и добавляется полносвязный слой для прогнозов вместо свёрточных предсказателей и не будет явно учитываться несколько соотношений сторон, то приблизительно воспроизведётся архитектура YOLO [39, 1].

Таким образом, резюмируя всё выше сказанное, можно отметить следующие особенности Single Shot детектора:

* Архитектура SSD позволяет детектировать объекты в реальном времени;
* Качество работы близко к Faster R-CNN;
* Детектирование происходит на разных масштабах, что позволяет локализовывать объекты разных размеров;
* Используется большое количество рамок по умолчанию, покрывающих входное изображение на разных масштабах;
* На этапе работы архитектура SSD300 осуществляет детектирование 7308 объектов, большая часть из которых впоследствии фильтруется;

Ключевой же особенностью является использование многомасштабных ограничительных рамок на выходах свёрточных слоёв, прикреплённых к нескольким картам признаков в верхней части сети. Экспериментально подтверждено в [37], что при соответствующих стратегиях обучения, большее количество тщательно выбранных анкеров приводит к повышению производительности.

Для решения задачи детектирования кисти на изображении Single Shot детектор подходит наилучшим образом, поскольку он имеет хорошее соотношение между скоростью работы и точностью. Этот же детектор также используется в библиотеке MediaPipe Hands, использующейся для поиска ключевых точек на руках. Рассмотрим её поподробнее.

2.3. Применение SSD в задаче детектирования кисти

Одним из применений Single Shot детектора в задаче поиске кисти на изображении является подход, реализованный в фреймворке MediaPipe от Google – MediaPipe Hands [**Ошибка! Источник ссылки не найден.**]. Для определения начального положения рук была разработана модель SSD, оптимизированная для использования в реальном времени. Поскольку обнаружение рук чрезвычайно сложная задача из-за огромного количества их конфигураций, то вместо детектора рук обучался детектор ладоней, так как получать прямоугольники негнущихся объектов, таких как ладони и кулаки, значительно проще, нежели обнаружение рук вместе с пальцами. Кроме того, поскольку ладони являются более мелкими объектами, алгоритм подавления немаксимумов хорошо работает даже для случаев пересечения двух рук, таких как рукопожатия. Более того, ладони могут быть смоделированы с помощью квадратных анкеров, игнорируя другие соотношения сторон, уменьшая количество анкеров в 3-5 раз. Во-вторых, кодирующий-декодирующий экстрактор признаков используется для большей осведомленности о контексте сцены даже для небольших объектов. Наконец, во время обучения минимизируется функция кросс-энтропийных потерь с динамическим масштабированием (focal loss), чтобы поддерживать большое количество анкеров, возникающих из-за большой дисперсии масштаба. С помощью описанных выше методов достигается средняя точность обнаружения ладони . Использование обычной функции кросс-энтропийных потерь и без декодера даёт точность всего .

После обнаружения ладони на всём изображении последующая модель выполняет точную локализацию 21 ключевой точки (рис. 17а) внутри обнаруженных областей руки посредством регрессии (как для определения цифры, написанной рукой), то есть прямого прогнозирования координат. Модель обучается последовательному внутреннему представлению позиции рук и устойчива даже к частично видимым рукам и самопересечениям. Так же стоит отметить, что кроме 21 ключевой точки модель так же определяет вероятность присутствия кисти на кадре и классификацию типа кисти – левой или правой.

|  |  |
| --- | --- |
|  |  |
| Рис. 17. Ключевые точки на руке, определяемые с помощью MediaPipe Hands (а). Данные для обучения (б). Сверху: реальные изображения рук, аннотированные вручную. Снизу: сгенерированные синтетические изображения рук с аннотациями. | |

Чтобы получить данные для обучения, вручную было промаркировано около изображений реального мира с координатой, как показано на рис. 17б. А также, чтобы лучше охватить возможные позиции рук и обеспечить дополнительное наблюдение за характером геометрии рук, была визуализирована высококачественная синтетическая модель руки на различных фонах, таким образом синтезируя набор из изображений. Для детектора ладони использовались лишь  000 изображений данного набора, поскольку этого достаточно для его обучения локализации кисти.

Таким образом, библиотека состоит из детектора ладоней, определяющего область изображения, на которой изображена рука и модели трекинга руки, предсказывающая ключевые точки на области, полученной детектором.

Основываясь на данной технологии, полученные ключевые точки можно использовать в дальнейших приложениях, например, в распознавании жестов.

3. РАЗРАБОТКА ИНТЕРФЕЙСА ДЛЯ УПРАВЛЕНИЯ ЖЕСТАМИ

Прежде чем начать разрабатывать эргономичный интерфейс для управления путём распознавания, необходимо обобщить все ранее отмеченные выкладки и сформировать окончательный алгоритм работы человеко-машинного взаимодействия с помощью распознавания жестов.

3.1. Методика управления жестами

Общая схема алгоритма работы интерфейса изображена на рис. 18.

|  |
| --- |
|  |
| Рис. 18. Общая схема алгоритма работы интерфейса. |

Изображение с кистью, считанное одиночной камерой, поступает на вход Single Shot детектора ладоней. Затем отделённое от всего остального кадра изображение кисти поступает на вход искусственной нейронной сети MediaPipe Hands, которая определяет положение ключевых точек с помощью регрессии. Массив ключевых точек сохраняется в буфер, первое время лишь наполняя его, и на вход искусственной нейронной сети классификации статических жестов, которая сопоставляет массив точек с жестом и возвращает результат в виде номера жеста. Если буфер точек, работая по принципу очереди, заполнился то, все данные в нём поступают на вход искусственной нейронной сети классификации динамических жестов, и если эта сеть определила, что массив сопоставляется с каким-либо динамическим жестом, то возвращает номер динамического жеста. Приоритет у динамического жеста выше, чем у статического. Рассмотрим подробнее реализацию методики.

3.2. Требования к реализации

Прежде чем выдвинуть требования к реализации интерфейса человеко-машинного взаимодействия необходимо вывести основные требования для эффективной работы систем «человек — техника», следующих из *инженерной психологии*:

1. Наличие высокопроизводительной и надёжной техники;
2. Конструкция техники и организация производственного процесса должны позволять человеку реализовывать все технические возможности;
3. Человек должен быть способным по своим качествам реализовывать все эти возможности, добиваться высокой производительности труда и обеспечивать выполнение производственных операций.

Требования к разрабатываемому человеко-машинному интерфейсу относятся ко второму пункту, то есть должно быть достигнуто согласование человека и техники как элементов единой системы. Под согласованием понимается максимальное приспособление техники к человеку (по параметрам конструкции и технологического процесса), максимальное приспособление человека к технике (по параметрам профессиональной пригодности и профессиональной подготовленности) и рациональное распределение функций между человеком и автоматическими устройствами в системах «человек – техника».

Отсюда следует, что одним из требований к реализации интерфейса человеко-машинного взаимодействия на основе распознавания жестов является *эргономичность*, включающая в себя *гибкость* использования. Здесь под гибкостью понимается лёгкость и быстрота добавления новых жестов обычным пользователем, поскольку в зависимости от машины, сферы применения и количества действий количество жестов может варьироваться, а также изменяться их вид – могут добавляться новые или удаляться/переназначаться старые. Гибкость использования реализуется посредством реализации двух режимов работы: «*обучение*», включающее в себя добавление новых жестов, и «*применение*», идентифицирующее жест и посылающее связанный с этим жестов сигнал.

Таким образом в режиме «обучение» нужно учесть следующее:

* Выбор режима работы перед запуском программы;
* Переключение между записью статического и динамического жеста;
* Непрерывный мониторинг уже записанных жестов и количества записей по ним;
* Вывод текущего записанного жеста;

Для реализации алгоритма будет использоваться язык программирования Python [47]. Выбор обусловлен тем, что Python имеет большое количество развивающихся библиотек для работы с нейронными сетями и простой синтаксис, позволяющий быстро и качественно реализовывать идеи, в том числе с использованием Jupyter Notebooks [31].

Спроектированный вид интерфейса в режиме «обучение» изображён на рис. 19. В верхней части выводится количество записей по каждому из жестов, предоставляя возможность выравнивать количество данных по ним. При нажатии на клавишу

* ‘q’ или Escape – программа завершает свою работу;
* цифру, то есть ‘0’-‘9’ – начинается запись или записывается жест по данной цифре;
* ‘c’ – переключаются режимы одиночной и непрерывной записи статического жеста. В первом режиме координаты ключевых точек записываются только при нажатии на соответствующую цифру. Последний же позволяет сохранять координаты ключевых непрерывно, а не только при нажатии на цифру. Данный подход позволяет намного быстрее собирать большое количество данных, непрерывно модифицируя и перемещая руку с жестом;
* ‘s’ – включается/выключается запись динамического жеста, которая продолжается до нажатия какой-либо другой клавиши. При последовательной записи динамических жестов удобно пользоваться только цифровыми клавишами, поскольку они одновременно останавливают предыдущую запись и начинают новую.

|  |
| --- |
|  |
| Рис. 19. Вид программы в режиме "обучение. |

Таким образом, собрав необходимое количество данных, необходимо на их основе обучить модель нейронной сети распознавать конкретных жест исходя из ключевых точек.

3.3. Подготовка и обучение модели распознавания статических жестов

Подготовка данных для обучения модели заключается в преобразовании координат к необходимому нормализованному виду, как, например, на рис. 20. Исходные значения координат поступают с привязкой к расположению ключевой точки на изображении, то есть в пикселях. Поскольку жест не всегда может находиться в конкретном данном расположении, то необходимо отвязать их от изображения. Для этого все координаты пересчитываются относительно ключевой точки, находящейся на запястье, и нормируются.

|  |
| --- |
|  |
| Рис. 20. Преобразование исходных координат. |

Затем, набор массивов координат, привязанных к конкретным жестам, подаётся на вход искусственной нейронной сети с предварительно отобранной структурой, показанной на рис. 21.

|  |
| --- |
|  |
| Рис. 21. Структура нейронной сети для распознавания статических жестов. |

Более подробно структуру можно записать следующим образом:

* Входной слой ()
* Дропаут
* Скрытый слой () с функцией активации ReLU:
* Дропаут
* Скрытый слой () с функцией активации ReLU
* Выходной слой () с функцией активации Softmax, где количество жестов.

*Исключение* или *дропаут* – метод регуляризации искусственных нейронных сетей, предназначен для уменьшения переобучения сети за счёт предотвращения сложных коадаптаций отдельных нейронов на тренировочных данных во время обучения, характеризует исключение определённого процента случайных нейронов на разных итерациях во время обучения нейронной сети. Такой приём значительно увеличивает скорость обучения, качество обучения на тренировочных данных, а также повышает качество предсказаний модели на новых тестовых данных.

В качестве функции оптимизации была выбрана функция Adam. поскольку это один из самых эффективных алгоритмов оптимизации в обучении нейронных сетей. Он сочетает в себе идеи среднеквадратичного распространения корня (RMSProp) и оптимизатора импульса. Вместо того чтобы адаптировать скорость обучения параметров на основе среднего первого момента (среднего значения), как в RMSProp, Adam также использует среднее значение вторых моментов градиентов. В частности, алгоритм вычисляет экспоненциальное скользящее среднее градиента и квадратичный градиент.

В качестве примера были выбраны 3 жеста с вытянутым одним, двумя и тремя пальцами соответственно и заранее сгенерированный набор из 3138 точек для этих жестов.

Для разработки и обучения искусственной нейронной сети была выбрана библиотека TensorFlow v2 [60]. По сравнению с похожей библиотекой PyTorch, TensorFlow показывает себя намного производительней, а значит тренировка модели займёт меньшее время.

Обучение для него проходило на 97 эпохах (рис. 22). Доля правильных ответов (accuracy) , потери (loss) . Основные метрики по каждому классу представлены в таблице 6.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис. 22. Изменение метрики accuracy (а) и loss (б) на тренировочных и валидирующих данных во время обучения. | |

|  |  |  |  |
| --- | --- | --- | --- |
| Таблица 6. | | | |
| Показатели метрик классификации статических жестов. | | | |
| Класс | Точность (precision) | Полнота (recall) | F-мера (f1-score) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

На матрице ошибок для задачи распознавания статических жестов (рис. 23) показано соотношение истинных значений (снизу) с ответами нейросети (справа).

|  |
| --- |
|  |
| Рис. 23. Матрица ошибок для задачи распознавания статических жестов. |

Интуитивно понятной, очевидной и почти неиспользуемой метрикой является accuracy – доля правильных ответов алгоритма:

Для оценки качества работы алгоритма на каждом из классов по отдельности вводятся метрики precision (точность) и recall (полнота):

Precision можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а recall показывает, какую долю объектов положительного класса из всех объектов положительного класса нашёл алгоритм. Существует несколько различных способов объединить precision и recall в агрегированный критерий качества. -мера (в общем случае ) – среднее гармоническое precision и recall:

Переменная в данном случае определяет вес точности в метрике, и при это среднее гармоническое. -мера достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю.

Проанализировав результаты, можно сделать вывод, что при тестировании нейросеть показала неплохие результаты, а значит её можно применять в задаче классификации статических жестов по ключевым точкам.

3.4. Расширение модели распознавания статических жестов

Расширение модели распознавания статических жестов заключается в обучении модели на большом наборе данных. Примером такого набора может быть набор HANDS статических жестов для взаимодействия человека и робота [43]. Этот набор был создан для исследования взаимодействия между человеком и роботом и состоит из пространственно и временно выровненных кадров RGB и глубины. Он содержит 12 видов статических жестов, выполненных как правой, так и левой рукой, и 3 вида статических жестов с двумя руками, таким образом, всего 26 уникальных классов.

Пять испытуемых (2 женщины и 3 мужчины) выполняли жесты, каждый из которых был на разном фоне и в разных условиях освещения. Для каждого жеста было собрано 150 кадров RGB и соответствующие им 150 кадров глубины, всего 2400 кадров RGB и 2400 кадров глубины на человека.

Данные были собраны с помощью камеры Kinect V2, откалиброванной для пространственного выравнивания данных RGB с данными о глубине. Временное выравнивание было выполнено в автономном режиме с использованием MATLAB, выравнивая кадры с максимальным временным расстоянием 66 мс. Примеры жестов изображены на рис. 24.

|  |
| --- |
|  |
| Рис. 24. Примеры фотографий каждого класса из набора данных HANDS. |

На каждом изображении были найдены ключевые точки и сопоставлены с классом жеста. Таким образом, общий набор составил записей по всем классам.

Тренировка нейросети (рис. 27) на эпохах на этом наборе жестов c оптимизатором Adam, а затем валидация на тестовой выборке из него показала точность и потери , что недостаточно для корректной классификации жестов, поэтому было решено её модифицировать.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис. 25. Изменение метрики accuracy (а) и loss (б) на тренировочных и валидирующих данных во время обучения нейросети на большой выборке. | |

Улучшенная структура искусственной нейронной сети для классификации статических жестов содержит большее количество скрытых слоёв и нейронов в них и изображена на рис. 26. Все слои, кроме последнего, как и прежде имеют функцию активации ReLU.

|  |
| --- |
|  |
| Рис. 26. Улучшенная архитектура нейросети. |

Обучение улучшенной нейросети проходило на 156 эпохах (рис. 27). Проверив работу нейросети на тестовых данных, доля правильных ответов (accuracy) составила , а потери (loss) – . Основные метрики по каждому классу представлены в таблице 7.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис. 27. Изменение метрики accuracy (а) и loss (б) на тренировочных и валидирующих данных во время обучения улучшенной нейросети. | |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Таблица 7. | | | | | | | | |
|  | | Показатели метрик расширенной модели классификации. | | | | | | | | |
| Класс | Точность (precision) | | Полнота (recall) | F-мера (f1-score) |  | Класс | Точность (precision) | Полнота (recall) | F-мера (f1-score) |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |

Матрица ошибок для задачи распознавания жестов с помощью улучшенной нейросети представлена на рис. 28.

|  |
| --- |
|  |
| Рис. 28. Матрица ошибок для задачи распознавания жестов с помощью улучшенной нейросети. |

Проанализировав результаты, можно сделать вывод, что улучшенная нейросеть успешно обучена и её можно применять в задаче классификации большего количества жестов.

3.5. Подготовка и обучение модели распознавания динамических жестов

По аналогии с подготовкой данных для распознавания статических жестов для обучения модели распознавания динамических жестов с каждым массивом координат точек необходимо проделывать все те же действия, описанные ранее (рис. 29).

|  |
| --- |
|  |
| Рис. 29. Подготовка данных для обучения нейросети распознавания динамических жестов. |

Поскольку динамический жест является последовательностью координат различной длины, то каждую запись жеста необходимо привести к одному общему размеру. В данном случае можно дополнить каждую из записей незначащими нулями справа до максимального размера среди всех записей.

Собранный набор данных для обучения включает в себя 3 класса динамических жестов: «вращение указательным пальцем по часовой стрелке», «вращение указательным пальцем против часовой стрелки» и «какой-то другой жест».

Необходимость последнего обусловлена тем, что при работе система непрерывно собирает распознанные координаты в буфер и так же непрерывно распознаёт динамический жест. Поскольку в текущий момент времени динамического жеста может не быть, но при этом может быть статический, то необходимо отделять «полезные» динамические жесты от побочных. Всего было собрано записей, распределение между классами соответственно и записей. Графическое изображение последовательности координат указательного пальца (точка на рис. 17а) в записях изображено на рисунке 30а, а на рисунке 30б изображены те же координаты, приведённые к относительным запястью (точке ). После сбора и нормализации данных происходит фильтрация сильно длинных последовательностей, чтобы остальные наборы координат дополнялись нулями несущественно.

|  |  |
| --- | --- |
|  |  |
| *а)* | *б)* |
| Рис. 30. Графическое изображение примеров последовательности координат указательного пальца (а) относительно запястья (б) ( «вращение по часовой стрелке», «против»).  Формат легенды: <номер записи>\_<класс>. | |

Затем, все эти наборы координат подаются на вход искусственной нейронной сети, проиллюстрированной на рис. 31. В структуре нейросети для классификации динамических жестов присутствует LSTM-слой с 50 выходами, слой дропаута 20%, скрытый слой с 100 нейронами и функцией активации ReLU и выходной слой с количеством нейронов, равным количеству динамических жестов (в текущем случае – 3), и функцией активации Softmax. При обучении будет так же использоваться функция оптимизации Adam.

|  |
| --- |
|  |
| Рис. 31. Структура нейронной сети для распознавания динамических жестов. |

Обучение сети проходило на эпохах (рис. 32). Доля правильных ответов (accuracy) составила , а потери (loss) – . Основные метрики по каждому классу представлены в таблице 8.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис. 32. Инициализированный фон (а) и изображение кисти на нём (б). | |

|  |  |  |  |
| --- | --- | --- | --- |
| Таблица 8. | | | |
| Показатели метрик классификации динамических жестов. | | | |
| Класс | Точность (precision) | Полнота (recall) | F-мера (f1-score) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Матрица ошибок для задачи распознавания динамических жестов изображена на рисунке 23.

|  |
| --- |
|  |
| Рис. 33. Матрица ошибок для задачи распознавания динамических жестов. |

Таким образом, можно сделать вывод, что нейросеть классификации динамических успешно обучена и её можно применять в задаче классификации динамических жестов.

3.6. Оценка работы интерфейса

В качестве практического применения интерфейса и тестирования распознавания жестов было выбрано управление курсором мыши, соотношения между жестами и действием которого распределено следующим образом:

* Жест 1 (показывается только указательный палец, остальные собраны в кулак) – перемещение курсора, где верхняя точка указательного пальца (точка на рис. 17а) отвечает за его координаты на экране;
* Жест 2 (по аналогии с 1 показывается только указательный и средний палец) – нажатие левой кнопки мыши;
* Жест 3 (по аналогии с 1 и 2 показывается указательный, средний и безымянный пальцы) – нажатие правой кнопки мыши;

Интерфейс программы изображён на рис. 34.

|  |
| --- |
|  |
| Рис. 34. Интерфейс программы при управлении курсором мыши. |

В верхней части отображается количество кадров в секунду (FPS, Frames Per Second) и номер распознанного жеста. В центральной части отображается область проекции экрана, то есть левый верхний угол данной области соответствует левому верхнему углу экрана и так далее. Перемещая указательный палец в данной области, перемещается курсор по экрану. Причина, по которой была выбрана именно область, а не всё изображение, состоит в том, что в крайних его частях плохо или вовсе не распознаётся жест, поскольку рука не полностью помещается в кадр.

При первом тестировании такой системы был найден недостаток такого подхода – из-за небольшой тряски руки курсор постоянно находится в движении и трудно корректно его позиционировать. В качестве решения был предложен метод *фильтра перемещения в малом*, заключающийся в игнорировании малых перемещений точки в пределах пикселей, решивший данную проблему.

При последующих тестированиях алгоритм работал достаточно точно и быстро, в большинстве случаев верно определяя ключевые точки и классифицируя их, выдавая заданный жест. Неточности возникали при больших наклонах руки – это связано прежде всего с неудачными данными для обучения. При увеличении количества различных ситуаций и количества данных алгоритм работал еще точнее.

В качестве основного недостатка данной системы можно выделить неустойчивость при резких передвижениях руки в кадре. Это вызвано тем, что алгоритм предсказывает следующее положение руки в малой окрестности, а при резком перемещении срабатывает детектор, которому необходимо немного больше времени для работы распознавания.

ЗАКЛЮЧЕНИЕ

В результате данной работы были рассмотрены основные теоретические выкладки по решению задачи распознавания жестов, подробно описаны существующие подходы и проведено их сравнение. В качестве датчика была выбрана одиночная камера, считывающая изображение жеста.

Также в работе были подробно изучены теоретические аспекты актуального способа детектирования объектов на основе искусственных нейронных сетей – Single Shot детектора. В результате сравнения с аналогами был сделан вывод, что на текущий момент этот детектор наилучшим образом подходит для решения требуемой задачи с использованием средств глубокого обучения.

После этого были описаны особенности работы современной библиотеки MediaPipe Hands, использующей Single Shot детектор и регрессию для определения 21 ключевой точки на руке. На основе хороших результатов работы этой библиотеки, полученных в результате тестирования, ей были делегированы задачи детектирования и извлечения конфигурации руки.

Далее была разработана методика обучения и распознавания статических и динамических жестов для управления курсором мыши. Для этого:

* была разработана программа с графическим интерфейсом для быстрого сбора данных;
* были собраны два набора данных: первый – состоящий из 3000 записей ключевых точек для трёх динамических статических жестов, и второй – 549 записей для трёх классов динамических жестов;
* была спроектирована и обучена искусственная нейронная сеть, классифицирующая *статические* жесты со средней точностью по классам равной 96%;
* была спроектирована и обучена искусственная нейронная сеть с рекуррентным слоем LSTM, классифицирующая *динамические* жесты со средней точностью по классам равной 99%;
* была разработана программа управления курсором мыши для тестирования методики, в результате использования которой исправлены незначительные недостатки методики и выявлены особо крупные.

По итогу работы были выявлены основные недостатки разработанной методики, заключающиеся в необходимости изменения архитектуры нейронной сети при увеличении классов жестов и в нестабильности работы из-за использования одиночной камеры.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Нагапетян В.Э., Хачумов В.М.. Автоматическое преобразование жестов русской ручной азбуки в текстовый вид // Искусственный интеллект и принятие решений. 2013. с. 59-66.
2. Adib F. et al. 3D tracking via body radio reflections // 11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14). – 2014. – С. 317-329.
3. Adib F. et al. Capturing the human figure through a wall // ACM Transactions on Graphics (TOG). – 2015. – Т. 34. – №. 6. – С. 1-13.
4. Adib F., Katabi D. See through walls with WiFi! // Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM. – 2013. – С. 75-86.
5. Allen B., Curless B., Popović Z. Articulated body deformation from range scan data // ACM Transactions on Graphics (TOG). – 2002. – Т. 21. – №. 3. – С. 612-619.
6. Arici T. et al. Robust gesture recognition using feature pre-processing and weighted dynamic time warping // Multimedia Tools and Applications. – 2014. – Т. 72. – №. 3. – С. 3045-3062.
7. Babenko B., Yang M. H., Belongie S. Visual tracking with online multiple instance learning // 2009 IEEE Conference on computer vision and Pattern Recognition. – IEEE, 2009. – С. 983-990.
8. Bauer A., Wollherr D., Buss M. Human–robot collaboration: a survey // International Journal of Humanoid Robotics. – 2008. – Т. 5. – №. 01. – С. 47-66.
9. Bay H., Tuytelaars T., Gool L. V. Surf: Speeded up robust features // European conference on computer vision. – Springer, Berlin, Heidelberg, 2006. – С. 404-417.
10. Belongie S., Malik J., Puzicha J. Shape matching and object recognition using shape contexts // IEEE transactions on pattern analysis and machine intelligence. – 2002. – Т. 24. – №. 4. – С. 509-522.
11. Breazeal C. et al. Humanoid robots as cooperative partners for people // Int. Journal of Humanoid Robots. – 2004. – Т. 1. – №. 2. – С. 1-34.
12. Celebi S. et al. Gesture recognition using skeleton data with weighted dynamic time warping // VISAPP (1). – 2013. – С. 620-625.
13. Cohen P. R., Levesque H. J. Persistence, intention, and commitment // Reasoning about actions and plans. – 1990. – С. 297-340.
14. Cohen P. R., Levesque H. J. Teamwork // Nous. – 1991. – Т. 25. – №. 4. – С. 487-512.
15. Comaniciu D., Ramesh V., Meer P. Real-time tracking of non-rigid objects using mean shift, in: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, IEEE, 2000, С. 142–149.
16. Dix A., Finlay J., Abowd G.D., Beale R. Human-Computer Interaction. - Third Edition, Pearson Education Limited: 2004. – С. 857
17. Elmezain M. et al. A hidden markov model-based continuous gesture recognition system for hand motion trajectory // 2008 19th international conference on pattern recognition. – IEEE, 2008. – С. 1-4.
18. Erhan D. et al. Scalable object detection using deep neural networks // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2014. – С. 2147-2154.
19. Felzenszwalb P., McAllester D., Ramanan D. A discriminatively trained, multiscale, deformable part model // 2008 IEEE conference on computer vision and pattern recognition. – IEEE, 2008. – С. 1-8.
20. Girshick R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2014. – С. 580-587.
21. Girshick R. Fast R-CNN. 2015. ICCV.
22. Google, Project Soli, in, Google, 2015. URL: <https://www.google.com/atap/project-soli/> (дата обращения: 23.04.2022)
23. Green S. A. et al. Human-robot collaboration: A literature review and augmented reality approach in design // International journal of advanced robotic systems. – 2008. – Т. 5. – №. 1. – С. 1.
24. Hariharan B. et al. Hypercolumns for object segmentation and fine-grained localization // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2015. – С. 447-456.
25. Haykin S. Kalman filtering and neural networks. – John Wiley & Sons, 2004. – Т. 47.
26. He K. et al. Spatial pyramid pooling in deep convolutional networks for visual recognition // IEEE transactions on pattern analysis and machine intelligence. – 2015. – Т. 37. – №. 9. – С. 1904-1916.
27. Hearst M. A. et al. Support vector machines // IEEE Intelligent Systems and their applications. – 1998. – Т. 13. – №. 4. – С. 18-28.
28. Hochreiter S., Schmidhuber J. Long short-term memory // Neural computation. – 1997. – Т. 9. – №. 8. – С. 1735-1780.
29. Howard A. G. Some improvements on deep convolutional neural network based image classification. 2013.
30. Howe N., Leventon M., Freeman W. Bayesian reconstruction of 3d human motion from single-camera video // Advances in neural information processing systems. – 1999. – Т. 12.
31. Jupyter. URL: <https://jupyter.org/> (дата обращения 10.05.2022)
32. Kadous M. W. Auslan sign recognition using computers and gloves // Deaf Studies Research Symposium. – 1998.
33. Kalman R. E. A new approach to linear filtering and prediction problems, Journal of Fluids Engineering, 82, 1960, С. 35–45.
34. Katsuki Y., Yamakawa Y., Ishikawa M. High-speed human/robot hand interaction system // Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts. – 2015. – С. 117-118.
35. Krüger J., Lien T. K., Verl A. Cooperation of human and machines in assembly lines // CIRP annals. – 2009. – Т. 58. – №. 2. – С. 628-646.
36. Kwon J., Lee K. M., Park F. C. Visual tracking via geometric particle filtering on the affine group with optimal importance functions, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, С. 991–998.
37. Liu W. et al. Ssd: Single shot multibox detector // European conference on computer vision. – Springer, Cham, 2016. – С. 21-37.
38. Liu W., Rabinovich A., Berg A.C. ParseNet: Looking wider to see better. 2016. ICLR.
39. Long J., Shelhamer E., Darrell T. Fully convolutional networks for semantic segmentation //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2015. – С. 3431-3440.
40. Lowe D. G. Object recognition from local scale-invariant features // Proceedings of the seventh IEEE international conference on computer vision. – IEEE, 1999. – Т. 2. – С. 1150-1157.
41. Matsumoto Y., Zelinsky A. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement // Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580). – IEEE, 2000. – С. 499-504.
42. Mitra S., Acharya T. Gesture recognition: A survey // IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). – 2007. – Т. 37. – №. 3. – С. 311-324.
43. Nuzzi C. et al. HANDS: an RGB-D dataset of static hand-gestures for human-robot interaction // Data in Brief. – 2021. – Т. 35. – С. 106791.
44. Oikonomidis I., Kyriazis N., Argyros A. A. Efficient model-based 3D tracking of hand articulations using Kinect // BmVC. – 2011. – Т. 1. – №. 2. – С. 3.
45. Parasuraman R., Sheridan T. B., Wickens C. D. A model for types and levels of human interaction with automation // IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans. – 2000. – Т. 30. – №. 3. – С. 286-297.
46. Peterson L. E. K-nearest neighbor // Scholarpedia. – 2009. – Т. 4. – №. 2. – С. 1883.
47. Python.org. URL: <https://www.python.org>/ (дата обращения 10.05.2022)
48. Redmon J. et al. You only look once: Unified, real-time object detection //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – С. 779-788.
49. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS. 2015.
50. Rublee E. et al. ORB: An efficient alternative to SIFT or SURF // 2011 International conference on computer vision. – IEEE, 2011. – С. 2564-2571.
51. Schapire R. E. The boosting approach to machine learning: An overview // Nonlinear estimation and classification. – 2003. – С. 149-171.
52. Schapire R. E., Freund Y. A decision-theoretic generalization of on-line learning and an application to boosting // Journal of computer and system sciences. – 1997. – Т. 55. – №. 1. – С. 119-139.
53. Schölkopf B., Smola A. Support Vector Machines, Encyclopedia of Biostatistics, 1998.
54. Sermanet P. et al. Overfeat: Integrated recognition, localization and detection using convolutional networks. 2014. ICLR.
55. Smeulders A. W. M. et al. Visual tracking: An experimental survey // IEEE transactions on pattern analysis and machine intelligence. – 2013. – Т. 36. – №. 7. – С. 1442-1468.
56. Smith J. et al. Electric field sensing for graphical interfaces // IEEE Computer Graphics and Applications. – 1998. – Т. 18. – №. 3. – С. 54-60.
57. Starner T. E. Visual Recognition of American Sign Language Using Hidden Markov Models. – Massachusetts Inst Of Tech Cambridge Dept Of Brain And Cognitive Sciences, 1995.
58. Starner T., Weaver J., Pentland A. Real-time american sign language recognition using desk and wearable computer based video // IEEE Transactions on pattern analysis and machine intelligence. – 1998. – Т. 20. – №. 12. – С. 1371-1375.
59. Suarez J., Murphy R. R. Hand gesture recognition with depth images: A review //2012 IEEE RO-MAN: the 21st IEEE international symposium on robot and human interactive communication. – IEEE, 2012. – С. 411-417.
60. TensorFlow. URL: <https://tensorflow.org/> (дата обращения 14.05.2022)
61. Thurau C., Hlavác V. Pose primitive based human action recognition in videos or still images // 2008 IEEE conference on computer vision and pattern recognition. – IEEE, 2008. – С. 1-8.
62. Uijlings J. R. R. et al. Selective search for object recognition // International journal of computer vision. – 2013. – Т. 104. – №. 2. – С. 154-171.
63. Vygotsky L. S., Cole M. Mind in society: Development of higher psychological processes. – Harvard university press, 1978.
64. Wachs J. P. et al. Vision-based hand-gesture applications // Communications of the ACM. – 2011. – Т. 54. – №. 2. – С. 60-71.
65. Wan E., Van Der Merwe R. The unscented Kalman filter for nonlinear estimation, in: Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000, IEEE, 2000, С. 153–158.
66. Wang R. Y., Popović J. Real-time hand-tracking with a color glove // ACM transactions on graphics (TOG). – 2009. – Т. 28. – №. 3. – С. 1-8.
67. Zhang F. et al. MediaPipe Hands: On-device Real-time Hand Tracking. 2020. CVPR.
68. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. Object detectors emerge in deep scene cnns. 2015. ICLR.