W0Министерство образования и науки Российской Федерации

Государственное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)»;

Центр «ПУСК» МФТИ

Программа «Науки о данных»

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

по направлению 01.04.02 «Прикладная математика и информатика»

Тема: «ML-моделирование перевода с русского жестового на русский звуковой»

(рабочее название)

Студент:

Кудрявцева П. Д.

Научный руководитель:

Каприелова М. С.

Москва 2024

**Содержание**

[Введение 3](#_Toc147702595)

[Глава 1. Обзор литературы и доступных данных по теме 8](#_Toc147702596)

[1.1 Особенности жестовых языков и РЖЯ в частности 8](#_Toc147702597)

[1.2 Наборы данных в области распознавания и перевода с жестовых языков 12](#_Toc147702598)

[1.3 Подходы к решению задачи в зарубежной и отечественной литературе 15](#_Toc147702599)

[Глава 2. Анализ данных 20](#_Toc147702600)

[2.1 Данные из корпуса РЖЯ 20](#_Toc147702601)

[2.2 Датасет Slovo 21](#_Toc147702602)

[Список литературы 22](#_Toc147702603)

Введение

Общение – важная составляющая в жизни любого человека. Все мы испытываем потребность делиться с другими своими мыслями, чувствами, точками зрения, хотим быть понятыми и услышанными. Также общение важно и для развития человека – мы многому учимся именно через общение с другими людьми: родными, близкими, учителями, друзьями, коллегами.

Глухие и слабослышащие люди зачастую сталкиваются с определёнными сложностями в общении с людьми слышащими. Основное средство коммуникации между глухими и слабослышащими – это жестовые языки. Далеко не каждый слышащий человек знаком с жестовым языком своих соотечественников. Для глухих же научиться понимать (посредством чтения по губам) и говорить (не слыша самих себя) – весьма сложная задача, которая требует кропотливой работы как с его стороны, так и со стороны преподавателя. Более того, в некоторых случаях (например, при общении с врачом в маске) человек оказывается лишённым даже возможности прочитать по губам.

Статистические данные свидетельствуют о постоянном росте числа новорожденных с нарушениями слуха. По данным Всемирной федерации глухих, более 300 жестовых языков используются более чем 700 миллионами людей по всему миру. По данным ВОЗ, количество глухих и слабослышащих людей в мире увеличивается год от года:

* порядка 278 миллионов человек в 2005 году,
* порядка 360 миллионов человек в 2015 году (+14%),
* порядка 466 миллионов человек в 2019 году (+13%),
* к 2050 году ВОЗ прогнозирует, что это число удвоится.

Традиционным посредником в общении между глухими/слабослышащими людьми и слышащими людьми выступает сурдопереводчик. По законодательству РФ, услуги сурдоперевода частично оплачиваются за счёт федерального бюджета. Однако объём услуги, которую глухой человек может получить за счёт государства относительно небольшой – максимум 40 часов в год. Кроме того, сурдопереводчики – достаточно редкая профессия: по данным Министерства просвещения РФ[[1]](#footnote-1), в 2020-м году в России насчитывалось около 900 сурдопереводчиков, при этом на каждого из них приходилось до 100 тысяч глухих. Также посредничество постороннего человека, может быть не всегда желательно для глухого человека: например, в вопросах, касающихся его здоровья.

Со стороны медицины также постоянно ведутся разработки технологических решений, помогающим глухим и слабослышащим людям ориентироваться в мире звуков. На настоящее время самым эффективным решением в этом направлении является кохлеарная имплантация. Однако она подходит не всем. Кроме того, слышат люди с имплантами несколько иначе, чем слышащие, и вынуждены какое-то время учиться воспринимать и различать звуки с помощью импланта.

Ещё одним мостиком для общения между глухими и слышащими людьми мог бы стать автоматический переводчик с жестового языка на звуковой и в обратном направлении. Несмотря на впечатляющие результаты, достигнутые учёными и разработчиками в этом направлении, необходимое качество такого перевода в настоящий момент не достигнуто. Глухие предъявляют следующие требования к такой системе автоматического перевода [Гриф и др., 2021]:

* точность перевода не менее 90%,
* качество визуализации должно быть максимально приближено к движениям человека-сурдопереводчика.

Почему же при текущем уровне развития машинного перевода с одного жестового языка на другой, для жестовых языков удовлетворительный результат пока ещё не достигнут? С какими сложностями сталкиваются исследователи, работающие с жестовыми языками?

Человек, говорящий на жестовом языке, использует для передачи своего сообщения не только движения рук (которые сами по себе допускают определённую вариативность), но также и движение корпусом (например, при сравнении), выражение лица (в РЖЯ, например, при вопросе необходимо нахмурить брови), «проговаривание» слов губами (чтение по губам может помочь различить похожие между собой жесты, однако не все глухие задействуют губы), движения головой (например, в РЖЯ при вопросе голова как правило движется немного вперед в сторону собеседника), позиции тела (например, при пересказе разговора двух и более лиц). Таким образом, каждый из указанных компонентов несёт в себе определённую смысловую нагрузку, а значит, при упрощении задачи только до движений руками, теряется существенная часть информации, а необходимость учитывать все факторы приводит к усложнению задачи.

С развитием ИИ, и особенно глубокого обучения, исследователи всё больше внимания стали уделять построению именно интеллектуальных систем перевода для жестовых языков. Среди работ последнего десятилетия чётко прослеживается популярность методов глубокого обучения. Например, работы на базе набора данных RWTH-PHOENIX-Weather 2014 чаще всего строят распознавание жестов на базе свёрточной нейронной сети (CNN). Те же работы, которые сфокусированы не на распознавании отдельных жестов, а на переводе непрерывной жестовой речи, чаще всего дополняют CNN лингвистической моделью (как правило, BLSTM) [Koller, 2020]. Тут следует отметить, что на данный момент большая часть работ всё-таки ориентирована на распознавание отдельных изолированных жестов, а не на перевод непрерывной жестовой речи.

В качестве входных данных сейчас наиболее популярны RGB-изображения или видео. Также есть варианты получения данных с сенсоров: электронные перчатки, технологии захвата движения, Kinect), но они существенно более дорогие, и сложно представить их массовое использование при работе с автопереводчиком.

Степень популярности того или иного языка у исследователей определяется в основном наличием общедоступного размеченного набора данных. Лидером по количеству публикаций в настоящий момент выступает американский жестовый язык. Но это верно только для небольших наборов данных (со словарём менее 200 жестов). На более крупных наборах лидирует уже китайский жестовый, а на наборах со словарём более 1000 жестов (RWTH-PHOENIX-Weather 2014) – немецкий жестовый [Koller, 2020].

Немануальные составляющие жестового языка (такие как: движения головы, корпуса, губ, бровей) чаще всего учитываются в работах именно по немецкому жестовому языку [Koller, 2020]. Исследователи РЖЯ также довольно часто принимают во внимание немануальные параметры. Однако в последние годы в мировой практике есть тренд к работе с полнокадровыми изображениями (когда говорящий на жестовом языке полностью попадает в кадр и все его движения рассматриваются как единое целое).

Подробнее о существующих в настоящий момент наборах данных, основных наработках и о трендах в подходах к распознаванию и переводу с жестовых языков (в том числе РЖЯ) см. главу 1.

Исследований по РЖЯ в настоящий момент существенно меньше, чем по американскому, немецкому или китайскому жестовым языкам. Одна из главных проблем тут – отсутствие общедоступных размеченных наборов данных, достаточно больших по объему и словарному запасу. Но это не значит, что нужно отказываться от исследований по тем наборам, которые доступны на данный момент.

**Объект исследования настоящей работы:** возможности автоматизации перевода с русского жестового языка (РЖЯ) на звуковой русский язык.

**Предмет исследования:** разработка подходов к автоматическому распознаванию непрерывной речи на РЖЯ и её переводу на звуковой русский язык методами глубокого обучения.

**Цель исследования:** обучить нейронную сеть (или набор нейронных сетей), которая будет показывать приемлемою точность перевода непрерывной жестовой речи в видеоформате на русский звуковой язык.

**Задачи исследования:**

1. Изучить международную практику в области интеллектуального машинного перевода с жестовых языков, а также наработки отечественных исследователей в отношении РЖЯ.
2. Выбрать несколько наиболее популярных/показывающих хорошую точность архитектур для перевода непрерывной жестовой речи и применить их к поставленной задачи. И применить их к выбранному набору данных.
3. Провести сравнительный анализ полученных результатов.
4. Выбрать модель с лучшим результатом и постараться определить, за счёт каких факторов результат получился лучше.

В качестве набора данных для исследования возьмём Корпус русского жестового языка [Электронный ресурс][[2]](#footnote-2), разработанный под руководством С. И. Бурковой в Новосибирском государственном техническом университете. Подробнее о характеристиках корпуса и его сравнение с другими доступными наборами данных, подходящими для решения поставленной задачи, см. главу 1.

* методологический аппарат исследования: TBD;
* основные пункты новизны, выносимой на защиту: TBD;
* теоретическая и/или практическая значимость исследования: TBD;
* апробация исследования (если таковая имеется) : TBD;
* краткая характеристика структуры диссертации: TBD.

Глава 1. Обзор литературы и доступных данных по теме

В этой главе более детально погрузимся в тему исследования и особенности поставленной задачи. Сначала рассмотрим особенности жестовых языков и их отношение к звуковым языкам. Затем исследуем доступные наборы данных, как по РЖЯ, так и по другим жестовым языкам, сравним их между собой, выделим основные преимущества и недостатке, а также подробнее остановимся на Корпусе русского жестового языка С. И. Бурковой, на основе которого будет вестись дальнейшее исследование. В последней части главы посмотрим, какие подходы применяют международные и отечественные исследователи к решению подобных задач.

* 1. Особенности жестовых языков и РЖЯ в частности

Жестовые языки выполняют ту же функцию средства коммуникации, что и звуковые языки. Это полностью самостоятельные языки со своей грамматикой, морфологией и фонологией: то есть жестовая речь – это совсем не пословесный перевод со звукового языка, а совершенно другой язык со своими собственными правилами. Когда глухие люди учатся читать на основном языке страны, где они живут, они, по сути, учатся читать на иностранном для них языке. В качестве иллюстрации возьмём предложение: «*Папа приготовил шашлык»*. На русском жестовом (если перевести отдельные жесты в слова) это предложение будет сформулировано так: *«Шашлык папа готовить всё»*.[[3]](#footnote-3)

Развиваются жестовые языки также независимо от звуковых. В результате бывает так, что в странах с общим звуковым языком, жестовые языки – разные. Например, американский жестовый язык (ASL) гораздо больше похож на французский жестовый (LSF), чем на британский жестовый (BSL), при том, что звуковой язык у США и Великобритании - один (английский).

В жестовых языках, подобно звуковым, формируются и свои диалекты. В случае с РЖЯ, например, исследователи отмечают различия между Москвой и Санкт-Петербургом[[4]](#footnote-4), между Москвой и Новосибирском[[5]](#footnote-5), также в литературе можно встретить казахско-русский жестовый язык (K-RSL). К сожалению, РЖЯ не так хорошо изучен лингвистами на данный момент, как звуковые языки, и какой-то формализованной информации по основным диалектам РЖЯ нет. Вероятно, это отчасти связано с тем, что официальное признание РЖЯ получил совсем недавно - 30 декабря 2012 (по ФЗ «О внесении изменений в статьи 14 и 19 Федерального закона “О социальной защите инвалидов в Российской Федерации”»).

В жестовой речи помимо собственно РЖЯ используется калькирующая жестовая речь (КЖР), дактиль. В дактильной азбуке каждой букве звучащего языка ставится в соответствие определённая конфигурация (и иногда движение) руки (дактилема). КЖР, как правило, используют для имён, названий, каких-то новых слов, для которых говорящий не знает жеста. Основная же часть коммуникации ведётся посредством жестов, а не дактилем.

Остановимся подробнее на составляющих жестовой речи. За счёт чего говорящий на жестовом языке коммуницирует смысл и своё отношение к сказанному? Современные исследователи жестовых языков выделяют тут 5 основных компонентов:

* *Форма (конфигурация)* руки: как расположены между собой пальцы, согнуты они или выпрямлены, расположены рядом или расставлены в стороны и т.п.
* *Локация* руки: где относительно тела говорящего расположен жест. Локацию подразделяют на место и сеттинг. Место – одна из крупных областей, где может исполняться жест (голова, лицо, шея, грудь, талия, нейтральное жестовое пространство). Сеттинг – детализация положения жеста внутри этой крупной области (если в качестве места выбрать лицо, сеттингами могут быть, напрмимер, правая щека, левый глаз, губы, подбородок и т.п.).
* *Ориентация* ладони – как расположена ладонь в пространстве. Например, ладонь может быть направлена от говорящего (или наоборот), вниз, в сторону и т.д.
* *Движение* руки: форма траектории, по которой движется рука, направление этого движения, поворот ладони (изменение ориентации), изменение конфигурации руки.
* *Немануальные* составляющие: движения корпусом (часто делаются при сравнении), движения головой, плечами, особенности мимики лица (например, нахмуренные брови при вопросе, улыбка при выражении радости), маусинг (проговаривание губами слов – не у всех, и не всегда есть связь со звуковым языком). Положение всего тела в пространстве также может использоваться говорящим, например, при рассказе о группе лиц.

Значение жеста может поменяться иногда при изменении даже одного из этих параметров.

В учебнике С.И. Бурковой [Буркова, 2019] всё многообразие жестов классифицируется по следующим признакам:

* Одноручные и двуручные жесты – в зависимости от того, участвует в исполнении жеста только ведущая рука или обе.
* Статичные и динамичные жесты – по признаку отсутствия или наличия движения руками.
* Симметричные и асимметричные жесты – это деление характерно для двуручных жестов (одинаково или по-разному работают руки в процессе исполнения).
* Синхронные и поочерёдные жесты – в зависимости от того одновременно задействуются руки или по очереди.

На исполнение жестов (кроме симметричных) оказывает влияние ведущая рука. У правшей – это правая рука, у левшей – левая. Соответственно, такие жесты в исполнении левши будут зеркальным отображением этих же жестов в исполнении правши. А значит, при обучении модели распознавания жестов, вероятно, имеет смысл учесть и зеркальные отображения исходных изображений, чтобы нивелировать возможное влияние фактора ведущей руки.

Рассмотрим, какие особенности жестовой речи, по мнению учёных ([Гриф и др., 2021], [Кагиров, Рюмин, 2022]), повышают сложность задачи распознавания и перевода жестовых языков и, в частности, РЖЯ:

* Отсутствие единого стандарта жестовой речи, наличие диалектов, одновременное использование РЖЯ и КЖР.
* Отсутствие достаточно больших по объёму наборов данных для распознавания непрерывной речи на жестовом языке.
* Разная скорость жестикуляции у говорящих.
* Фон влияет на то, как выглядит изображение.
* Эпентеза – дополнительные переходные движения между жестами, которые появляются в непрерывной речи.
* Комбинаторные изменения (вариативность) параметров жестов в непрерывной речи и вытекающая из этого проблема определения границ между жестами.

Остановимся подробнее на составляющих вариативности жестов в непрерывной речи и их аналогах в звуковых языках (табл. 1).

Таблица 1. Комбинаторные изменения в жестовых и звуковых языках

|  |  |
| --- | --- |
| Жестовые языки | Звуковые языки |
| Ассимиляция (уподобление параметров одного жеста параметрам соседнего жеста):   * антиципация - пассивная рука готовится заранее (одноручный жест перед двуручным), * персеверация - пассивная рука вместо возвращение в положение покоя сохраняет своё положение в предыдущем двуручном жесте (одноручный жест после двуручного), * эллипсис - исчезновение некоторых формантов из конфигураций и локализаций между жестами. | Ассимиляция (уподобление артикуляции звуков), аккомодация (частичное приспособление артикуляции смежных звуков), качественная редукция звуков. |
| Редукция жеста (исполнение движения с меньшей траекторией). | Редукция звука (сокращение длительности звука, изменение его качественных характеристик, и иногда даже выпадение). |
| Метатеза (взаимная замена начальной и конечной локализации при исполнении жестов с двойной локализацией). | Метатеза (взаимная перестановка звуков или слогов). |
| Редупликация, выпадение движения при использовании сочетания жестов или повторяющегося движения. | Диереза (выпадение звуков), гаплология (выпадение звуков по причине диссимиляции). |
| Варьируются положение выделенных пальцев (согнуты в разной степени), ориентация ладони, локализация и амплитуда движения. | Диссимиляция (расподобление артикуляции звуков, утрата ими общих фонетических признаков), качественная редукция звука. |
| Эпентеза (добавление жестов). | Эпентеза (вставка звуков). |

Таким образом, мы рассмотрели основные особенности жестовых языков, показали, что РЖЯ и русский звуковой – это разные (по сути иностранные друг для друга) языки, и подробнее остановились на факторах, представляющих определённую сложность с точки зрения задачи распознавания и перевода жестовых языков. Теперь перейдем к вопросу данных, доступных для решения подобных задач.

* 1. Наборы данных в области распознавания и перевода с жестовых языков

Большая часть работ на тему распознавания и перевода жестовых языков опирается в своих исследований на один из общедоступных «контрольных» наборов данных. Контрольными эти наборы называют потому, что благодаря их открытости, наличию разметки и популярности у других авторов, они используются сообществом для сравнения результатов исследований. В таблице 2 приведён сравнительный анализ основных из таких наборов данных, предназначенных именно для решения задач перевода. Для задач распознавания подходит большее количество наборов данных, их основное отличие - в отсутствии перевода на тот или иной звуковой язык.

Здесь важно затронуть вопрос транскрибирования жестовых языков. Первая система аннотации бала предложена Уильямом Стоуки [Stokoe, 1960] для описания американского жестового языка с помощью графических символов. До этого момента жесты рассматривали как неделимые единицы. Затем стали появляться и другие методы. В настоящий момент чаще всего для транскрибирования жестовых языков используют **глоссирование**.

Согласно Большому толковому словарю, **глосса** – это «перевод или объяснение непонятного слова, выражения, помещённое на полях или в самом тексте древних и средневековых рукописей; научный комментарий законов или судебных решений». В случае с аннотированием жестовых языков в соответствие каждому жесту в транскрипции записывается слово, которое он обозначает, иногда вместе с дополнительными общепринятыми символами для обозначения значимых лексико-грамматических особенностей (например, знак # используется для слов, произнесенных дактилем).

В работе [Гриф и др., 2021] предложена разметка для выделения эпентезы и границ жестов, где каждый фрагмент непрерывной жестовой речи разбивается на составляющие: «нет жеста», «идёт жест», «начальное движение», «переходное движение», «конечное движение».

Таблица 2. Контрольные наборы данных для тренировки систем перевода с жестовых на звуковые языки

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Набор данных** | **Язык** | **Тематика** | **Разметка** | **Объём** |
| RWTH-PHOENIX-Weather 2014 T | Немецкий жестовый | Прогноз погоды за 2009 – 2011 на немецком телеканале PHOENIX | Глосс-аннотация;  Письменный перевод на немецкий | выпусков прогноза погоды  глосс  2,887 слов  8,257 видео |
| CSL-Daily | Китайский жестовый | Повседневная жизнь (путешествия, покупки, медицинская помощь) | Глосс-аннотация;  Письменный перевод на китайский | 2,000 глосс  2,343 слов  20,654 видео |
| ASLG-PC12 | Американский жестовый (ASL) | Тексты на английском проекта «Гутенберг» преобразованы в глоссы американского жестового. | Глосс-аннотация;  Письменный перевод на английский | 800 млн. слов/глосс |
| How2Sign | Американский жестовый (ASL);  Английский текст | Перевод на американский жестовый язык видео-инструкций из набора данных How2 | Глосс-аннотация;  Письменный перевод на английский;  Глубина;  2D и 3D – скелеты; | 80 часов видео  16 тыс. слов |

Из таблицы 2 видно, что есть большая проблема с наличием качественных наборов данных для обучения интеллектуальных систем перевода с жестовых языков. Из 4-х популярных наборов данных только для китайского жестового языка собраны данные по темам повседневного общения – наиболее релевантная сфера для применения таких переводчиков. По большинству же жестовых языков, в том числе РЖЯ, отсутствуют даже такие, хоть и относительно ограниченные, но размеченные и общедоступные наборы данных. И в этом одна из основных причин, которая препятствует прогрессу в данном направлении.

Поскольку ни в одном из контрольных наборов данных РЖЯ не представлен, подходящие для исследования данные необходимо искать где-то ещё. Для этого обратимся к публикациям исследователей, работающих над системами распознавания и перевода собственно РЖЯ, а не абстрактного жестового языка. Такими системами достаточно активно в настоящий момент занимаются:

* Новосибирский государственный технический университет (НГТУ);
* Лаборатория речевых и многомодальных интерфейсов СПБ ФИЦ РАН;
* Центром развития технологий AI в Sber AI.

Соберём основные характеристики наборов данных, на которые они ссылаются в своих публикациях в аналогичную таблицу (табл. 3).

Таблица 3 Данные, упоминаемые в публикациях по теме распознавания жестов и перевода с РЖЯ

|  |  |  |  |
| --- | --- | --- | --- |
| **Набор данных** | **Тематика** | **Разметка** | **Объём** |
| Корпус РЖЯ С.И. Бурковой (НГТУ)  <http://rsl.nstu.ru/> | Повседневная | Глоссы для правой руки;  Глоссы для левой руки;  Перевод на русский | 175 видео |
| База данных РЖЯ поликлинического предназначения | Первичный приём у врача-терапевта | Сегменты движения и удержания  Морфологическая (глоссы)  Перевод на русский | 85 фраз;  7 часов видео |
| TheRuSLan | Покупки в супермаркете | Скелет (25 точек);  Перевод на русский язык;  Глоссы (по словам)  Классы жестов; | 154 жеста  8 часов 56 минут видео |
| Сурдосервер 2.0  http://surdoserver.ru/ | Видеоуроки | Нет |  |
| Slovo (Sber AI)  <https://github.com/>  hukenovs/slovo.git | Повседневная бытовая | Глоссы (по словам) | 9.2 часа видео  20400 видео  1000 жестов |
|  |  |  |  |
|  |  |  |  |

Таким образом, из пяти потенциальных источников, наиболее доступным и подходящим для исследования в рамках данной работы представляется Корпус русского жестового языка, разработанный в НГТУ под руководством Бурковой С.И.: датасеты Sber AI, TheRuSLan и база РЖЯ поликлинического назначения закрыты для внешних пользователей, а на Сурдосервере отсутствует разметка. Главные преимущества Корпуса РЖЯ Бурковой – в доступности данных (по запросу открывают доступ) и в наличии профессиональной разметки. Недостатки – в его малом размере (с точки зрения машинного обучения) и в отсутствии возможности выгрузить данные единым архивом.

* 1. Подходы к решению задачи в зарубежной и отечественной литературе

В статье [Koller, 2020] годом начала работы над машинным переводом жестовых языков указан 1983 год со ссылкой на работу [Grimes, 1983]. За это время общепринятые подходы, применяемые исследователями для подобных задач, претерпели ряд изменений. В этой части остановимся подробнее на этих ключевых изменениях и трендах, на том, какие методы и подходы чаще всего применяются к обучению интеллектуальных систем распознавания и перевода с жестовых языков, и на достигнутых результатах.

Задачи, решаемые в этой области, можно разделить на 2 типа:

* распознавание жестов (SLR) - на выходе получаем глоссы
* перевод с жестовых языков (SLT) - на выходе получаем связный текст на целевом языке.

Как правило, для решения задачи перевода поверх задачи распознавания надстраивается дополнительная нейросеть, ответственная за перевод глосс в предложения. Задача распознавания жестов (SLR) также делится на два подтипа: распознавание отдельных жестов (Isolated SLR) и распознавание жестов в непрерывной речи (CSLR). [Koller, 2020] отмечает, что работ по распознаванию изолированных жестов почти в два раза больше. При распознавании непрерывной речи как раз добавляются сложности с выделением эпентезы, распознаванием границ жестов, ассимиляцией жестов и прочими комбинаторными изменениями.

Всю большую задачу перевода с языка жестов можно декомпозировать на следующие подзадачи [Adeyanju et al., 2021]:

* сбор данных,
* предобработка данных,
* сегментация изображения,
* извлечение признаков,
* классификация (тут заканчивается задача распознавания, SLR),
* перевод (тут заканчивается SLT).

Остановимся подробнее на каждом этапе.

1. **Сбор данных**

На этом этапе собирается и размечается датасет. В таблице 4 приведены основные виды входных данных, используемых в таких исследованиях, их преимущества и недостатки.

Таблица 4. Подходы к фиксации жестовой речи

|  |  |  |
| --- | --- | --- |
| **Метод** | **Преимущества** | **Недостатки** |
| RGB - видео | неинтрузивный;  наиболее естественный с точки зрения взаимодействия селовека с компьютором (через камеру);  недорогой. | фон изображения может влиять на результат нежелательным образом;  нет информации о глубине;  качество изображения может различаться у разных камер. |
| RGB-D, Kinect | добавляется информация о глубине:  3D – траектория движения;  легко отделяется фон. | Kinect – дорого и нельзя использовать на улице |
| Разноцветные перчатки | недорогой | требует, чтобы говорящий надевал перчатки (интрузивный метод) |
| Электронные перчатки | Фон не оказывает влияние. | Интрузивный;  Дорого. |
| Технологии захвата движения | Улавливает много деталей движения; | Интрузивный (говорящей надевает сенсоры)  Дорого |

В работе [Koller, 2020] показаны следующие тренды в отношении методов фиксации жестовой речи:

* RGB-видео набирают популярность после 2005 года, тогда же начинается постепенный сдвиг от интрузивных методов (сенсоры) в пользу неинтрузивных (видеозапись).
* RGB-D становятся популярными после 2010 года с появлением сенсоров Kinect.

В данном исследовании этот этап пропущен, так как используются данные из готового набора.

1. **Предобработка изображений** может оказать существенное влияние на конечный результат.

Распространённые варианты предобработки видеоизображений в задачах распознавания жестовой речи:

* Повышение качества изображения. Основная цель здесь – повысить качество изображения так, чтобы его было легче обрабатывать (машине или человеку). В обзоре [Adeyanju, 2021] выделены 4 таких метода с их преимуществами и недостатками (табл. 5).

Таблица 5 Сравнение методов повышения качества изображения

|  |  |  |
| --- | --- | --- |
| **Метод повышения качества изображения** | **Преимущества** | **Недостатки** |
| Выравнивание гистограммы (HE, Histogram equalization) | Простое применение  Высокая эффективность для чёрно-белых изображений | Меняется яркость изображения  Может повыситься контрастность фонового шума.  Сложно провести различие между шумом и полезным признаком |
| Адаптивное выравнивание гистограммы (AHE) | Даёт результат лучше, чем HE.  Подходит для локального увеличения контраста и границ в части изображения | Меняется яркость изображения  Усиливает шумы |
| Адаптивное выравнивание гистограммы с ограниченным контрастом (CLAHE) | Сохраняет яркость  Меньше шума по сравнению с HE и AHE  Подходит для локального увеличения контраста и границ в части изображения | Даёт плохие результаты на изображениях с повышенной яркостью и с несбалансированной контрастностью |
| Логарифмическое преобразование (Logarithmic transformation) | Узкий диапазон низких значений яркости преобразуется в более широкий диапазон, и наоборот. | Не для всех изображений  При применении к светлым изображениям теряется информация |

Самый популярный метод – выравнивание гистограммы (HE).

* Восстановление изображения. Эти техники позволяют восстановить чёткость размытого, испорченного шумами изображения. Для этой цели применяют разного рода фильтры: фильтр средних значений, медианный фильтр, фильтр Гаусса, адаптивный фильтр, фильтр Винера.
* Масштабирование и кадрирование дают возможность сократить время расчётов и объём памяти, необходимой для хранений изображений.
* Преобразование цветного изображения в чёрно-белое также упрощает расчёты и сокращает потребность в ресурсах, сохраняя =при9 этом важные свойства изображения. Однако теряется цветовая информация. Используются также двумерные черно-белые изображения: каждый пиксел представлен не одним, а двумя каналами (интенсивностью белого и интенсивностью чёрного).

1. **Сегментация изображения**

На этом этапе из изображения извлекают сегменты, представляющие интерес с точки зрения поставленной задачи. Например, можно попытаться отделить говорящего от фона. Здесь также существует богатый выбор методов:

* Пороговая обработка – достаточно простой и один из самых популярных методов сегментации.
* Выделение границ объектов на изображении.
* Сегментация по областям изображения.
* Кластерный метод.
* Сегментация с помощью нейросетей (BPNN, FFNN, MLFF, MLP, CNN, SOM).

Согласно исследованию [Koller, 2020], до 2005 года наибольший интерес для исследователей представляла форма руки, и чуть меньше – движения и локация руки, то есть исключительно мануальные составляющие. Позже, с появлением больших наборов данных (со словарём от 1000 жестов), набирают популярность полнорамочные исследования – когда на изображении говорящий учитывается полностью, а значит неманульные компоненты тоже учитываются. В целом наблюдается тренд в сторону глобальных признаков (полнорамочные изображения) вместо выделения отдельных значимых признаков вручную.

Но это только общий тренд, а не полный отказ от локальных признаков в пользу глобальных. Например, в работе [Ryumin, Karpov, 2017] фокус делается на движения губ и их помощь в распознавании отдельных жестов. В статье [Chen et al., 2023] авторы строят модель на основе двух потоков: собственно изображения и выделенные на этих изображениях ключевые точки. В исследовании [Гриф и др., 2021] также использован приём с построением дополнительной разметки в виде скелетной модели (на каждом изображении были выделены 33 ключевые точки с помощью системы MediaPipe Pose).

1. **Извлечение признаков**

Здесь на выделенных на предыдущем шаге сегментах выявляются наиболее значимые для корректной классификации жеста признаки. При применении свёрточных нейронных сетей (CNN) этот этап совмещён со следующим (классификацией). Но есть и другие подходы к выделению таких признаков: метод главных компонент (PCA), дескрипторы Фурье (FD), гистограмма направленных градиентов (HOG), масштабно-инвариантное преобразование признаков (SIFT), ускоренные надежные функции (SURF).

1. **Классификация**

На этом этапе на основе выделенных ранее признаков каждому жесту ставится в соответствие его смысловое значение (глосса). Для классификации применяются следующие популярные подходы: алгоритм K-ближайших соседей (KNN), искусственная нейронная сеть (ANN), метод опорных векторов (SVM), скрытые модели Маркова (HMM), свёрточные нейронный сети (CNN), коннекционистская временная классификация (CTC), нечёткие логические системы, ансамбли и т.д.

1. **Перевод**

Теперь полученные смысловые единицы (глоссы) необходимо перевести в связный текст на языке перевода. Здесь используются: сети с долговременной и кратковременной памятью (LSTM), двунаправленные LSTM (BLSTM), двунаправленные управляемые рекуррентные нейроны (BGRU), механизм внимания (Attention), модель Transformer, mBART

Глава 2. Анализ данных

Сравним, как организованы видео и аннотации к ним в двух выбранных датасетах.

По количеству уникальных слов и фраз лидирует корпус. Преимущество датасета Slovo - в том, что каждое отдельное слово, буква, фраза всегда записано на 20 разных видео, что может оказаться удобнее для целей машинного обучения. В корпусе вполне могут оказаться как слова, присутствующие только на одном видео, так и очень популярные слова.

Разметка в обоих датасетах позволяет определить начало и конец сказанного слова/фразы с той лишь разницей, что в Slovo указаны номера кадров, а в корпусе – миллисекунды.

Ниже подробнее остановимся на каждом датасете.

* 1. Данные из корпуса РЖЯ

На сегодняшний день в корпусе представлено всего **173** видеотекста, продолжительностью **от 16 секунд до 29 минут 32 секунд**. Это относительно длинные видео: не единичные слова или фразы, а полноценные рассказы из связанных между собой фраз. Кто-то рассказывает историю из жизни, кто-то пересказывает сюжет из мультфильма, кто-то – увиденное на картинке. Если поделить все эти файлы на более короткие (на основе временных отрезков, размеченных в аннотации), то получится **51 743** видеофрагмента. Это суммарно как отдельные глоссы, так и полноценные фразы, сформированные из этих глосс: всего порядка **5,2 тыс. уникальных глосс** и **7,8 тыс. уникальных фраз** (более 1 глоссы).В записи видео принимали участие 40 различных носителей РЖЯ из Новосибирска, Москвы и Узбекистана.

Видео различаются по размеру изображения:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Высота изображения** | **Ширина изображения** | **Кол-во**  **видеофайлов** | **Кол-во**  **видеофрагментов** | **Кол-во**  **кадров** |
| **480** | **640** | 70 | 13 391 | 382 991 |
| **480** | **854** | 1 | 1 371 | 37 199 |
| **576** | **720** | 22 | 8 830 | 199 521 |
| **576** | **1024** | 5 | 5 627 | 129 835 |
| **720** | **1280** | 75 | 22 524 | 474 224 |
| **Всего** | | **173** | **51 743** | **1 223 770** |

Видео различаются по количеству кадров в секунду (FPS):

|  |  |  |  |
| --- | --- | --- | --- |
| **FPS** | **Кол-во**  **видеофайлов** | **Кол-во**  **видеофрагментов** | **Кол-во**  **кадров** |
| **23-24** | **1** | 253 | 8 202 |
| **25** | **80** | 30 365 | 618 892 |
| **29-30** | **92** | 21 125 | 596 676 |
| **Всего** | **173** | **51 743** | **1 223 770** |

* 1. Датасет Slovo

Датасет состоит из 20 000 видео от 194 различных носителей РЖЯ. На каждом видео – некоторое часто используемое в повседневном общении слово или фраза. Начало и конец жеста на видео размечены временным интервалом.

Видео разделены на 1000 классов по количеству уникальных букв, слов, фраз РЖЯ. Из них 68 – это фразы (более одного слова).

Кол-во кадров в секунду (FPS) у всех видео одинаковое и равно 30, но сами изображения гораздо больше разнятся по ширине и высоте:

|  |  |  |  |
| --- | --- | --- | --- |
| **Высота изображения** | **Ширина изображения** | **Кол-во**  **видео** | **Кол-во**  **кадров** |
| 358 | 490 | 1 | 45 |
| 494 | 1 | 38 |
| 498 | 1 | 29 |
| 508 | 1 | 45 |
| 514 | 1 | 28 |
| 518 | 1 | 31 |
| 522 | 1 | 35 |
| 534 | 1 | 37 |
| 552 | 1 | 37 |
| 560 | 1 | 53 |
| 580 | 1 | 41 |
| 780 | 8 | 448 |
| 360 | 360 | 258 | 11 956 |
| 440 | 1 | 30 |
| 480 | 1 | 64 |
| 522 | 1 | 136 |
| 528 | 1 | 46 |
| 542 | 1 | 54 |
| 550 | 1 | 39 |
| 558 | 1 | 35 |
| 560 | 2 | 99 |
| 640 | 2 436 | 13 5347 |
| 720 | 5 | 275 |
| 752 | 7 | 362 |
| 786 | 2 | 147 |
| 404 | 358 | 1 | 37 |
| 420 | 358 | 1 | 40 |
| 424 | 360 | 1 | 47 |
| 426 | 358 | 1 | 42 |
| 436 | 358 | 1 | 68 |
| 548 | 360 | 1 | 38 |
| 576 | 360 | 4 | 298 |
| 622 | 358 | 5 | 229 |
| 636 | 360 | 117 | 6 144 |
| 640 | 360 | 15 023 | 756 961 |
| 720 | 360 | 236 | 10 460 |
| 752 | 360 | 23 | 822 |
| 772 | 360 | 32 | 1 907 |
| 780 | 358 | 152 | 6 549 |
| 786 | 360 | 52 | 2 775 |
| 800 | 360 | 15 | 856 |
| 840 | 358 | 1 599 | 74 715 |
| **Всего** | | **20 000** | **1 011 445** |

Глава 3. Подготовка данных

* 1. Подготовка кадров для свёрточной сети

Свёрточным нейронным сетям на входе требуются данные одинаковой размерности. В датасете Slovo видео имеют разный формат, кроме того люди находятся не всегда в центре кадра.

Шаги обработки:

* Поиск лица в кадре.
* Квадрат с учётом размеров лица (должен попасть корпус и руки, которые могут быть и над головой).
* Если в вырезанной часте не хватает пикселей – дополняем их чёрным фоном.
* Берём кадры только между метками начала и конца жеста
* Отбираем нужное количество кадром (если кадров меньше, равномерно дублируем случайную выборку из кадров).

Глава 4. Обучение моделей

* 1. Свёрточная нейронная сеть

Модель: r3d\_18(weights='KINETICS400\_V1')- 18 layer Resnet3D model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель | Обучение | Кол-во классов | Lr=0.0001 | Lr=0.00001 |
| Resnet3D 18 слоёв | Только последний слой | 30 |  |  |
|  | Только последний слой | 1000 |  |  |
|  | Все слои | 30 |  |  |
|  | Все слои | 1000 |  |  |

Список литературы

**Нормативно-правовые акты**

Постановление Правительства РФ от 25 сентября 2007 г. N 608 "О порядке предоставления инвалидам услуг по переводу русского жестового языка (сурдопереводу, тифлосурдопереводу)".

Федеральный закон от 30 декабря 2012 г. N 296-ФЗ "О внесении изменений в статьи 14 и 19 Федерального закона "О социальной защите инвалидов в Российской Федерации".

**Книги**

Большой толковый словарь русского языка. / ред.: С.А. Кузнецов. - Первое издание: СПб.: Норинт - 1998.

Введение в лингвистику жестовых языков. Русский жестовый язык: учебник / ред.: С.И. Буркова, В.И. Киммельман. – Новосибирск: Изд-во НГТУ, 2019. – 356 с.

**Статьи**

Гриф М.Г., Элаккия Р., Приходько А.Л., Бакаев М.А., Раджалакшми Е. Распознавание русского и индийского жестовых языков на основе машинного обучения // Системы анализа и обработки данных. - 2021. - №3(83). - С. 53-74.

URL: <https://journals.nstu.ru/vestnik/catalogue/contents/view_article?id=27523>

Гриф М.Г., Королькова О.О., Приходько А.Л. Распознавание жестовой речи с учётом комбинаторных изменений жестов // Материалы XXI Международной научно-методической конференции «Информатика: проблемы, методы, технологии». - В.: Вэлборн, 2021. С. 1387-1393.

Кагиров И.А., Рюмин Д.А. База данных русского жестового языка поликлинического предназначения: лингвистические особенности материала и аннотирования. // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2022 - №20(3). – С. 90-108.

URL: <https://doi.org/10.25205/1818-7935-2022-20-3-90-108>

[Adaloglou](https://paperswithcode.com/author/nikolas-adaloglou) N., [Chatzis](https://paperswithcode.com/author/theocharis-chatzis) T., [Papastratis](https://paperswithcode.com/author/ilias-papastratis) I., [Stergioulas](https://paperswithcode.com/author/andreas-stergioulas) A., [Papadopoulos](https://paperswithcode.com/author/georgios-th-papadopoulos) G.T, [Zacharopoulou](https://paperswithcode.com/author/vassia-zacharopoulou) V., [Xydopoulos](https://paperswithcode.com/author/george-j-xydopoulos) G.J., [Atzakas](https://paperswithcode.com/author/klimnis-atzakas) K., [Papazachariou](https://paperswithcode.com/author/dimitris-papazachariou) D., [Daras](https://paperswithcode.com/author/petros-daras) P. A Comprehensive Study on Deep Learning-based Methods for Sign Language Recognition

URL: https://arxiv.org/pdf/2007.12530v2.pdf

Adeyanju I.A., Bello O.O., Adegboye M.A. Machine learning methods for sign language recognition: A critical review and analysis // Intelligent Systems with Applications – 2021. – Vol.12. – P.2667-3053.

URL: <https://doi.org/10.1016/j.iswa.2021.200056>

Achraf Othman and Zouhour Tmar. “English-ASL Gloss Parallel Corpus 2012: ASLG-PC12, The Second Release”. Fourth International Conference On Information and Communication Technology and Accessibility ICTA’13, Hammamet, Tunisia, October 24-26, 2013.

Duarte, Amanda and Palaskar, Shruti and Ventura, Lucas and Ghadiyaram, Deepti and DeHaan, Kenneth and Metze, Florian and Torres, Jordi and Giro-i-Nieto, Xavier. A Large-scale Multimodal Dataset for Continuous American Sign Language / Conference on Computer Vision and Pattern Recognition (CVPR). – 2021. URL: https://arxiv.org/pdf/2008.08143v2.pdf

Koller O. Quantitative survey of the state of the art in sign language recognition // URL: [https://arxiv.org/abs/2008.09918v2 - 2020](https://arxiv.org/abs/2008.09918v2%20-%202020).

Ryumin D., Karpov A. Parametric representation of the speaker`s lips for multimodal sign language and speech recognition. // The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences – 2017. – Vol.XLII-2/W4. – P.155-161

Chen Y., Zuo R., Wei F., Wu Y., Liu Sh., Mak B. Two-Stream Network for Sign Language Recognition and Translation. // 2023. URL: https://arxiv.org/pdf/2211.01367v2.pdf

**Интернет-источники**

«В России отмечают День сурдопереводчика» / Пресс-служба [Минпросвещения России](https://edu.gov.ru/), 2020. URL: <https://edu.gov.ru/press/3084/v-rossii-otmechayut-den-surdoperevodchika/>

Онлайн-курс русского жестового языка (РЖЯ) / РООИ «Перспектива».

URL: <https://edu.perspektiva-inva.ru/courses/kurs-russkogo-zhestovogo-yazyka/>

Разработка модели машинного перевода с русского жестового языка / AIRI Seminars.

URL: https://youtu.be/OHQ8MmIJmEM

Сайт негосударственного образовательного частного учреждения дополнительного профессионального образования «Учебно-методический центр Всероссийского общества глухих» (НОЧУ ДПО «УМЦ ВОГ»). URL: https://nochuvog.ru/

Сайт общероссийской общественной организации инвалидов «Всероссийское общество глухих» (ВОГ). URL: <https://voginfo.ru/>

**Наборы данных**

Корпус русского жестового языка [Электронный ресурс] / Руководитель проекта С. И. Буркова – Новосибирск – 2012–2015 – Режим доступа: [http://rsl.nstu.ru/](http://rsl.nstu.ru/site/signlang)

CSL-Daily: Chinese Sign Language Corpus / University of Science and Technology of China (USTC) // Multimedia Computing & Communication, SLR Group.

URL: <http://home.ustc.edu.cn/~zhouh156/dataset/csl-daily/>

English-ASL Gloss Parallel Corpus 2012: ASLG-PC12

URL: <https://achrafothman.net/site/english-asl-gloss-parallel-corpus-2012-aslg-pc12/>

How2Sign (A Large-scale Multimodal Dataset for Continuous American Sign Language) URL:  <http://how2sign.github.io/>

WTH-PHOENIX-Weather 2014 T: Parallel Corpus of Sign Language Video, Gloss and Translation / RWTH Aachen University // Human Language Technology & Pattern Recognition Group

URL: <https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/>

Kapitanov, A., Kvanchiani, K., Nagaev, A., & Petrova, E. (2023). Slovo - Russian Sign Language Dataset and Models*.*

URL: https://arxiv.org/abs/2305.14527

1. «В России отмечают День сурдопереводчика» / Пресс-служба [Минпросвещения России](https://edu.gov.ru/), 2020. URL: https://edu.gov.ru/press/3084/v-rossii-otmechayut-den-surdoperevodchika/ [↑](#footnote-ref-1)
2. http://rsl.nstu.ru//site/signlang [↑](#footnote-ref-2)
3. Пример взят из онлайн-курса русского жестового языка (РЖЯ) РООИ «Перспектива»: https://edu.perspektiva-inva.ru/lessons/3-vvedenie-v-grammatiku/ [↑](#footnote-ref-3)
4. https://nochuvog.ru/russkij-zhestovyj-yazyk-rzhya/ [↑](#footnote-ref-4)
5. https://voginfo.ru/vog/2015/05/30/rslidioms/ [↑](#footnote-ref-5)