В первый год мы должны были сделать следующее

1. Обзор литературы по распознаванию рукописного текста
2. Сбор данных
3. Проектирование системы

Второй год – разработка собственной модели оптического распознавания рукописного текста и его отладка

1. Разработка моделей и проведение опытов на собранном датасете
2. Выбор определенной модели
3. Тестирование и отладка модели

Третий год – построение программного обеспечения

1. Разработка программного продукта на основе нашей модели
2. Проектирование пользовательского интерфейса системы
3. Тестирование программного продукта

Вывод

Что удалось реализовать на данный момент? Что еще предстоит сделать?

Написано несколько статей.

Собран и аннотирован приличный датасет рукописного текста на русском и английском языках.

За это время отработана и отшлифована методология сбора данных. Данный процесс по мере возможности был автоматизирован. Для сбора данных были использованы новейшие технологии и сервисы. К примеру, были отсканированы контрольные работы студентов, с использованием ранее написаных модулей эти работы были сегментированы на слова. Далее был написан telegram bot, который высылает картинку слова, а пользователь печатает в ответ слово, которое там написано. Таким образом мы аннотируем данные.

**СОДЕРЖАНИЕ**

1. Введение (5 стр.)

* В чем заключается задача? Почему это важно?
* В чем заключается новизна нашей научно-прикладной работы?
* Может ли наше решение быть применено на практике?

2. Проектирование системы (5 стр.)

3. Обзор литературы по распознаванию рукописного текста (5 стр.)

4. Методология (20 стр.)

4.1. Сбор данных (10 стр.)

4.2. Выбор моделей (5 стр.)

4.3. Калибрация моделей распознавания (5 стр.)

4.4. Валидация моделей распознавания

4.5. Корректировка результатов распознавания с использованием априорных знаний

5. Разработка программного обеспечения (10 стр.)

5.1. Архитектура системы

5.2. Графический пользовательский интерфейс

5.3. Валидация программного обеспечения

5.4. Характеристика программы и лицензия (1 стр.)

5.5. Тестирование и применение

Программа будет – веб-сервис. Монетизация по подписке на месячной основе. Клиент покупает годовую или же месячную лицензию.

6. Заключение (5 стр.)

**1. ВВЕДЕНИЕ**

*В чем заключается наша задача?*

Задача заключается в распознавании рукописного текста на почтовом конверте. Текст предполагается на любом из трех языков (казахский, русский, английский). Распознавание рукописого адреса может быть применен для своего рода автоматизация процесса сортировки или же банального ввода данных в систему почтовой службы. Данный модуль может быть использован как в комплекте с аппаратно-программным комплексом по сортировке писем, так и для ручной обработке писем. Согласно открытым источникам *(*[*http://zviazda.by/ru/news/20191009/1570601704-kak-sortiruyut-pochtovye-otpravleniya*](http://zviazda.by/ru/news/20191009/1570601704-kak-sortiruyut-pochtovye-otpravleniya)*),* письма со стандартными габаритами отправляются на автоматическую сортировку. Письма нестандартного размера, веса или толщины отправляются на ручную сортировку. Под «нестандартным» размером здесь понимается размер конверта, который не является регламентированным в конкретном регионе. Подобное расхождение стандартов может происходить в силу того, что практически у каждой страны своя почтовая служба, а также множество частных компаний, предоставляющих услуги почты. Это привело к тому, что размеры некоторых конвертов могут не соответствовать международным стандартам. Если даже они могут быть регламентированными стандартами, то авто-сортировщик может быть не приспособлен к работе с определенными размерами. В контексте данной задачи мы концентрируемся больше на ручной сортировке писем, так как считаем, что данный аспект является наиболее уязвимым в процессее сортировки. Однако, мы хотели бы еще раз подчеркнуть, что решение, которое разработано в данной работе может быть легко использовано в качестве модуля распознавания в аппаратно-программном комплексе автоматической сортировки.

К примеру, оператору поступило письмо, которое он сомневается в какую стопку положить. Он сканирует лицевую часть конверта, и система выдает ему автоматизированный ответ на основе адреса получателя. Система работает в реальном времени.

Всю задачу можно разделить на два основных подраздела:

* Входящая корреспонденция (т.е. почтовое отделение является пунктом назначения)
* Исходящая корреспонденция (т.е. почтовое отделение принимает письмо на отправку).

Во втором случае почтовому оператору необходимо вбить адрес в систему (в случае если письму присваивается трэкинговый номер). Обычно этот процесс занимает относительно продолжительное время, а также допускаются множетвенные ошибки при наборе текста, т.к. не все операторы владеют навыками скоропечатания. Данная система была бы полезна именно для этого этапа работы почтового сервиса, что значительно ускорило бы уровень обслуживания клиента. Уровень обслуживания клиента частично также определяется временем, затраченным на его обслуживание. В результате увеличения качества обслуживания клиентов рейтинг компании и доверие клиентов к компании могут возрасти.

*Как работает почта?*

*Автоматическая сортировка писем появилась в СССР еще в 60-е годы. Тогда было введено в эксплуатацию около 180 письмо-сортировочных машин, а население приучили писать индексы на конвертах. Но с началом 90-х годов Россия перешла на европейский формат написания почтовых адресов, к которым машины не были приспособлены. В итоге «Почта» вынуждена была вернуться к ручной сортировке.*

*Корреспонденцию внутри одного региона сначала направляют в центр – это быстрее, чем осуществлять сортировку и регистрацию писем на месте. Вся бумажная корреспонденция автоматически делится на стандартные (по габаритам) и нестандартные письма. После этого их автоматически укладывают лицом в одну сторону с учетом направления текста в адресной части, затем машина наносит на письма календарный штемпель.*

*Рассортированные по размерам письма складывают в контейнеры и отправляют в машину для кодирования. В начале процесса кодировочные аппараты преобразуют цифровой или буквенный адрес в условный код, который печатается на письме. Далее закодированную информацию о почтовом адресе сканируют устройством для считывания символов (оборудование в том числе поддерживает распознавание кириллицы). В случае невозможности распознать индекс или адрес отсканированное изображение поступает на монитор операторам, которые вводят информацию вручную. После этого письма автоматически разделяют по адресным ячейкам. Выемка ящиков с отсортированной почтой и подготовка документов для отправки осуществляется вручную.*

*На участок ручной сортировки писем поступает почта, которая не подлежит обработке в автоматической машине, например, из-за дефекта упаковки, неправильно написанного адреса или недопустимого вложения внутри письма (не с точки зрения цензуры, а потому что могут повредить машину или упаковку). Корреспонденция сортируется по индексам. В зависимости от адреса оператор раскладывает их по определенным ячейкам.*

*На участке сортировки обрабатывают около 8 тыс. посылок за час. Оператор устанавливает их на линию, после чего проходит взвешивание. Если вес допустим по нормативам (до 20 кг), посылка считывается сканером и попадает на ленту, если нет — машина останавливается, и подключается ручной труд.*

*После того как посылка прошла по конвейерной ленте, она выпадает в определенную ячейку, которой присваивается штрих-код. Это сделано для того, чтобы посылки не терялись в пути следования. Вся информация об отправлении, таким образом, попадает в единую систему Почты России.*

*Рассортированную почту загружают в машины и развозят по отделениям. Грузовики приезжают сюда 2 раза в сутки: утром и вечером. Посылки и письма заносят в разные комнаты и регистрируют. В отделе доставки письма вручную штампуют, нанося дату, время приема и координаты почтового отделения. Далее оператор раскладывает их по ящикам с номером участка, откуда газеты и письма забирают почтальоны.*

*Заказные письма и посылки взвешивают и с помощью прибора со штрих-кодом заносят в специальную программу. Она позволяет отслеживать статус посылки (посмотреть его можно на официальном сайте Почты России, при условии, что вы знаете ваш номер почтового отправления). Далее выписывают почтовые извещения, а сами посылки уносят на склад. Срок хранения — месяц. Если получатель так и не обнаруживается, то их возвращают адресату.*

*Срок получения письма или посылки может сократить сам получатель, если будет придерживаться определенных правил. Во-первых, корректно указать индекс: 30% отправлений задерживаются из-за орфографических ошибок и опечаток. Во-вторых, если вы не можете забрать посылку в рабочий день и ждете выходных, то срок увеличивается до 18:00. Сотрудники предлагают воспользоваться, например, системой почтоматов: при заказе товара в интернет-магазине выберите способ доставки: «Почта России» — через почтоматы и отметьте ближайший к вам. Когда товар поступит в ячейку, вам придет СМС-уведомление и код доступа. На сегодняшний момент в городе действуют 19 почтоматов, один из которых, на Почтамтской улице, работает 24 часа.*

*В чем заключается новизна нашей научно-прикладной работы?*

Во-первых, наш научный вклад заключается в том, что был собран и аннотирован довольно обширный датасет рукописных слов на казахском и русском языках. (Сколько конкретно слов собрано и аннотированно?). По количеству собранных и аннотированных слов, насколько нам известно, нашему датасету нет аналогов.

Во-вторых, для сбора данных были использованы новейшие технологии, такие как использование телеграм ботов для аннотации данных. Другими словами, подобный метод аннотации рукописного текста практически не применялся до этого. Он показал свою эффективность на практике. Он позволяет ускорить скорость аннотации, посредством расширения аудитории людей готовых посвятить несколько минут на аннотацию слов.

Во-третьих, современые модели компьютерного зрения по разпознаванию были адаптированы и калиброваны на собранном датасете. В основном эти модели применялись в общем контексте распознавания объектов в различных сценах, печатного или же рукописного английского текста. Однако, ни одна из современных моделей не применялась конкретно для распознавания рукописного почтового адреса. Тот факт, что эти модели были целенаправлено калиброваны на распознавание казахского и русского алфавита увеличивает значимость вклада и подчеркивает его важность.

В-четвертых, в данном проекте были использованы априорные знания, а имеено контекст задачи (структура лицевой части конвертов и содержание текста). Мы твердо уверены, что использование априорных знаний о положении текста, свойственном в почтовой кореспонденции, содержании каждого из полей дает преимущество перед более общей задачей распознавания рукописного текста. Другими словами, используя эти знания, точность распознавания может быть значительна повышена. В данной работе предложен эффективный метод коррекции результатов распознавания, который позволяет сократить пространство поиска, опираясь на априорные знания.

*Может ли наше решение быть применено на практике?*

**2. ПРОЕКТИРОВАНИЕ СИСТЕМЫ**

**Уровень абстрации 1**

Основная задача проекта – это сортировка входящей в Казахстан письменной корреспонденции.

**Уровень абстрации 2**

**Уровень абстрации 3**

Конверты очень сильно отличаются друг от друга: по размеру, цвету и содержанию. К примеру, на конвертах может быть разное количество марок, паразитный текст или положение областей получателя-отправителя могут быть смещены или вообще отсутствовать. Следовательно, невозможно обеспечить 100% точность работы системы. В связи с этим, в случае некорректного автоматического срабатывания системы, мы предоставляем оператору возможность вручную указывать область снимка, где находится область получателя.

**Уровень абстрации 4**

Письменная корреспонденция может быть как внутренней, так и международной, поэтому нам необходимо рассматривать несколько языков: казахский, русский, английский.

Для большей точности рекоммендуется иметь отдельную модель распознавания для каждого языка.

Допустим мы разбили область снимка с данными получателя на слова. Предположим, что в этом наборе слов обязательно есть название страны. В случае, если страна не указана, то по умолчанию корреспонденция считается внутренней.

Внутренняя переписка может вестись на казахском и русских языках. Переписка по странам СНГ может вестись на русском языке. Международная переписка – на английском.

С учетом этого допущения, для определения языка нам необходимо иметь словарь ключевых слов для каждого языка. В данном случае, ключевыми словами будут считаться названия стран ( слов если учитывать сокращения и т.д.).

Следовательно, мы пробегаемся по всем картинкам, распознавая слова с использованием одной из моделей распознавания, и сравниваем с каждым словом в базе данных ключевых слов. С использованием определенной метрики выбираем наиболее вероятную модель распознавания языка.

Допуская, что модель распознавания корректно выбрана и страна корректно определена. Для каждой страны мы храним словарь ключевых слов, который включает в себя названия провинций, областей, городов (<1000 слов). Этого должно быть достаточно для сортировки писем для межгородских сообщений.

Следовательно, для коррекции результатов распознавания мы можем находить степень соответсвия ключевым словам и заменять соответствующими словами.

**2020 год**

этот год в основном работали над моделями, результаты сравнения описали в приложенных статьях. По Интерфейсу ничего нового нет, только то что реализовал в прошлом году Жулдызжан. Про интерфейс можешь найти в отчете прошлого года.

Запустили новый метод сбора данных с использованием телеграм бота. Отсканировали экзаменационные листы студентов, сейчас около 2000 страниц А4, планируется еще около 4000. Сегментировали слова и и запустили аннотацию. До этого делали телеграм бот, где был обратный принцип, мы отправляли слова пользователям, а они должны были прикреплять фото с рукописным текстом. Но этот метод не пошел, так как очень много телодвижении.

2000 листов сегментировали 180000 слов