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 наш вклад и новизна научно-прикладной работы?
* Может ли наше решение быть применено на практике?

2. Аппаратно-программные комплексы

Siemens mail sorting solutions

2. Проектирование системы (5 стр.)

2.2. Графический пользовательский интерфейс

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. Валидация программного обеспечения

5.3. Характеристика программы и лицензия (1 стр.)

5.4. 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).

*В чем актуальность данной задачи?*

Вопрос распознавания рукописного текста является широко и активно изучаемым вопросом в научном мире. На данный момент не существует готовых решений, которые могли бы гарантировать высокую точность распознавания рукописного текста, что не позволяет большим компаниям, таким как Siemens (Рис. 1) или Blueshift (Рис. 2), внедрить ее в массовое производство. Их решения автоматической сортировки писем построенны на распознавании печатного адреса, но не рукописного.



### Рисунок 1. Один из аппаратно-программных комплексов Siemens для сортировки писем.

### (<https://www.siemens-logistics.com/en/mail-sorting/letter-sorting-and-sequencing>)

Все письма, которые не имееют печатного адреса сортируются вручную (Рис. 3). Как показывает практика, ручная сортировка имеет множество недостатков. К ним относится большие временные затраты, а также ошибки, связанные с человеческим фактором.

В данной работе мы хотели бы изучить потенциальные решения того, как можно было бы повысить эффективность ручной сортировки или же сократить усилия, которые она требует, что позволило бы сократить время обслуживания и повысить точность. Если говорить более конкретно, то мы рассматриваем задачу распознавания рукописного адреса как некий предварительный шаг, необходимый для дальнейшей автоматической сортировки. К примеру, распознавание рукописого адреса может быть использован для банального автоматического или полу-автоматического ввода данных в систему почтовой службы. Это позволило бы ускорить процесс ручной сортировки письменной корреспонденции, так как, имея адреса отправителя и получателя в электронном виде, оператор почтовых услуг мог бы представить в любой удобной для него форме (штрихкод, печатный текст и т.д.). Допустим, оператор решает печатать электронный текст стикерах, которые затем помещаются на письма, поверх рукописного текста. Этот предварительный шаг позволяет подготовить письма для дальнейшей автоматической сортировки. Следовательно, наши наработки могли бы быть потенциально использованы в комплекте с аппаратно-программным комплексом по сортировке писем в качестве модуля предобработки.

### 

### Рисунок 2. Vantage Sorting System, установленная в Минске.

### http://zviazda.by/ru/news/20191009/1570601704-kak-sortiruyut-pochtovye-otpravleniya

### Кі­ра Ера­шэ­віч — сар­та­валь­нік паш­то­вых ад­праў­лен­няў і дру­ка­ва­ных срод­каў ма­са­вай ін­фар­ма­цыі цэ­ха пісь­мо­вай ка­рэс­пан­дэн­цыі і гіб­рыд­най пош­ты вы­твор­час­ці «Мін­ская пош­та» РУП «Бел­пош­та».

### Рисунок 3. Ручная сортировка писем нестандартного размера или же с рукописным почтовым адресом

### http://zviazda.by/ru/news/20191009/1570601704-kak-sortiruyut-pochtovye-otpravleniya

Всю задачу можно разделить на два основных подраздела:

* Входящая корреспонденция
* Исходящая корреспонденция

В первом случае, опреатору может понадобиться внести данные о письме в реестр (что оно принято в отделении) и доставить конечному адресату. Внесение данных в реестр, к примеру, может понадобиться когда письмо заказное или же необходимо указывать его передвижение и статус в онлайн системе трэкинга корреспонденции. Если у письма имеется штрихкод, то проблем не возникает и необходимо просто отсканировать штрихкод. Если же у письма нет штрихкода, то оператор вынужден вбивать данные в систему вручную, что может занять у него некоторое время и потенциально могут быть допущены ошибки.

Во втором случае почтовому оператору необходимо вбить адрес в систему (в случае если метаданные письма вносятся в реестр компании и генерируется штрихкод или же трэкинговый номер). Обычно процесс вбивания адреса в ручную занимает относительно продолжительное время, а также допускаются множетвенные ошибки при наборе текста, т.к. не все операторы владеют навыками скоропечатания.

Наше решение, построенное на распознавании рукописного адреса, могло бы быть полезно именно для этого этапа работы почтового сервиса. С его использованием провайдер почтовых услуг мог бы значительно ускорить процесс обслуживания клиентов. Так как Уровень обслуживания клиента частично также определяется временем, затраченным на его обслуживание. В результате увеличения качества обслуживания клиентов рейтинг компании и доверие клиентов к компании могли бы со временем возрасти.

*Как предположительно выглядит процесс сортировки?*

*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).*

*Срок получения письма может сократить сам получатель, если будет придерживаться определенных правил. Корректно указать индекс: 30% отправлений задерживаются из-за орфографических ошибок и опечаток.*

*В чем заключается вклад и 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.

**Siemens mail sorting solutions**

<https://www.siemens-logistics.com/en/parcel-logistics/reading-and-coding>

One of our core competences is the extraction and interpretation of all relevant sorting and routing information for mail and parcels, including weight and dimensions. Different laser scanners, line-scanning cameras, scales and dimensioning systems provide a fully integrated, service-based solution.

**ARTread – High-performance Reading**

Our automatic address reading ARTread sets benchmarks in terms of speed, flexibility and accuracy. Even for items without routing information or an ID barcode, or when the corresponding information is not available, it is possible to determine the routing or the delivery point information. Mail and parcel reading technology from Siemens can decipher machine as well as hand-written addresses in various alphabets. It also recognizes non-address attributes such as stamps, stickers and labels, as well as markings for dangerous goods and handling instructions.

**ARTid – Fingerprint/Parcel ID**

Our fingerprint technology makes use of the fact that no two mail pieces or parcels look exactly alike. Optical characteristics such as surfaces, stamps and logos are recorded, along with the mail piece’s size, volume and weight. The image-based identification technology thus recognizes a mail piece or parcel throughout the entire process chain. The image can be used for in-hub and cross-border tracking, as well as for third-party parcel tracking.

**Video Coding: Fully flexible and non-hierarchical manual encoding**

Scanned images that cannot be interpreted by the automatic reading function are manually coded by video-coding operators. Video coding can be performed online when the parcel or letter is processed in the sorting machine, or offline when, for example, the item is being transported to the subsequent sorting center.

**Automatic Learning System: Continuous improvement of address recognition**

Reading performance is highly dependent on the quality of the underlying address directory. We therefore recommend the use of our learning system to continuously update address directories. This system “learns” typical spelling errors and updated addresses – and it also identifies new elements frequently occurring on parcels or mail.

**Label Learning – mastering the increasing diversity of address labels**

Innovative Label Learning from Siemens enables customers to easily add new reading tasks and train the reading system to recognize new label layouts. It can resolve corrupted or non-readable barcodes and extract additional parcel information, or resolve conflicting identifier barcodes on reused parcel packaging. It is even possible to process international and third-party parcel or mail labels.

The Intelligent Data Entry Workplace (IDEW) is a flexible, compact address-reading and manual-encoding station for parcels and mail. The solution incorporates various types of image-capturing devices, like a top camera or a handheld scanner combined with label printer. Optionally, a scale can be integrated to semi-automatically obtain the relevant parcel data.

#### Advanced Color Recognition (ACR)

Colorful stamps, blue meter marks, black PPIs, tiny 2D codes and even no marks at all – recognition machines operating with Advanced Color Recognition (ACR) can handle this and more with ease. For many years and with great success, Siemens has been relying on ACR technology that, in the meantime, achieves facing rates close to 100%. The latest scanner generation CLSC 3004 provides images at full C4 height for facing and all other recognition tasks, up to address reading with true 256 dpi resolution. ACR is the solution of choice for outstanding facing rates.

#### Additional Functionalities

Address reading (OCR) and offline video coding for letters, printing and verifying ID and/or destination bar code for letters, revenue protection, and franking/billing service for letters can also be integrated into the CFC 3004.

## Intelligent and proven solutions – from the first to the last mile

From pre-processing to sequencing – our portfolio of proven mail sorting systems offers you high throughput and availability for processing letters or flats. Siemens mail sorting systems perform every sorting task, whether postcards or parcels up to a thickness of 32 millimeters.

### Benefits

* Innovative technology and field proven modules for outstanding operating efficiency, process quality and system performance
* Highest throughput
* High machine availability
* Low maintenance requirements
* Highly configurable functionality for letter sorting and sequencing on minimized floor space
* Optimal ergonomic working conditions
* No stacker walls on both levels
* Compact Sequence Sorter (CSS) and Reader Sorter (CRS) with C4 capability

**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