**УДК** 004.6

СМИРНОВ С.В.

# ПОДСИСТЕМА МАССОВОГО РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЙ АРХИВНЫХ ДОКУМЕНТОВ

*Смирнов С.В.* **Подсистема массового распознавания изображений архивных документов.**

**Аннотация.** В настоящей статье вначале описываются особенности и проблематика массового распознавания архивных документов. Рассматриваются ключевые проблемы проектирования такого рода систем, приводятся примеры и предлагаются различные варианты их решения. Далее приводится концептуальная схема построения электронного архива и отдельная схема организации входящей в его состав подсистемы автоматического распознавания. Описываются основные компоненты, функции и бизнес-процессы, протекающие в системе.

**Ключевые слова:** автоматическое распознавание изображений, электронный архив, оцифровка архивных документов.

*Smirnov S.V.* **Subsystem of mass image recognition of archival documents.**

**Abstract.** The description of main features and problems of mass recognition of archival documents is presented in the beginning of the paper. Key design problems are considered, the examples are shown and different kinds of solutions are advised. After that, the scheme of electronic archive and the scheme of recognition subsystem are shown with the description of their components, functions and business processes.

**Keywords:** optical character recognition, automatic recognition of images, electronic archive, digitization of archival documents.

**1. Введение.** Одной из главных задач любого архива, музея, библиотеки является обеспечение сохранности и предоставление инструментов для оперативного поиска и обработки хранящихся документов. В случае, когда основная доля всех документов представлена на бумажных носителях, данные задачи усложняются трудностями поддержки специальных условий хранения, задачами непрерывного процесса контроля качества хранящихся документов, а также необходимостью ведения объемного справочно-поискового аппарата и сложностями реализации процедур поиска запрашиваемых данных. Проблемы сохранности успешно решаются путем создания электронных образов документов с помощью современных сканирующих устройств. После снятия электронной копии документа и размещения ее в хранилище электронного архива (ЭА) для изображений текстовых документов на передний план выходят задачи перевода отсканированного текста в его электронную форму, построение поисковых индексов и формирование электронных версий оригинала документа.

Очевидно, что задачи перевода многомиллионных фондов архивных документов от электронных образов к текстовому виду требуют

огромных временных затрат. С целью уменьшения количества документов, требующих обработки, можно фильтровать их по показателям востребованности, частоте обращений, признакам ценности или состоянию оригинала. Далее каждой группе документов следует назначать приоритеты, в порядке которых они будут поступать на площадку сканирования и ввода и т.д. Но оцифровка даже в таком порядке вряд ли обеспечит должных темпов. Поэтому для снижения временного порога попадания документа в поисковый индекс можно на первоначальном этапе не переводить его в полноценный электронный вид, а снабжать стартовыми поисковыми реквизитами. Таким образом, можно будет приступать к использованию документов сразу после этапа сканирования, а в дальнейшем прикреплять к ним полнотекстовое представление на этапах ввода и распознавания.

Из описанного процесса следует, что скорость сканирования документов, в общем случае, намного превышает скорость их преобразования в полнотекстовый вид, дополненный средствами разметки, обеспечивающей дополнительные сервисные функции (возможность поиска и др.), и как следствие возникает острая потребность в проектировании и разработке средств автоматизации в виде систем массового распознавания. Эксплуатация таких систем должна обеспечивать следующие преимущества:

- увеличение скорости распознавания документов;
- увеличение объема обрабатываемых документов;
- возможности быстрого поиска документов;
- экономию времени, затрачиваемого на поиск информации по изображениям, путем применения поискового инструментария вкупе с распознанным представлением документов;
- альтернативу ручному процессу распознавания изображений документов;
- уменьшение затрат на содержание штата операторов ручного ввода.

**2. Особенности и проблематика массового распознавания архивных документов.** Отличительными особенностями массового распознавания архивных документов являются:

- сверхбольшие объемы обрабатываемых документов;
- разбиение всего объема документов на большие тематические группы, обладающие общими свойствами;
- высокие требования к пропускной способности системы;

- отсутствие практической возможности проведения детальной верификации и корректировки всех результатов распознавания;
- важность проведения автоматической оценки и контроля качества результатов распознавания.

Также процессы проектирования и внедрения системы массового распознавания отличают следующие проблемные области:

- определение целей и предназначения результатов распознавания;
- проведение анализа качества исходных документов и их характеристик;
- вычисление показателя точности распознавания и установка допустимого порога;
- разработка инструментов автоматического контроля качества;
- проектирование с учетом масштаба и продолжительности проекта [1].

**2.1 Цель и предназначение результатов распознавания.** Существует много вариантов использования результатов распознавания, и они далеко не ограничиваются созданием лишь полностью идентичной копии оригинала документа. Во многих случаях качество результатов автоматического распознавания проигрывает качеству ручного ввода текста, и, как следствие, для исторических архивных документов данный способ распознавания является малопригодным. Однако, в зависимости от целей разрабатываемой системы, требования к достоверности результатов могут существенно варьироваться, и во многих случаях применение средств автоматизации оправдывается даже при невысоком качестве распознавания.

Потенциальные виды использования результатов распознавания включают [2]:

- Индексирование — результат распознавания рассматривается как простой текст и в дальнейшем подается на вход поисковой системы. Текст используется как основа для полнотекстового поиска. Причем, конечному пользователю не отображаются фактические результаты оптического распознавания, а выдается найденный образ документа без обозначения вхождения поисковой фразы.

  Данный вид не требователен к точности распознавания и одновременно предоставляет хорошие поисковые возможности.
- Отображение с подсветкой результатов — в данном режиме распознанный текст обрабатывается также как и в предыду-

щем случае, отличие же заключается в подсистеме отображения поисковых результатов. В результатах поиска пользователю предоставляется изображение с выделенными фрагментами вхождений поисковой фразы. Очевидно, что в данном случае неточности распознавания будут обнаружены с большей вероятностью, тем не менее, данный подход намного более дружелюбен к пользователю, чем его предшественник.

- Выдача результатов в виде текста — вместо оригинального документа пользователю отображается текст, полученный в результате распознавания. Главным показателем в данном режиме является качество распознавания. Если слова будут сильно искажены, то пользователь не сможет получить искомой информации и потеряет доверие к системе. Таким образом, точность должна быть очень высокой, что практически не может быть достигнуто без привлечения человеческого ресурса, и, как следствие, ведет к значительным временным и финансовым затратам.
- Текстовое представление с разметкой — отображение результатов распознавания редко производится без форматирования и разметки текста с целью сохранения исходной структуры и деталей расположения элементов. В дополнение размеченный xml документ может содержать дополнительные атрибуты, тэги или ссылки на родственные документы.

Таким образом, тщательный анализ исходных данных наравне с выработкой стратегии представления результатов позволяет однозначно определить требования к критериям достоверности распознавания и необходимость разработки и применения ручных этапов верификации и корректировки с целью получения удовлетворительных результатов.

**2.2 Качество и характеристики исходных документов.** Проведение тщательного анализа исходного набора документов позволит определить степень пригодности к автоматическому распознаванию. Например, рукописный текст не сохраняет однородности начертания символов и малопригоден для оптического распознавания символов (OCR — optical character recognition), а сильно поврежденная бумага может привести к нулевой производительности автоматических методов распознавания.

Прошедшие первичный отбор документы должны быть исследованы на предмет выявления характеристик, влияющих на процесс распознавания, таких как используемый язык, шрифт, макет, наличие графических изображений, таблиц и т.п. Выявление данных характе-

ристик позволит выстроить рациональный процесс обработки документов, выбрать подходящие методы, алгоритмы и модули распознавания. Более того, при наличии устойчивых наборов характеристик исходные документы могут быть разбиты на типы. Каждый тип, обладая определенным составом признаков и коэффициентом пригодности к распознаванию, может обрабатываться различными методами и движками или априори считаться неподходящим для распознавания. Примерами таких типов могут быть полноразмерные графические документы, документы с частыми вхождениями спецсимволов, научных или математических формул, документы с нестандартным лексическим содержанием, написанные на исторических языках или содержащие большой объем узкотематической терминологии.

**2.3 Показатель точности распознавания системы.** Большинство производителей систем оптического распознавания заявляют о стабильном уровне точности распознавания при определенных параметрах конфигурации. Оценка точности, в большинстве случаев, рассчитывается как процентное соотношение корректно распознанных символов к общему количеству символов. Более того, оценка зачастую производится на «идеальных» документах и с большей долей вероятности будет отличаться от реальной оценки при обработке исторических документов.

Следовательно, перед этапом запуска системы массового распознавания архивных документов следует установить допустимый порог точности и провести визуальную проверку полученных результатов на тестовой выборке документов с целью проверки, насколько разработанная система удовлетворяет заданным требованиям к качеству. Это позволит выявить потенциальные недостатки, ошибки и исключит возможные затраты на перезапуск системы в случае провала испытаний. В долгосрочных проектах уместно будет включить этап определения показателя точности в бизнес-процесс работы системы и производить его регулярно.

При вычислении показателя точности следует обратить внимание на разницу значимости показателя, вычисленного в отношении символов и в отношении слов. С точки зрения поиска и затрачиваемых усилий при корректировке, показатель, вычисленный относительно слов, представляет большую полезность и значимость.

**2.4 Автоматический контроль качества распознавания.** В связи с тем, что главной чертой систем массового распознавания является сверхбольшой объем документов и отсутствие возможности произвести проверку каждого документа вручную, важнейшим процессом яв-

ляется автоматическое определение показателя качества проведенного распознавания. Наличие такой оценки позволит установить определенную шкалу градации, по которой будет определяться дальнейшая переадресация документа. Например, при обработке документов научно справочного аппарата архивов может применяться следующая шкала:

- от 99% — работа с документом может считаться законченной, документ считается пригодным для индексации и отображения пользователю;
- 95%–99% — документ переводится на этап верификации и ручной корректировки, до тех пор, пока он не пройдет корректировку, документ считается недостоверным, но, тем не менее, присутствует в поисковом индексе, и пользователи работают с ним;
- 85%–95% — документ переводится на этап верификации и ручной корректировки, не попадет в поисковый индекс, и пользователи не получают доступ к нему;
- до 85% — документ отбраковывается и считается непригодным к распознаванию автоматическими средствами.

**2.5 Проектирование с учетом масштаба и продолжительности проекта.** При планировании разработки системы массового распознавания важно осознавать, что решения, хорошо зарекомендовавшие себя в краткосрочных проектах объемом в десятки тысяч страниц, могут быть неприменимы к долгосрочным проектам размером в десятки миллионов страниц. Даже у краткосрочных проектов может возникнуть ситуация, когда подходы, обеспечивающие стабильные результаты в начальной фазе, могут давать серьезные сбои по мере разрастания рамок проекта.

В дополнение к вышесказанному, в долгосрочных проектах следует учитывать скорость развития технического прогресса, потому как изначально выбранные программно-аппаратные решения со временем могут устаревать и требовать обновления. Даже в случае работоспособности выбранных изначально средств внедрение новых или замены существующих компонентов могут принести значительный выигрыш как в производительности, так и в стоимости всего проекта.

Таким образом, важно еще на этапе проектирования выстроить хорошо расширяемую систему с возможностями адаптации и принять во внимание потенциальные риски при составлении планов и расчетах бюджета проекта.

**3. Место подсистемы распознавания в электронном архиве.** Учитывая вышеописанную проблематику, рассмотрим концепцию построения электронного архива с входящим в его состав модулем массового распознавания документов. На рисунке 1 представлена схема организации и взаимодействия основных компонентов электронного архива, вовлеченных в процесс массового распознавания.

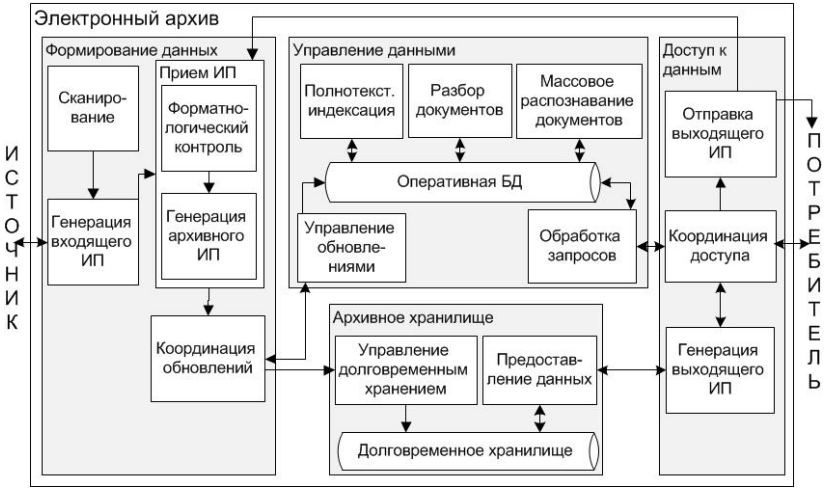


Рис. 1. Схема организации компонентов электронного архива.

На вход генератора входящего информационного пакета (ИП, submission information package [3]) поступают электронные образы архивных документов с этапа сканирования и от других источников информации. В процессе формирования входящего информационного пакета поступающие образы снабжаются уменьшенными копиями для предварительного просмотра и объединяются в группы. Каждая группа изображений помимо содержательной информации и данных о ее физическом размещении содержит в себе описательную и справочно-поисковую информацию.

Далее сформированный входящий ИП подвергается форматно-логическому контролю и классификации: проверяются форматы представления информации, качество передаваемых изображений (разрешение, размеры, цветность и т.п.), достаточность описательной информации, принадлежность к тем или иным категориям и типам информации, хранимой в архиве. После прохождения проверок и клас-

сификаций входящий ИП преобразуется в набор архивных ИП, которые уже непосредственно предназначаются для размещения в оперативных базах данных электронного архива.

Таким образом, поступившие на хранение в ЭА данные сразу могут быть доступны для поиска и просмотра потребителю, но в большинстве случаев первоначальной справочно-поисковой информации оказывается недостаточно для организации высококачественного и оперативного обнаружения информации. Как следствие, полученный набор документов обрабатывается модулями разбора, ввода, распознавания и индексации для расширения поисковой базы, упорядочивания и увеличения вероятности быть найденным среди всей россыпи поступившей информации.

В задачи модуля распознавания не входит построение идентичной электронной копии документа, которая могла бы использоваться без прикрепленных к ней образов. В качестве поискового инструмента достаточной является привязка текстового представления к области изображения по заданным координатам. Данное условие существенно снижает требования к сложности реализации модуля оптического распознавания, необходимости проведения затратных по времени ручных верификаций и корректировок, а также позволяет избежать осуществления процедур детального анализа макета изображения.

Работа с распознанным текстом, неотделенным от изображений, в свою очередь снижает порог поступления документа в рабочую базу системы за счет нестрогих требований к качеству распознавания (никакая информация не будет утеряна, так как всегда будет доступна для изучения непосредственно с образа). Таким образом, можно приступать к работе с документом по достижению установленного порога качества распознавания, а, впоследствии, после ручной верификации и корректировки, лишь обновлять базу данных. Применение данной стратегии увеличивает скорость пополнения поисковой базы архива, что, в свою очередь, дает прирост производительности и скорости поиска документов.

После этапов разбора, распознавания и тщательной проверки администраторами ЭА электронные версии документов могут группироваться в ИП и отправляться вновь на вход компонента приема данных, но уже не в качестве поисковых материалов, а как проверенные и готовые для размещения в долговременном хранилище документы.

**4. Архитектура подсистемы распознавания.** Представленная к рассмотрению подсистема массового распознавания состоит из четырех функциональных областей, связанных между собой единой базой

данных. Компоненты данных областей предназначены для выполнения следующих задач:

- импорт и экспорт данных для распознавания;
- подготовка системы к обработке поступивших данных;
- выполнение процессов распознавания;
- обработка результатов распознавания.

Архитектура описываемой подсистемы представлена на рисунке 2.

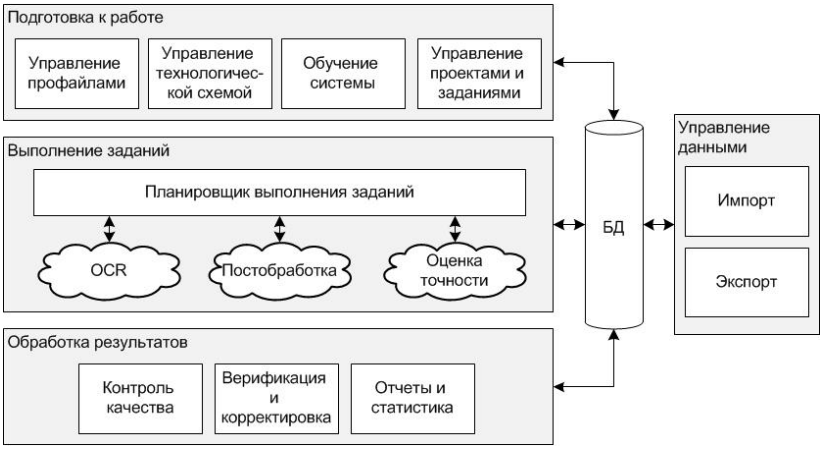


Рис. 2. Архитектура подсистемы массового распознавания.

**4.1 Управление данными и подготовка системы.** В обязанности компонента импорта данных входит выборка назначенных к распознаванию изображений документов из оперативной базы данных ЭА и перенаправление их в блок подготовки к обработке. После прохождения всех этапов распознавания полученные текстовые представления документов экспортируются из подсистемы распознавания обратно в оперативную базу данных ЭА для последующей полнотекстовой индексации и отправки пользователям архива.

На этапе подготовки электронные образы объединяются в группы по различным признакам, зависящим от специфики электронного архива, такими признаками могут быть: тематика документы, языковая принадлежность, шрифты, разметка, структура документа.

Каждой однородной группе документов назначается технологическая схема, детально описывающая последовательность предстоящей обработки. Данная технологическая схема представляет собой пошаго-

вую инструкцию, где для каждого шага содержатся соответствующие настройки в виде профайлов. Упомянутые профайлы создаются во время тренировочных процессов, проводимых перед началом обработки новых групп изображений. Результатами таких тренировочных прогонов являются файлы настроек различных компонентов распознавания от предварительной обработки до коррекции результатов оптического распознавания и методов оценки точности.

Принимая во внимание уникальность и особенности архивных документов, легко предположить, что базовый набор алгоритмов, методов, словарей и компонентов оптического распознавания с малой долей вероятности подойдет для обработки специфических изображений с древнерусским лексиконом или рукописными шрифтами. Отсюда вытекает необходимость в разработке модуля обучения системы распознаванию документов с неординарными свойствами.

**4.2 Выполнение процессов распознавания.** Центральным компонентом данного функционального блока является планировщик выполнения заданий. Он принимает задания, организует их постановку в очередь, инициирует запуск и производит диспетчеризацию среди всего набора модулей обработки. Планировщик обеспечивает выполнение задания строго по указанной в задании технологической схеме.

Особенностью построения данного блока является сервис ориентированная организация основных вычислительных модулей. Такая «облачная» архитектура обеспечивает возможности расширения, масштабируемости, подключения новых модулей и простоту управления вычислительными ресурсами системы, что является немаловажным фактором для дальнейшего развития системы и адаптируемости к различным средам эксплуатации и развертывания.

Процесс распознавания разделяется на три уровня: оптическое распознавание символов, постобработка результатов оптического распознавания и оценка точности проведенного распознавания. Каждый из таких уровней может обладать произвольным количеством реализаций и делится на модули, имплементирующие базовые интерфейсы, объединенные через общую шину электронного взаимодействия.

Рассмотрим более детально каждый из уровней обработки.

Модуль оптического распознавания символов отвечает за перевод изображений текста в машиночитаемую и редактируемую форму. В общем случае, принципы работы OCR основываются на структурном анализе изображений и разбиении этих изображений на более мелкие участки с целью обнаружения текстовых зон. Среди этих зон OCR идентифицирует отдельные строки текста, а среди строк выделяет от-

дельные слова и символы. Далее происходит поиск совпадения найденного символа с символами из предустановленного набора шрифтов и так далее для всех символов слова. После этого полученное слово прогоняется через словари с целью обнаружения совпадающих кандидатов [4]. Данным образом обрабатывается весь текстовый отрезок изображения, и в результате формируется результат распознавания в виде файла с описанием координат текстовых регионов, распознанным вариантом и дополнительными реквизитами, такими как степень уверенности отдельных слов и символов. Следует также отметить, что прежде чем поступить на вход компоненты OCR, изображение проходит ряд предварительных обработок: бинаризацию, сглаживание фоновых структур, устранение шумовых дефектов, выравнивание угла наклона, устранение искажений и сегментацию структуры документа.

На следующем уровне обработки результат оптического распознавания подвергается автоматической корректировке. Причем немаловажную роль на данном этапе играет информация о контексте распознавания. Известные подходы контекстной постобработки [5–8] включают статистические и лингвистические методы, использующие Скрытые Марковские Модели (СММ) [9], конечные автоматы, нейронные сети, N-граммы символов и слов, алгоритмы нечеткого отображения строк [5]. Также используются методы, использующие специальную внешнюю информацию, комбинированные методы и подходы, основанные на эвристиках [10].

Последним этапом обработки документов является проведение оценки точности распознавания. Вычисление данной оценки основывается на вероятностях достоверности распознавания, поступающих от модуля OCR и вероятностях подбора верных кандидатов для замены на этапе посткоррекции.

**4.3 Обработка результатов распознавания.** В область обработки результатов распознавания входят компоненты ручной верификации и корректировки, контроля качества и компоненты построения статистических отчетов.

Верификация и корректировка результатов распознавания может понадобиться в случаях, когда требуется стопроцентная точность распознавания или достоверное построение макета документа. Очевидно, что достичь таких результатов без применения ручных методов обработки не представляется возможным.

Контроль качества распознавания позволяет администраторам системы оценить степень пригодности различных модулей, настроек и технологических схем для обработки тех или иных типов документов

и своевременно реагировать в случае неудовлетворительных результатов.

Ни одна система не может обойтись без сбора и анализа статистических данных. Результаты данного анализа могут послужить сигналами, например, к оптимизации программных компонент, модернизации аппаратных средств или проведению профилактических работ.

**5. Заключение.** Электронный архив с включенной в его состав подсистемой массового автоматического распознавания документов обладает рядом существенных преимуществ над архивными системами, в которых распознавание либо отсутствует, либо осуществляется вручную. Данными преимуществами являются: высокие темпы перевода документов в электронную форму, возможности автоматического построения мощного и качественного поискового аппарата, быстрота поиска и доступа к электронным образам документов.

Перед построением такого рода систем необходимо найти решения и ответы на целый ряд задач. Предложенная в статье архитектура и организация подсистемы массового распознавания в комплексе электронного архива носит концептуальный характер и описывает возможные подходы к преодолению основных проблем, с которыми можно столкнуться, и о решении которых необходимо задуматься на этапах проектирования и планирования разработки.

В заключение необходимо отметить, что дальнейшая реализация предложенной подсистемы потребует проектирования и разработки форматов хранения, представления и обмена полученными электронными версиями документов, выработки методик оценки точности и контроля качества, детального анализа существующих и разработки новых методов оптического распознавания и алгоритмов коррекции результатов распознавания.

## Литература


1. *Anderson N.* IMPACT Best Practice Guide: Optical Character Recognition – Part 1. 2010. URL: http://www.impact-project.eu/uploads/media/IMPACT-ocr-bpg-pilot-s1.pdf (дата обращения: 06.06.2012)
2. *Tanner S.* Deciding Whether Optical Character Recognition is Feasible. 2004. URL: http://www.odl.ox.ac.uk/papers/OCRFeasibility_final.pdf (дата обращения: 06.06.2012)
3. ISO 14721:2003. Space data and information transfer systems -- Open archival information system -- Reference model.
4. *Anderson N.* IMPACT Briefing Paper: Optical Character Recognition. 2010. URL: http://www.impact-project.eu/uploads/media/IMPACT-ocr-bp-pilot-1b.pdf (дата обращения: 06.06.2012)
5. *Kukich K.* Techniques for automatically Correcting Words in Text // ACM computing survey Computational Linguistic. 1992. В. 24. №4. С. 377–439.

6. *Mailburg M.* Comparative Evaluation of Techniques for Word Recognition Improvement by Incorporation of Syntactic Information // 4th International Conference Document Analysis and Recognition (ICDAR ‘97). Август 1997. С. 784.
7. *Beitzel S., Jensen E., Grossman D.* A Survey of Retrieval Strategies for OCR Text Collections // Proc. of 2003 Symposium on Document Image Understanding Technology. Апрель 2003.
8. *Sholomov D.L.* Interpreting the Indistinctly Recognized Textual Constructions // Pattern Recognition and Image Analysis. 2003. В. 13. №2. С. 353–355.
9. *Bouchaffra D., Govindaraju V., Srihari S.* Postprocessing of Recognized Strings Using Nonstationary Markovian Models // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997. В. 21. №10. С. 990–999.
10. *Шоломов Д.Л., Постников В.В., Марченко А.А., Усков А.В.* Пост-обработка результатов OCR распознавания, использующая частично определенный синтаксис // Труды ИСА РАН. 2005. Т. 16. С. 146–163

**Смирнов Сергей Владимирович** — соискатель ученой степени канд. техн. наук; СПИИРАН. Область научных интересов: автоматическая обработка изображений документов и текстов. Число научных публикаций — 4. serge.smir@gmail.com; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7 911 2430840. Научный руководитель — С.В. Кулешов.

**Smirnov Sergey Vladimirovich** — competitor of PhD in Technics; SPIIRAS. Research interests: automatic processing of document images and texts. The number of publications — 4. serge.smir@gmail.com; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7 911 2430840. The scientific adviser — S.V. Kuleshov.

Рекомендовано лабораторией автоматизации научных исследований СПИИРАН, зав. лаб., д.т.н., профессор Александров В.В.

Статья поступила в редакцию 07.05.2012.

## РЕФЕРАТ

*Смирнов С.В.* **Подсистема массового распознавания изображений архивных документов.**

В статье описывается актуальная проблема массового распознавания изображений архивных документов. Скорость сканирования документов, в общем случае, намного превышает скорость их преобразования в полнотекстовый вид, дополненный средствами разметки, обеспечивающей дополнительные сервисные функции (возможность поиска и др.), и как следствие возникает острая потребность в проектировании и разработке средств автоматизации в виде систем массового распознавания.

Но процессы разработки систем массового распознавания архивных документов сопряжены с рядом проблем и требуют тщательной проработки в следующих проблемных областях:

- определение целей и предназначения результатов распознавания;
- проведение анализа качества исходных документов и их характеристик;
- вычисление показателя точности распознавания и установка допустимого порога;
- разработка инструментов автоматического контроля качества;
- проектирование с учетом масштаба и продолжительности проекта.

Как отправная точка для решения вышеперечисленных проблем в статье рассматривается концептуальная схема построения электронного архива с включенной в его состав подсистемой массового распознавания.

Рассматриваемая подсистема массового распознавания состоит из четырех функциональных областей, отвечающих за импорт и экспорт данных для распознавания, подготовку системы к обработке поступивших данных, выполнение процессов распознавания и обработку результатов распознавания.

Центральным компонентом системы является планировщик выполнения заданий, который принимает задания, организует их постановку в очередь, инициирует запуск и производит диспетчеризацию среди всего набора модулей обработки.

Особенностью построения блока выполнения заданий является сервис ориентированная организация основных вычислительных модулей. Сам же процесс распознавания разделяется на три уровня: оптическое распознавание символов, постобработка результатов оптического распознавания и оценка точности проведенного распознавания. Каждый из таких уровней может обладать произвольным количеством реализаций и делится на модули, имплементирующие базовые интерфейсы, объединенные через общую шину электронного взаимодействия.

В заключении приводится перечень ключевых моментов, требующих дальнейшего исследования. Ими являются форматы представления результатов распознавания, методики оценки точности, методы оптического распознавания и алгоритмы коррекции результатов распознавания.

## SUMMARY

*Smirnov S.V.* **Subsystem of mass image recognition of archival documents.**

The paper describes the actual problem of mass image recognition of archival documents. Speed of document scanning, in general, is much higher than the speed of conversion to full-text form, supplemented by means of a markup, which provides supplementary service functions (the ability to search, etc.) and as a consequence there is an urgent need for designing and development of automatic systems for mass recognition.

The system development processes of mass recognition of archival documents involves a number of problems and requires careful consideration in the following areas of concern:

- definition of objectives and purpose of the recognition results;
- analysis of the quality of the source documents and their characteristics;
- calculation of the rate of recognition accuracy and definition of the allowable threshold;
- development of tools for automatic control of quality;
- design appropriated to the scale and duration of the project.

The conceptual scheme of the electronic archive and its subsystem of mass recognition are described later in the paper.

The recognition subsystem consists of four functional areas responsible for the import and export of data, preparation for processing of incoming data, execution of the processes of recognition and processing of recognition results.

The central component of the system is the scheduler, which fulfills a task, initiates the start-up and produces a dispatch through the entire set of processing modules.

The main feature of the execution block is service-oriented organization of the main computational modules. The process of recognition is divided into three levels: optical character recognition, post-correction of results and evaluation of the accuracy. Each of these levels can have an arbitrary number of implementations and is divided into modules, which implements the basic interface, and is integrated through a common communication bus.

The conclusion of the paper contains the list of key points that require further investigations: presentation formats of recognition results, methodology for assessing the accuracy, optical character recognition techniques and algorithms for the correction of recognition results.