Методы оптического распознавания текста на арабском языке[1]

О. А. Берникова, к. фил. н., *А. А. Бояров*, *О. И. Редькин*, д. фил. н., *А. А. Сенов*[2]
Санкт-Петербургский государственный университет
"Лаборатория анализа и моделирования социальных процессов: политический ислам / исламизм: теория и практика в сравнительной и исторической перспективе"
bernikova@mail.ru, andrei.boiarov@gmail.com, oleg_redkin@mail.ru, alexander_senov@gmail.com

Одним из самых известных применений искусственного интеллекта является оптическое распознавание символов. Эта технология нашла широкое применение в различных прикладных областях. В последнее время были получены значительные результаты для распознавания латинских и славянских символов, однако арабский язык имеет свою специфику, не позволяющую применять такие подходы. В этой работе проведено исследование методов оптического распознавания текстов на арабском языке.

Ключевые слова: оптическое распознавание символов, арабский язык, машинное обучение, кластеризация.

1. Введение

Оптическое распознавание символов (optical character recognition, OCR) — одно из самых известных и широко используемых видов реализации технологий искусственного интеллекта. Он применяется для перевода документов и книг в электронный вид, извлечения информации из изображения. OCR используется для считывания и обработки информации со сканированных документов, фотографий, рукописных текстов и т. д. Современные системы оптического распознавания символов применяют различные технологии обработки изображения и комбинации разнообразных алгоритмов машинного обучения, таких как нейронные сети, машина опорных векторов (SVM), методы кластеризации, логистическая регрессия и т. д.

Актуальность разработки методов оптического распознавания

[1]Работа проведена в рамках проекта СПбГУ "Математическая модель распознавания и процессинга текстов на восточных языках на основе сегментации релевантных составляющих" (Мероприятие 1, Шифр ИАС 0.37.102.2011).

арабского текста обусловлена комплексом факторов. В настоящее время арабская графика, помимо собственно арабского языка, лежит в основе письменности персидского, урду, пушту, дари, кашмири, пенджаби, синдхи, хауса, фула, курдского (в Иране и Сирии), уйгурского, а также ряда других языков и используется в ареале Северной и Западной Африки, Ближнего Востока, южной и юго-восточной Азии с населением около 1 млрд. человек. В прошлом географическая дистрибуция арабской графики была еще шире. Несмотря на это, имеющиеся решения для ее распознавания не отвечают в полной мере запросам пользователей. Эффективность существующих продуктов распознавания для арабского языка, таких как Sakhr (см. http://www.sakhr.com/ocr.aspx), в значительной степени зависит от структуры анализируемых материалов и функционирует с минимальным количеством ошибок лишь при условии работы с "идеальным" текстом (набранным с помощью одного из наиболее распространенных шрифтов, лишенным огласовок и т.д.). Очевидно, что проблема распознавания символов в тексте на арабском языке сложнее, чем в текстах на латинице или кириллице. Данное обстоятельство во многом обусловлено проблемами, как лингвистического, так и технологического характера. Трудность распознавания арабской графики обусловлена и большим количеством дериватов, "слитным" характером письма, допускающим различную длину соединительных линий, возможность реализации точек в стороне от буквы, наличие лигатур, слитное написание ряда предлогов и частиц [16].

Распознавание рукописных текстов на арабском языке представляет собой еще более трудную задачу. При этом важность разработки такого рода технологии очевидна. Средневековые рукописи считаются не только одним из важнейших элементов арабо-мусульманского наследия, но и являются важным источником по истории Ближнего Востока и Северной Африки, а также Средней Азии, Кавказа, Европы и России, что нашло свое отражение в работах В.В. Бартольда, И.Ю. Крачковского и А.Б. Халидова и В.И. Беляева. Значение рукописного наследия стран мусульманского востока, написанного, как правило, на арабском языке, трудно переоценить, и работа по его изучению является одной из центральных задач современного востоковедения. Принципы текстологического, и, в более узком смысле этого слова, лингвистического анализа рукописного текста базируются на четко выверенной и

проверенной временем парадигме описания манускриптов. Такого рода парадигма предусматривает их хронологизацию, лексический и терминологический анализ, атрибутику с одной из рукописных школ, изучение исторического или лингвистического контекста, сопутствующего их созданию. Традиционно описание рукописей ведется исключительно на основе субъективного восприятия исследователя с участием человеческого фактора без определения детальных, метрических характеристик рукописного текста. Тогда как такие формальные показатели, как цветовые характеристики текста, его "ритмика", индивидуальные особенности почерка переписчиков или авторов манускриптов могут быть объектом описания на основании характеризующих их цифровых параметров.

До последнего времени цифровая обработка рукописей сводилась к их сканированию, форматированию полученного изображения и размещению в сети Интернет с созданием локальных баз данных, при этом изображение ограничивалось форматом 2D. Вместе с тем, необходимость ввода в научный оборот дополнительного материала требует нового подхода, предварительного исследования и обработки рукописей, что обуславливает необходимость новых технологических решений, осуществляемых в результате совместной работы специалистов в области компьютерного программирования и арабистов–исламоведов, филологов и лингвистов.

Библиотеки и научные центры Санкт-Петербурга располагают значительными рукописными фондами, так, в Восточном отделе библиотеки им. Горького СПбГУ находится более 900 арабских рукописей. Наличие такого большого банка данных делает возможным использование современных средств интеллектуального анализа данных (Data Mining), на основе которых могут быть автоматически получены объективные количественные и качественные характеристики арабографических текстов, такие как стиль письма, дата написания, авторство. Также может быть произведен автоматический анализ всего банка текстовых данных для выявления полезной аналитической информации. Например, могут быть обнаружены взаимосвязи между отдельными группами текстов, проанализирована эволюция какого-либо из стилей, построена общая тенденция развития языка. Такие задачи являются актуальными для современной арабистики. Кроме того, они позволяют расширить сферы применения некоторых алгоритмов интеллектуального анализа данных, исследовать их поведение и получить полезный

опыт, важный для дальнейших исследований.

В результате работы был предложен и программно реализован алгоритм распознавания печатных арабских текстов фиксированного шрифта и размера. Основная идея метода заключается в извлечении признаков из изображения при помощи *обучения признаков без учителя (unsupervised feature learning)*, и последующей тренировки классификатора на выборке размеченных изображений в пространстве извлеченных признаков.

2. Предыдущие исследования

Исследования в области оптического распознавания символов продолжаются на протяжении более шестидесяти лет. Однако, все наиболее значимые результаты были получены в основном для задач распознавания латинских символов и цифр [2], в то время, как распознавание арабских текстов ставит множество вопросов, к которым невозможен обычный подход. Исследовательской деятельностью в области автоматической обработки арабских печатных и рукописных текстов занимаются во всем мире уже несколько десятилетий подряд [3, 4]. Данная задача является довольно трудной, поскольку возможно несколько представлений одного и того же символа в различных шрифтах и стилях. А при обработке арабских текстов возникает также множество неопределенностей, связанных с особенностями языка [5].

Для сложных систем обработки и анализа различного рода изображений характерны три основных стадии: предобработка, которая включает в себя нормализацию изображения, выделение шумов и артефактов, выделение признаков, включающее в себя преобразование изображения, или его регионов, в количественные и/или качественные, чаще всего векторные признаки и анализ, который в зависимости от задачи может состоять из кластеризации, классификации, трансформации полученных векторов признаков, их обработки.

Перед тем как использовать какой-либо метод интеллектуального анализа, текстовые данные необходимо предварительно обработать. Для этого осуществляют сегментацию текста, а затем по сегментированным данным получают численные характеристики (признаки) текста. Сегментацию текста особенно трудно произвести во время обработки рукописного материала. Строки зача-

стую искривлены, расстояние между строками неодинаково, размеры букв варьируются - все это составляет трудность сегментации рукописных текстов. Наличие диакритических знаков в арабском языке еще более усложняет задачу. Существующие методы извлечения строк текста можно разделить на следующие группы: основанные на общем проецировании, основанные на локальной группировке и размытии и основанные на преобразовании Хафа [6–8]. В работе [9] задача извлечения строк из текста моделируется как более общая задача кластеризации. Нейро-эвристический подход к сегментированию арабского текста был предложен в работе [10].

Качество выделения признаков может оказать решающее влияние на общее качество системы выделения информации из текстов [11]. Поэтому задача выделения признаков особенно важна для построения надежной системы.

Отличительной чертой рукописной арабской вязи и древних арабских текстов в частности являются ее геометрические особенности, разнящиеся ее от языков романской, славянской и других групп. Это оказывает значительно влияние на использование методов выделения признаков [12]. Для выделения геометрический особенностей используются различные методы выделения признаков. Зачастую используются методы выделения границ: морфологические преобразования, выделение максимумов градиента, методы первого порядка, дифференциальное выделение границ. Отдельно стоит упомянуть оператор Кэнни [13], который при своей простоте остается одним из наиболее качественных и используемых методов выделения границ [14]. Для линий арабской вязи характерными являются признаки кривизны и пересечения линий письма, которые находятся с помощью Лапласиана Гауссиана, Дифференциала Гауссиана, Детерминант Гессиана, определения кривизны кривой, детектор Харриса (более полное описание методов выделение структурных особенностей приведено в работе [15]). Признаки формы также информативны для арабской вязи, например: цепные коды, дескрипторы Фурье, гистограмма направлений. Отличительной особенностью арабской вязи является пропорции содержания петель, литер с нижним и верхним выносным элементом, верхних и нижних точек.

Для построение качественных систем обработки информации, в том числе и распознавании арабских текстов, популярным подходом является комбинирование различных признаков. Так в опи-

санной в работе [16] системе для определения авторства арабских рукописей значимыми показаны следующие характеристики: распределения углов и их направлений, инварианты моментов, общая площадь текста, его длинна и ширина, высота символов, расстояние от центральной линии от нижней и верхней точки каждого слова.

Упомянутая центральная линия является еще одной важной особенностью арабской вязи, и ее положение используется для построения различных признаков распределения символов относительно центральной линии. В работе [17] положение центральной линии и производные признаки использованы для распознавания рукописных текстов. Кроме того, положение центральной линии само по себе характерно для различных стилей письма и является важным признаком. Определение центральной линии в рукописном тексте является сложной задачей и включает в себя построение признаков, в том числе упомянутых выше. Для решения этой задачи в статье [18] использовался приближенный полигональный скелет и граф связных компонент, показавший высокую эффективность.

В последнее время все большую популярность получают методы полуавтоматического выделения признаков с использованием алгоритмов машинного обучения без учителя. Так, например, в статье Andrew Ng с соавторами [19] признаки, полученные на основе кластеризации изображений, успешно использовались для распознавания как печатного так и рукописного текста.

3. Проведенные исследования

Задача оптического распознавания текста включает в себя такие подзадачи, как определение областей, содержащие текст, нормализация текста, разделения строк, обработка каждой из строк по-отдельности. Каждая из этих подзадач может осложняться дополнительными факторами: наличием шума или повреждений изображения, качеством изображения, использованием различных шрифтов. Особняком стоит наличие рукописного текста, при котором задача усложняется на порядок.

Было проведено исследование подходов оптического распознавания в случае изображения, содержащего печатный текст, полученного при помощи фотографии или сканера. В его ходе выявились сложности, присущие как самой задаче распознавания текста с изображения (выделение и нормализация текста), так и арабскому

тексту, как это было описано выше (невозможность разбить текст на отдельные буквы). В результате, были предложены и исследованы различные методы решения возникших проблем.

3.1. Определение текста на изображении

При выделении текста на изображении возникает задача приведения области, содержащей текст в вертикальное положение, так как разбиение на строки в этом случае осуществляется значительно легче. Был предложен метод, который определяет наиболее вероятный угол поворота, основываясь на границе (правой, в случае арабского языка) области изображения, содержащей текст. Затем интересующая нас часть поворачивается на найденный угол и очищается от шумов и артефактов. Полученное нормализованное изображение необходимо разделить на строки текста. Для решения этой задачи был исследован подход, основанный на методе гистограмм. По вертикальной оси изображения строиться гистограмма интенсивности черного цвета, на основе которой определяются наиболее вероятные границы строк текста, по которым и происходит их разделение.

3.2. Выделение центральной линии

Как было описано выше, выделение центральной линии является одной из важнейших задач в оптическом распознавании текста на арабском языке, так как центральная линия представляет собой характерный признак арабской вязи. Для решения этой задачи в рассматриваемой ситуации, когда строка содержит печатный текст, был исследован метод заключающийся в комбинированном применении морфологических преобразований и анализа на основе гистограмм. Данный подход сравнительно прост для реализации, однако обладает свойством робастности, и с большой долей вероятности дает нужный результат даже при наличии шума. В начале, к строке текста применяется морфологическое преобразование, тогда результирующие изображение содержит компоненты связности с ярко выраженными центральными линиями, что объясняется свойствами арабской вязи. Применяя к компонентам связности гистограммный анализ, определяем линию, с наибольшей вероятностью являющуюся центральной линией.

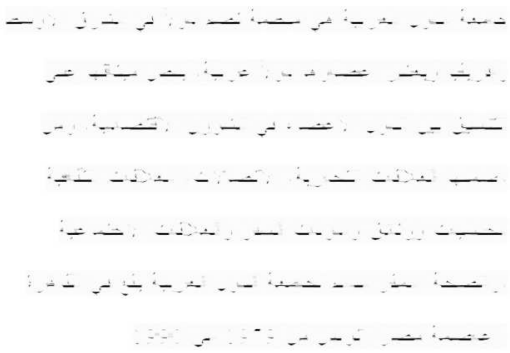

Рис. 1: Центральные линии компонент связности

3.3. Разбиение на слова и буквосочетания

Задача распознавания текста является частным случаем более общей задачи — классификации. Классификация — формализованная задача, в которой имеется множество объектов, разделчнных некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Классовая принадлежность остальных объектов неизвестна. Требуется построить алгоритм, способный классифицировать произвольный объект. В случае оптического распознавания текстов роль объектов играют изображения, а классов - буквы алфавита и символы. В рассматриваемой задаче количество возможных классов определяется количеством различных вариантов написания букв арабского алфавита. Нами был определена единица арабского текста, которая с одной стороны, может быть найдена на изображении простым образом, а с другой стороны, обеспечивала бы наибольшее качество и наиболее полную информацию для классификации арабских букв (и вариантов из написания) — *буквосочетаний.*

В арабском языке слова состоят из нескольких групп букв, внутри которых группы соединены вместе. Подобное объединение одной и более букв в единую неразрывную конструкцию называются

буквосочетанием. Известно, что буквы арабского алфавита могут иметь различное написание в зависимости от их положения: в начале, в середине, или в конце буквосочетания. Исходя из этого факта, было предложено вначале разделить строку на слова, которые в свою очередь разделить на буквосочетания, а затем проводить распознавание букв отдельно для каждого из полученных буквосочетаний. Таким образом, ставится две вспомогательные задачи: разделить изображение строки арабского текста на слова и разделить полученные слова на буквосочетания.

Стоит отметить, что буквосочетание — это минимальная связная текстовая единица (за исключением диакритических знаков), следовательно основной задачей ставится разделение изображения строки текста на части, содержащие в себе наименьшие компоненты связности (при этом не являющиеся диакритическими знаками и артефактами). Исходя из этого можно выделить два основных подхода для решения этой задачи:

- **выделение компонент связности** — поиск всех минимальных компонент связности из изображения (компонента связности — связный набор черных пикселей),
- **выделение непрерывных вертикальных промежутков** — поиск непрерывных кривых от верха до низа строки. Подобные кривые либо разделяют два буквосочетания, либо отделяют буквосочетания от левого/правого края строки.

Было реализовано оба этих метода и проведено их эвристическое сравнение.

Выделение компонент связности

Любое буквосочетание можно рассмотреть как объединение одного внешнего контура, очерчивающего буквосочетание, и нескольких (или ни одного) внутреннего контура, очерчивающих полости внутри буквосочетания. Под контуром понимается непрерывная граница между областями черных и белых пикселей. На практике было выявлено, что наличие артефактов на изображении и наклон текста, влияют на результаты данного метода, делая его непригодным для распознавания даже печатного арабского текста. Поэтому был разработан другой метод.

Выделение непрерывных вертикальных промежутков

Непрерывные вертикальные промежутки от низа до верха изображения могут представлять собой как пробелы — пустые прямоугольные области, так и искривленные промежутки, которые могут возникать между соседними буквосочетаниями внутри одного слова. Пробелы могут быть легко определены с помощью гистограммного метода. Точное определение искривленных промежутков — гораздо более сложная задача. Она была решена при помощи специального алгоритма поиска пути. Он заключается в поиске пустых областей на центральной линии и затем поиска непрерывных путей от центральной линии до нижней и верхней границы строки для каждого из найденных на центральной линии пустых областей.

Рис. 2: Пример разбиения строки на слова

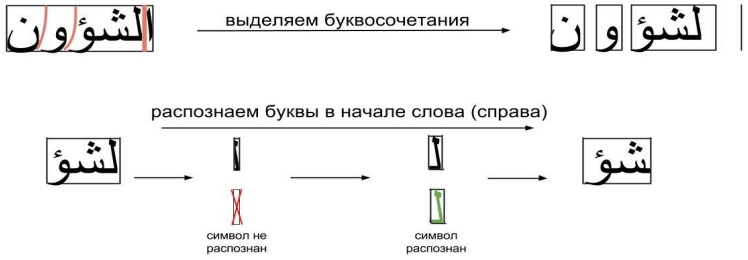

Рис. 3: Пример выделение из слова его буквосочетаний и их частичное распознавание

4. Машинное обучение в задаче OCR

Для полного распознавания текста был рассмотрен и реализован в прототипе простейший метод скользящего окна. Этот подход заключается в движении окна заданного размера справа налево вдоль строки, содержащей текст на арабском языке, и сравнении изображения в окне с известными заранее образцами символов арабского алфавита. Пусть $x^w = (x_1^w \dots x_N^w)$ — изображение в окне размера N, которое представляет собой массив символов. Аналогично $z^t = (z_1^t \dots z_N^t)$, $t = 1 \dots T$ — изображения образцов символов арабского алфавита размера N, где T — количество классов и соответствующих изображений-образцов. Тогда, при движении окна $w = 1 \dots L$

$$y = \min_t \frac{1}{N} \sum_{i=1}^{N} |x_i^w - z_i^t|$$

результатом классификации будет класс t^*

$$t^* = \arg \min_{t=1 \dots T} \frac{1}{N} \sum_{i=1}^{N} |x_i^w - z_i^t|,$$

если y меньше заданного порога. Таким образом классифицируется символ с ошибкой сравнения меньшей заданного порога. В идеальной ситуации этот метод показывает наилучший результат.

Рис. 4: Результат работы простейшего алгоритма для строки, рассмотрены только несколько наиболее частых символов

Однако, такой подход сильно зависит от размера и шрифта образцов и самого текста и при их сильном различии качество распознавания значительно падает. Также этот метод восприимчив к наличию шума в изображении, содержащем текст. Гораздо более надежным и универсальным представляется использования алгоритма машинного обучения, примененном к определенным ранее буквосочетаниям. В общем виде, необходимо уметь выделять из данных (изображения) признаки. Эти признаки передаются обученному на тренировочной выборке алгоритму, который классифицирует рассматриваемый объект.

Один из подходов выделения признаков для арабского языка описан в [5]. Такой метод позволяет успешно работать с несколькими относительно близкими шрифтами текста и размерами. Однако, в этом случае признаки выделяются искусственно, что представляет собой достаточно трудоемкую операцию. К тому же, при некоторых изменениях в данных, таких как сильно отличающийся шрифт, эти искусственные признаки могут оказать влияние на ухудшение результата. Для повышения универсализма и робастности операции распознавания текста был рассмотрен подход, основанный на автоматическом выделении признаков из данных, представленный в статье [19]. Была исследована возможность применения такого метода к текстам на арабском языке. Как было упомянуто во ведении, метод состоит в извлечении признаков из изображения при помощи обучения признаков без учителя, которое реализуется путем применения кластеризации K-средних (K-means) к набору маленьких изображений, полученных с обрабатываемого изображения. Такой подход позволяет повысить робастность алгоритма распознавания текста. Кластеризация, как метод обучения признаков без учителя, была выбрана как потому, что хорошо зарекомендовала себя в смежной задаче, описанной в статье [19], так и потому, что это направление активно развивается и уже на данный момент в нем имеются значительные наработки (смотри, например [20, 21])

Процесс обучения предложенного алгоритма, основанного на статье [19] состоит из трех этапов

1. При помощи алгоритма обучения признаков без учителя выделить основные признаки на основе фрагментов картинок фиксированного размера, полученных из тренировочных изображений.

2. Сверточное вычисление вектора признаков для изображений на основе признаков, полученных из предыдущего пункта.

3. Обучение алгоритма классификации (например SVM) на основе векторах признаков, полученных в предыдущем пункте.

Опишем формально первый этап. Пусть, мы разделили все картинки из тренировочного набора на фрагменты фиксированного размера (например, 8 на 8 пикселей), получив таким образом множество фрагментов $\widehat{X} = \{\hat{x}^{(i)}\}_{i=1}^{N}$, где $\hat{x}^{(i)} \in \mathbb{R}^{8\times 8}$. В качестве

алгоритма обучения признаков без учителя использован адаптированная версия алгоритма кластеризации K-means. Перед использованием алгоритма кластеризации, фрагменты $\hat{x}^{(i)}$ преобразуются в $x^{(i)} \in \mathbb{R}^{64}$ путем нормализации и линеаризации, формируя множество соответствующих векторов $X = \{x^{(i)}\}$. Далее, на полученном множестве векторов X применяется модифицированный алгоритм K-means, решающий следующую задачу оптимизации с ограничениями:

$$||Cs^{(i)} - x^{(i)}||_2 \xrightarrow{C, s^{(1)}, \ldots, s^{(N)}} \min$$
$$C = [C^{(1)} : \ldots : C^{(K)}] \in \mathbb{R}^{64 \times K}, \quad s^{(i)} \in \mathbb{R}^{64}$$
$$||C^{(j)}||_2 = 1 \quad \forall j = 1 \ldots K,$$
$$\text{только один из элементов } s^{(i)} \text{ не нулевой} \quad \forall i = 1 \ldots N.$$

Здесь столбцы матрицы C играют роль центроидов, а вектора $s^{(i)}$ играют роль маркеров, указывающих к какому кластеру принадлежит вектор $x^{(i)}$. Описанная минимизация достигается путем итераций поочередной оптимизации по матрице C и по векторам-маркерам $\{s^{(i)}\}_{i=1}^{N}$. Формально алгоритм минимизации можно представить в следующем виде

1. Шаг: инициализация алгоритма
   - фиксируется число кластеров K,
   - случайно инициализируется матрица $C \in \mathbb{R}^{64 \times K}$ таким образом, что $C^{(j)} = 1$, $\forall j = 1 \ldots K$.
2. Шаг: минимизация по $\{s^{(i)}\}_{i=1}^{N}$
$$s_k^{(i)} = \begin{cases} C^{(k)T} x^{(i)} & k = \arg\max_j C^{(j)T} x^{(i)} \\ 0 & else \end{cases}$$
3. Шаг: минимизация по C
$$D^{(j)} = \frac{\sum_{i=1}^{N} \langle s^{(i)}, e^{(i)} \rangle x^{(i)}}{\sum_{i=1}^{N} \langle s^{(i)}, e_i \rangle}$$
$$D^{(j)} = \frac{D^{(j)}}{||D^{(j)}||_2},$$
   гду $e^{(i)}$ — вектор с единицей на i-ой позиции и нулем на остальных.

4. Шаг: условие остановки алгоритма.

   Если совершено максимальное количество итераций, или изменение матрицы D мало, то алгоритм завершает работу, иначе он возобновляется с шага 2.

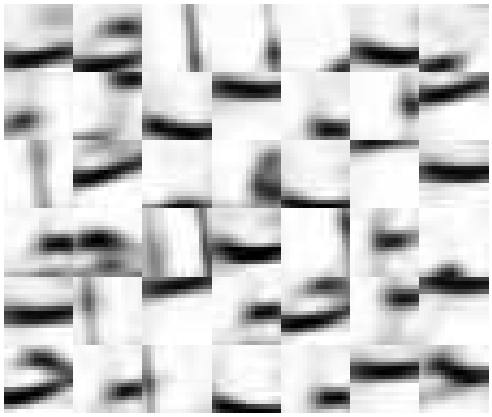

Рис. 5: Признаки, полученные автоматически путем кластеризации обучающего набора данных

Стоит отметить, что в признаках превалирует горизонтальная составляющая, что подчеркивает важную роль центральной линии в распознавании арабского текста.

Итоговый метод, полученный в ходе исследования поставленной проблемы можно записать следующим образом:

1. Выделение на входящем изображении областей, содержащих текст.
2. Нормализация изображения (поворот, очистка от артефактов).
3. Разделение текста на строки.
4. Выделение центральных линий.
5. Определение буквосочетаний.
6. Для каждого буквосочетания автоматически выделяются признаки и определяются символы алфавита.

Рис. 6: Результат работы алгоритма с автоматическим выделением признаков для строки. Рассмотрены наиболее частотные буквосочетания. Результат хуже, чем на 4

5. Заключение и перспективы дальнейших исследований

Представляется перспективным дальнейшее исследование в области выделения текста на изображении. Могут быть рассмотрены более сложные случаи, содержащие большее количество шума и артефактов. Также в дальнейшем необходимо продолжить исследования алгоритма распознавания текста на арабском языке, основанного на автоматическом выделении признаков. Естественно рассмотреть различные возможности улучшения и модификации метода, применение его к текстам с различными шрифтами и артефактами.

Интересна возможность повысить общую робастность системы за счет использования более устойчивых вариаций метода K-средних, так сильная зависимость результатов оригинального метода от начальных методов может выразится как в неоптимальном выборе кластеров, так и в затрудненной воспроизводимости результатов. В этом контексте интересно опробовать применение результатов из статьи [20], где приведена рекуррентная модификация метода K-средних и доказана ее состоятельность в условии почти произвольных помех. Является перспективным рассмотрение этого рекуррентного алгоритма в случае смеси гауссовских распределений, т.к. такое поведение данных является весьма правдоподобным в случае задачи распознавания букв. Интерес также представляет и решение задачи устойчивого выбора числа кластеров, которая является традиционной в кластеризации. Существует множество различных подходов к решению этой проблемы, как например: геометрический подход [23, 24], подход на основе индекса корреляции внешнего разбиения [25], методы основанные на повторных выборках [26, 27]) и др. Интересное развитие индексных методов с помощью аппроксимации полиномами Чебышева получено в статье [28]. Разнообразие и новизна перечисленных методов и подходов дает основания надеяться на возможность значительного повышения качества и ста-

бильности работы системы распознавания.

Оптическое распознавание текстов на арабском языке является актуальной проблемой. Методы, разработанные для латинских и славянских языков, испытывают трудности при применении их к арабской вязи. В процессе исследования были выявлены эти проблемы и предложены пути их решения, такие как определение центральной линии и буквосочетаний, использования автоматического выделения признаков для алгоритма машинного обучения. Таким образом, можно заключить, что в этой востребованной области были рассмотрены перспективные пути развития и подходы, открывающие широкое поле для дальнейших исследований.

Список литературы

[1] *Redkin O.I., Bernikova O.A.* Problems of the Arabic OCR: new attitudes // Proceedings of the 2013 International Conference on Artificial Intelligence. Las Vegas, 2013. P. 777–782.

[2] *LeCun Y., Botton L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition // Proc. IEEE 1998.

[3] *El-Sheikh T. S., Guindi R. M.* Computer recognition of arabic cursive script // Pattern Recognition. № 21(4). 1988. P. 293Ц-302

[4] *Hussain F., Cowell J.* Character recognition of arabic and latin scripts // Proc. IEEE International Conference on Information Visualisation. 2000. P. 51Ц-56.

[5] *Шалымов Д. С.* Автоматическое распознавание печатных текстов арабского языка // Стохастическая оптимизация в информатике. 2007. № 3. С. 124–137

[6] *Razak Z., Zulkiflee K.* Off-line handwritting textline segmentation: a review // International Journal of Computer Science and Network security. 2006. Vol. 8. № 7.

[7] *Papavassiliou V., Stafylakis T., Katsouros V., Carayannis G.* Handwritten document image segmenation into textlines and words // Pattern Recognition. 2010. Vol. 43. No 1.

[8] *Li Y., Zheng Y., Doermann D.* Detecting text line in handwritten documents // ICPR'06, 2006, P. 1030–1033

[9] *Kumar, Jayant, et al.* Handwritten arabic text line segmentation using affinity propagation // Proc. of the 9th IAPR International Workshop on Document Analysis Systems. ACM. 2010

[10] *Hamid A., Haraty R.* A neuro-heuristic approach for segmenting handwritten Arabic text // ACS/IEEE International Conference on Computer Systems and Applications. 2001.

[11] *Yang Y., Pedersen J. O.* A comparative study on feature selection in text categorization // Proc. Fourteenth International Conference on Machine Learning. 1997. P. 412–420.

[12] *Khorsheed M. S.* Off-line arabic character recognition. A review // Pattern Analysis and Applications, 2002. Vol. 5. P. 31–45.

[13] *Canny J.* A computational approach to edge detection // IEEE Trans. Pattern Analysis and Machine Intelligence, 1986. Vol. 8. P. 679–714.

[14] *Shapiro L. G., Stockman G. C.* Computer Vision. — London etc.: Prentice Hall, 2001. P. 326.

[15] *Lindeberg T* Feature detection with automatic scale selection // International Journal of Computer Vision. 1998. № 30(2), P. 77Ц-116.

[16] *Al-Maadeed S.* Text-dependent writer identification for arabic handwriting // Journal of Electrical and Computer Engineering. 2012.

[17] *El-Hajj R., Likforman-Sulem L, Mokbel C.* Arabic handwriting recognition using baseline dependant features and hidden Markov modeling // 8th International Conference on Document Analysis and Recognition. ICDAR 2005. Seoul, Korea, 2005.

[18] *Pechwitz M., Margne V.* Baseline estimation for arabic handwritten words // Proc. Eighth International Workshop Frontiers in Handwriting Recognition. 2002. P. 479–484.

[19] *Coates A., Carpenter B., Case C., Satheesh S., Suresh B., Wang T., David J. Wu, Andrew Y. Ng* Text detection and character recognition in scene images with unsupervised feature learning // ICDAR 2011. P. 440–445

[20] *Morozkov M., Granichin O., Volkovich Z., Zhang X.* Fast algorithm for finding true number of clusters. Applications to control systems // In: Proc. of the 24th Chinese Control and Decision Conference (CCDC). 2012. P. 2013–2018.

[21] *Граничин О.Н., Шалымов Д.С., Аврос Р., Волкович З.* Рандомизированный алгоритм нахождения количества кластеров // Автоматика и телемеханика. 2011. №4. С. 86–98.

[22] *Граничин О.Н., Измакова О.А.* Рандомизированный алгоритм стохастической аппроксимации в задаче самообучения // Автоматика и телемеханика. 2005. №8. С. 52–63.

[23] *Dunn J.C.* Well separated clusters and optimal fuzzy partitions // Journal Cybernetics. 1974. Vol. 4. P. 95–104.

[24] *Tibshirani R., Walther G., Hastie T.* Estimating the number of clusters via the gap statistic // Journals of the Royal Statistical Society: Series B, 2001. Vol. 63. № 2. P. 411–423.

[25] *Dudoit S., Frid lyand J.* A prediction-based resampling method for estimating the number of clusters in a dataset // Genome Biol. 2002. No. 3. P. 112–129.

[26] *Sugar C., James G.* Finding the number of clusters in a data set: An information theoretic approach // J. America Statistical Assoc. 2003. No. 98. P. 750–763.

[27] *Levine E., Domany E.* Resampling method for unsupervised estimation of cluster validity // Neural Computation. 2001. № 13. P. 2573–2593.

[28] *Avros R., Granichin O., Shalymov D., Volkovich Z., Weber G.-W.* Randomized algorithm of finding the true number of clusters based on Chebychev polynomial approximation (Chapter 6) // Data Mining: Found. & Intell. Paradigms, D.E. Holmes, L.C. Jain (Eds.), Berlin Heidelberg: Springer-Verlag, ISRL 23, 2012. Vol. 1, P. 131–155.