Информационные технологии в системном анализе

Контроль комплектности документов в системах массового ввода*

М. А. Алиев, В. В. Арлазаров, Д. Г. Слугин

Аннотация. В работе были описаны основные задачи контроля комплектности документов в системах массового ввода документов. Были введены базовые определения и приведена формальная постановка задачи. Предложен алгоритм решения задачи на основе теории конечных автоматов.

Ключевые слова: *системы массового ввода, обработка документов, теория конечных автоматов.*

Введение

В настоящее время системы массового ввода документов получили широкое распространение во многих областях народного хозяйства, как в государственных учреждениях, так и в частных компаниях. Буквально за последнее десятилетие многие процессы ввода и обработки бумажных документов кардинально изменились, активно применяются системы электронного документооборота, для которых важной частью является массовый ввод бумажных документов и перевод их в электронную форму посредством сканирования, фотографирования и т. п. От качества работы этого этапа во многом зависит работа всей системы в целом, так как потеря документов, ошибки распознавания и многое другое может поставить под сомнение целостность данных и корректность функционирования всего комплекса. В данной статье рассматривается одна из таких проблем, а именно – комплектация документов.

* Работа выполнена при частичной финансовой поддержке РФФИ, проект №13–07–12171.

1. Постановка задачи

Основную задачу систем массового ввода документов [1, 2] в указанных терминах можно описать следующим образом: на вход подается поток изображений, которые надо идентифицировать и разделить на документы. Таким образом, вначале производится идентификация изображений, затем комплектация документов. Благодаря такому разделению в системе уменьшается количество ошибок.

Введем следующие определения, необходимые нам для корректной постановки задачи:

Графический образ документа — совокупность графических образов всех его страниц или частей.

Тип страницы — шаблон распознанной страницы.

Комплектация документа — однозначное сопоставление графических образов документа соответствующим типам страниц. Без ограничения общности можно считать, что каждой странице ставится в соответствие один тип.

Задача комплектации — проверка полноты и непротиворечивости документа в данный момент времени работы системы исходя из соответствующих страницам типов. В идеале каждому входящему документу ставится в соответствие набор типов страниц, который мы будем называть *тип документа*.

Например, если у нас есть анкета социологического опроса, то заполненных документов (анкет) может быть большое количество, но все они соответствуют одному типу документа — исходной анкете.

В зависимости от постановки задачи могут быть заданы различные условия и ограничения на типы страниц документа, то есть такие условия, которым удовлетворяют все документы данного типа. Примеры условий:

- Задан тип первой страницы — известно, с какой страницы начинается документ;
- Задан тип последней страницы — известно, какой страницей заканчивается документ;
- Задан порядок следования страниц;
- Заданы типы обязательных страниц — известны страницы, которые точно входят в документ;
- Задано количество страниц документа.

2. Предлагаемая схема работы модуля комплектации

Рассмотрим схему работы модуля комплектации. На вход поступают бумажные документы, которые затем в виде наборов изображений поступают в модуль комплектации. Данный модуль можно представить в виде «черного ящика», который на выходе выдает комплекты изображений, соответствующие отдельным документам и их статус — комплектный, некомплектный, подозрительный и так далее. Если говорить более строго, на входе цепочка изображений, на выходе — пары из набора изображений и вероятностей что данный набор соответствует тому или иному типу документа. Ввиду возможных ошибок комплектации, таких как некомплектность исходных документов, повреждение изображений, отсутствие изображений из-за слипания страниц в устройствах ввода (например, сканере), полностью автоматизировать данный процесс практически невозможно, особенно когда речь идет о важных коммерческих или государственных документах. Наборы изображений, требующие сверки или ручной комплектации, поступают оператору для корректировки и принятия решения. При невозможности исправления ошибок комплектации проблемные изображения откладываются для принятия решения на следующем уровне компетенции для определения, являются ли исходные документы изначально некомплектными, произошла ли ошибка формирования изображений (слипание страниц) и нужен ли повторный ввод бумажных документов.

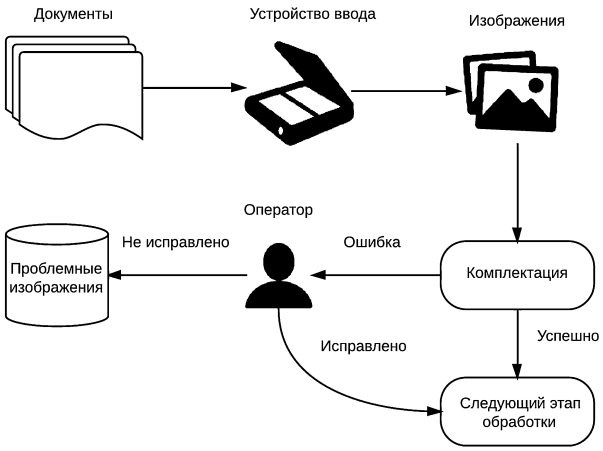

Рис. 1

3. Методы решения

От общей постановки задачи перейдем к наиболее часто встречающимся случаям в рамках систем массового ввода документов. Введем необходимые обозначения.

Пусть $D_1 \dots D_M$ список различных типов документов, которые могут поступать на вход системе. В основном мы имеем в виду бумажные документы, соответствующие тому или иному типу и которые состоят из набора страниц. Не сужая общности можно считать, что документ D_i состоит из следующих типов $\{F_{i1}, \dots, F_{ij}, \dots, F_{il_i}\}$. Обычно страницы документа можно пронумеровать по порядку согласно ГОСТ 6.38–90, что соответствует порядку их следования или сшивки. Если же порядок не предусмотрен, тогда просто нумеруем страницы по своему желанию. Для удобства можно считать, что каждое F_{ij} это положительное число – уникальный идентификатор типа страницы.

На вход системе поступает поток изображений страниц документов, обозначим эти изображения как I_n, где n — номер изображения в последовательности. Пусть у нас есть функция $\delta(I_n)$, которая определена следующим образом:

$$\delta\left(I_n\right) = \begin{cases} F_{ij}, & \text{если изображению } I_n \\ & \text{соответствует тип } F_{ij}, \\ \text{ошибка}, & \text{если тип не определен.} \end{cases}$$

Таким образом, для системы массового ввода у нас есть система идентификации, которая по изображению выдает результат, а именно какому типу оно соответствует. Существует большое количество описанных в литературе методов, например [3] и [4]. Введем следующие понятия:

- Счетчик скомплектованных страниц для документа D_i — d_i;

- Номер текущей идентифицированной страницы документа D_i по порядку (в случае его актуальности) — C_i;
- Функция порядка от текущего идентифицированного изображения для документа D_i

$$\text{Если } \delta(I) = F_{ij},$$

$$\text{то } W(I) = \begin{cases} \text{true}, & \text{если } C_i < j, \\ \text{false}, & \text{если } C_i \geqslant j. \end{cases}$$

Введем определение функции комплектности K набора изображений следующим образом

$$K\left(I_{n1}; \ldots; I_{nm}\right) = \\ = \begin{cases} D_i, & \text{если изображения содержат все типы страниц } F_{ij} \text{ документа,} \\ \text{ошибка}, & \text{если иначе.} \end{cases}$$

Это означает, что последовательность изображений содержит все типы страниц F_{ij} для конкретного i, т. е. из нее можно скомплектовать документ D_i. При достаточно большой последовательности функция может возвращать набор документов, если изображения содержат страницы нескольких документов.

В общем виде задача комплектации описывается моделью детерминированного конечного автомата [5] :
$M = (\Sigma, S, s_0, F, \sigma)$, где

- Σ — входной алфавит (конечное непустое множество входных символов);
- S — множество состояний;
- s_0 — начальное состояние ($s_0 \in S$);
- F — множество заключительных состояний $F \subset S$;
- σ — функция переходов, определенная как отображение: $\sigma : S \times \Sigma \to S$.

В данном случае входной алфавит — это все варианты типов поступающих изображений, а множество состояний — это все возможные варианты комплектности текущих незавершенных документов. Заметим, что в данной задаче начальное и заключительное состояния неважны, так как процесс может быть остановлен в любой момент, а потом продолжен с места прерывания.

Перейдем к основным задачам комплектации документов и опишем алгоритмы комплектации для каждого варианта. Очевидно, что алгоритм для каждого рассматриваемого случая и задает функцию переходов σ.

Допустим, что исходные документы комплектны, количество страниц документов равно количеству изображений, поступающих на вход и нет потери изображений (например, от слипания страниц в сканере).

Рассмотрим задачу комплектации при условии, что входной поток содержит документы одного типа D_1.

I. δ-функция всегда дает верный результат для страниц документа;

a) Документ D_1 имеет известное число страниц l_1.

В этом случае, когда количество пришедших на вход страниц равно количеству страниц в документе l_1, заканчиваем комплектацию. Отметим, что в данной схеме мы не можем получить некомплектный документ.

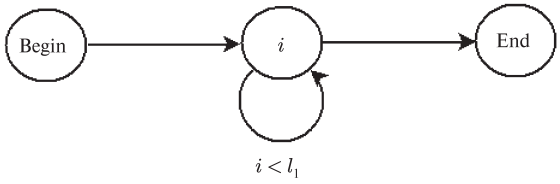

Рис. 2

b) Известен порядок следования страниц.

Рассмотрим случай, когда нам известна последовательность страниц на входе. Тогда алгоритм комплектации будет следующим:

1. Начинаем комплектовать первый документ с первого входящего изображения;
2. Собираем изображения в документ до тех пор, пока на вход не придет изображение с типом страницы не соответствующей порядку следования типов страниц документа;
3. Переходим к п. 1 для следующего изображения.

Используя данный алгоритм, мы можем уверенно собрать наборы изображений, соответствующие входным документам и определить их комплектность. Схема конечного автомата для этого случая изображена на рис. 3.

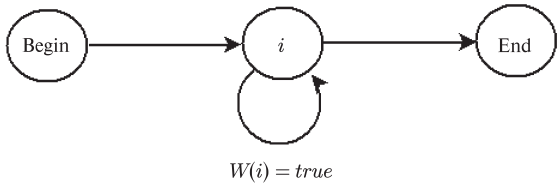

Рис. 3

Номером i обозначено состояние, при котором последней скомплектованной страницей является i-ая страница документа. Переходы не по порядку следования страниц отсутствуют, так как в этом случае мы заканчиваем комплектацию текущего документа и начинаем комплектовать следующий.

c) Известны страницы начала и/или конца документа.

Данный случай является ослаблением случая I.a. Здесь известна только страница начала и/или конца документа, сами страницы могут не иметь строгого порядка следования. Этот вариант имеет несколько отдельных модификаций — в зависимости от строгости условий присутствия соответствующих крайних страниц (только первая, только последняя, и первая и последняя, хотя бы одна из них и т. д.).

Схема конечного автомата для случая, когда должна присутствовать хотя бы одна краевая страница, изображена на рис. 4. Для остальных вариантов схемы получаются аналогично.

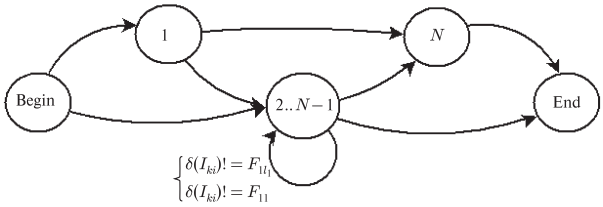

Рис. 4

Под номером 1 обозначен номер первой страницы документа, под номером N обозначен номер последней страницы документа, под $2..N-1$ обозначено любое состояние, при котором последней скомплектованной страницей является не первая и не N-я страница документа.

d) Порядок следования страниц не задан.

Часто встречаются ситуации, когда изначально неизвестна последовательность страниц входящих документов или последовательность нарушена. В таком случае, если документы следуют один за другим, мы можем комплектовать их следующим образом:

1. Начинаем комплектовать первый документ с первого входящего изображения;
2. Собираем изображения в документ до тех пор, пока на вход не придет изображение с типом страницы, которая уже встречалась;
3. Переходим к п. 1 для следующего изображения.

Схема конечного автомата для этого случая изображена на рис. 5:

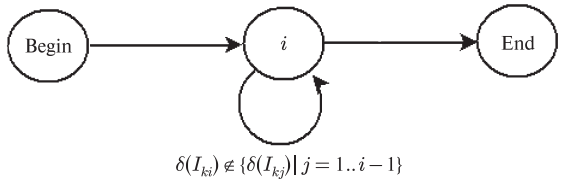

Рис. 5

II. Поток документов одного типа с наличием посторонних страниц.

Рассмотрим случаи аналогичный I, только на вход могут поступать посторонние страницы, для которых δ-функция выдает ошибку. Обычно посторонними страницами являются страницы других документов, которые на данный момент нам не интересны и типы которых мы не определяем намеренно либо не можем определить ввиду сложности документа. В этом случае алгоритм действия будет таким же с той особенностью, что при поступлении посторонних страниц счетчик скомплектованных страниц не увеличивается.

Так как алгоритмы уже описаны, то далее для этого раздела представим только схемы. В них под U понимаются посторонние страницы.

a) Документ D_1 имеет известное число страниц l_1.

В этом случае только с добавлением определенной страницы мы увеличиваем счетчик скомплектованных страниц.

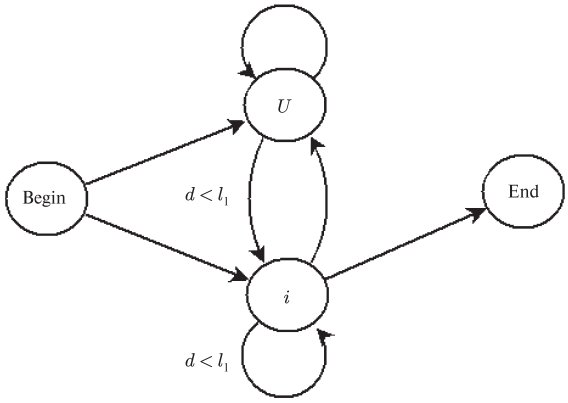

Рис. 6

b) Известен порядок следования страниц.

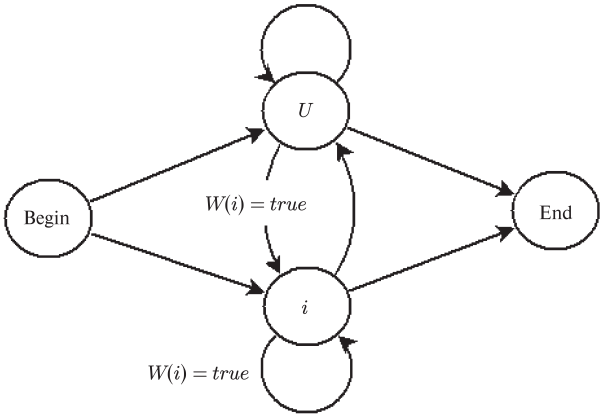

Рис. 7

c) Известны страницы начала и/или конца документа.

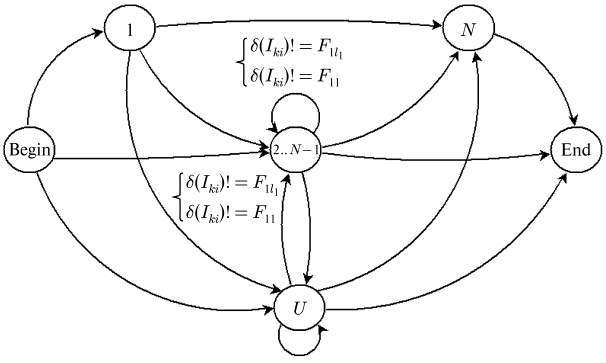

Рис. 8

Под $2..N-1$ обозначено любое состояние, при котором последней скомплектованной страницей является не первая и не N-я страница документа.

d) Порядок следования страниц не задан.
Как и на рисунке 5 в пункте I.d мы собираем изображения в документ до тех пор, пока не вход не придет изображение с типом страницы, которая уже встречалась, т. е. собираем изображения пока приходят уникальные для комплектации этого документа состояния.

III. Поток документов одного типа с возможностью δ-функции выдавать ошибку на страницах документа.
В таком случае комплектуем документы, используя неопределенные страницы как недостающие.

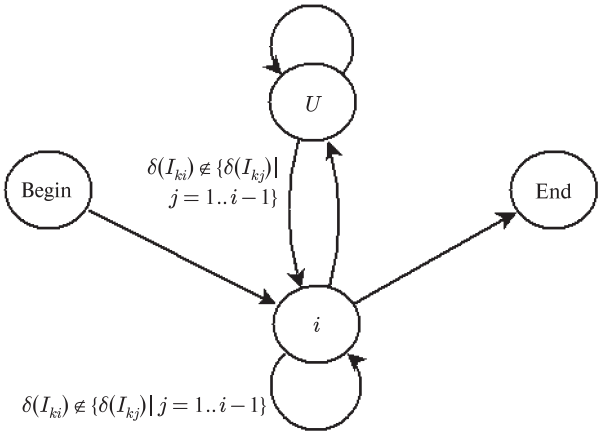

Рис. 9

a) Документ D_1 имеет известное число страниц l_1.
В этом случае с добавлением каждой страницы, как определенной, так и нет, мы увеличиваем счетчик скомплектованных страниц.

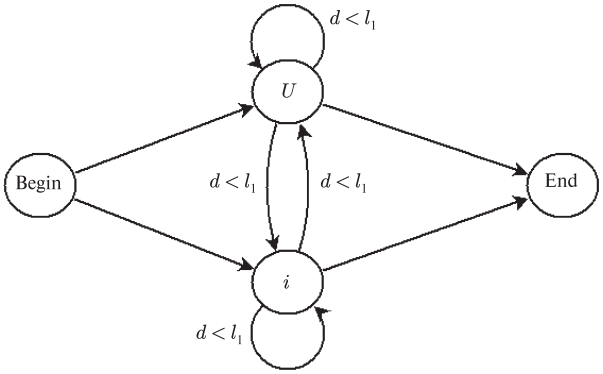

Рис. 10

b) Известен порядок следования страниц.
В случае известной последовательности страниц присваиваем неопределенной странице тип страницы по порядку их следования в документе.

c) Известны страницы начала и/или конца документа.

d) Порядок следования страниц не задан.

В случае b или с уверенно скомплектовать документ мы можем только в случае одной ошибочной страницы в последовательности, которой присваиваем недостающий тип. Если же таких страниц более одной, то ввиду невозможности точного сопоставления изображения типу такие документы отправляются на ручную комплектацию для окончательного принятия решения.

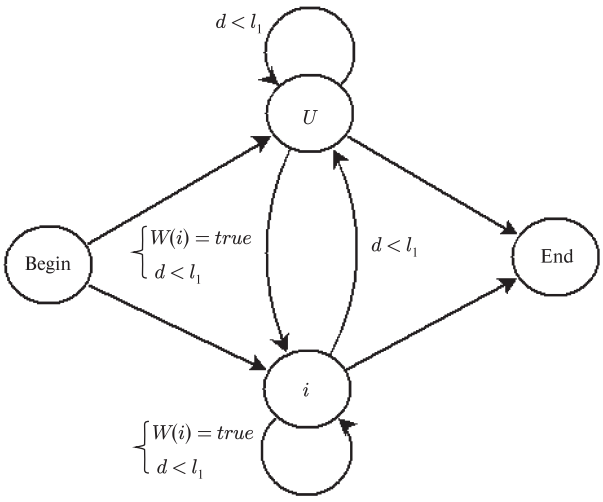

Рис. 11

IV. Поток документов разных типов.
Рассмотрим случай, когда входной поток содержит документы разных типов $D_1 \ldots D_N$. В зависимости от особенностей входного потока комплектовать документы можно следующим образом.

a) Входящие документы следуют последовательно, один за другим.
В таком случае комплектация осуществляется аналогично разделу I, только с дополнительным условием остановки процесса комплектации документа, когда на вход приходит тип страницы документа другого типа.

b) Страницы документов разных типов могут быть перемешаны между собой.

Комплектация осуществляется параллельно, для разных типов документов собираем последовательности изображений с соответствующими им типами. Алгоритмы комплектации будут такие же, как и в разделе I, с условием, что изображения с типом страницы документа, отличной от текущего документа, мы считаем посторонними страницами для его набора. Таким образом, страницы документа могут быть «размазаны» по достаточно большой последовательности изображений, но в результате все равно документ будет скомплектован правильно. Однако такой метод комплектации возможен в случае, когда δ-функция дает верный результат на изображениях входных документов. Если это не так, то для изображения, для которого не определен тип, сказать уверенно какому документу оно принадлежит невозможно. В таком случае последовательность изображений отправляется на ручную комплектацию или используются вероятностные модели оценки комплектности.

4. Случаи изначальной некомплектности документов и слипания страниц

Рассмотрим особенности комплектации документов в случае изначальной некомплектности документов и/или слипания страниц. Слипание страниц в сканере может быть представлено как удаление *последовательной* части изображений из общей последовательности. В таком случае природа некомплектности документов не является важной для модуля комплектации, можно считать, что изначально последовательность изображений была комплектна, но по определенным причинам из нее исчезло несколько подпоследовательностей. Понять истинную причину некомплектности может только оператор, имеющий на руках бумажные оригиналы документов. Сверяя их с проблемными изображениями, оператор определяет, был ли исходный документ действительно некомплектным или произошло слипание и документ надо отправить на повторный ввод.

Слипание страниц приводит не только к некомплектности документов и может нарушать не только вероятностные оценки комплектности, но и давать неверный результат даже в, казалось бы, очевидных случаях. Приведем простой пример — пусть на вход поступают документы одного типа, состоящие из двух страниц. Такой пример не является искусственным, наоборот документы, размещенные на одном листе с двух сторон (те же анкеты), используются повсеместно. Допустим произошло слипание страниц, и вторая страница одного документа следует за первой страницей другого. Комплектация пройдет успешно, однако очевидно, что результат некорректен. Например, если нам известно количество входных документов, то сверка числа ожидаемых и полученных даст нам сигнал ошибки, но определить само место ошибки при больших последовательностях изображений весьма затруднительно. Поэтому при комплектации могут использоваться интеллектуальные методы, такие как:

1. Кодирование страниц документов уникальными кодами и наборами данных.
Достаточно широко распространенный метод, когда на каждую страницу документа впечатывается или наклеивается уникальный код или штрих-код, содержащий информацию о документе.
2. Использование методов распознавания и сверки атрибутов.
Применяется в случаях, когда на страницах одного документа встречаются повторяющиеся атрибуты, например, ФИО. В таком случае использую лексические сопоставления можно

отличить похожие страницы разных документов.

3. Геометрические методы и анализ изображений.

   Используют особенности изображений, принадлежащих одному документу. Например, на многих договорах требуется подпись заявителя на каждой странице. Возможность находить и сравнивать подписи дает инструмент для отличия страниц документов одного типа.

5. Заключение

Рассмотрены основные задачи комплектации документов в системах массового ввода. Введены определения, постановка задачи, предложена модель. Авторами был разработан алгоритм на основании конечных автоматов [5], который в дальнейшем был применен и апробирован в процессах ввода пакетов документов для страховых компаний, Фонда социального страхования Российской Федерации и ряда финансовых организаций. Как показала практика применения, описанный авторами подход является достаточным для решения большого количества задач ввода. В дальнейшем авторами предполагается рассмотреть более общую модель, при которой δ-функция не дает нам однозначного ответа, какому типу соответствует изображение, а возвращает набор типов с вероятностями оценки.

Литература

1. *Арлазаров В. Л.*, *Емельянов Н. Е.* Системы обработки документов. Основные компоненты. Управление информационными потоками. Сборник трудов Института системного анализа РАН, М.: URSS, 2002.
2. *Casey D. Ferguson*, *Mohiuddin K. and Walach E.* Intelligent forms processing system, Machine Vision and Applications, vol. 5, № 3, pp. 1443–1455, 1992.
3. *Постников В. В.* Разработка методов наложения формы на графическое изображение документа // В сб. Интеллектуальные технологии ввода и обработки информации, М., 1998.
4. *Shimotsuji S. and Asano M.* Form Identification based on Cell Structure, Proc. of the 1996 Int. Conf. on Pattern Recognition (ICPR96), pp. 793–797, 1996.
5. *Белоусов А. И.*, *Ткачев С. Б.* Дискретная математика, М.: МГТУ, 2006.

Алиев Михаил Александрович. М. н. с. ИСА РАН. Окончил в 2008 г. МИСиС. Количество печатных работ: 1. Область научных интересов: обработка изображений. E-mail: aliev.michael@gmail.com

Арлазаров Владимир Викторович. Зав. лабораторией ИСА РАН. К. т. н. Окончил в 1999 г. МИСиС. Количество печатных работ: 13. Область научных интересов: распознавание образов, обработка изображений, системы массового обслуживания. E-mail: vva777@gmail.com

Слугин Дмитрий Геннадьевич. Н. с. ИСА РАН. Окончил в 2000 г. МГУ. Количество печатных работ: 7. Область научных интересов: распознавание образов, обработка изображений, электронный документооборот, распределенные вычисления. E-mail: slugindm@gmail.com