Обработка и анализ изображений и сигналов

Метод автоматической оценки качества цветовой сегментации в задаче упаковки изображений печатных документов*

Д. П. Николаев, Д. В. Полевой, Т. С. Чернов

Аннотация. Статья посвящена подходу к автоматизации контроля качества в системах сжатия изображений документов по технологии смешанного растрового содержимого (MRC). Построена модель качества изображения документа для случая его разбиения на текстовый и графический информационные слои. Предложен автоматический метод оценки качества цветовой сегментации изображения документа. Экспериментально показано, что предложенный метод имеет высокий уровень корреляции с экспертными оценками. Метод был успешно использован для автоматического поиска оптимальных параметров цветовой сегментации.
Ключевые слова: *сжатие изображений документов, автоматическая оценка качества, цветовая сегментация изображений, оптическое распознавание символов (OCR).*

Введение

В последние годы многие предприятия переходят на электронные архивы, содержащие вместо бумажных документов их цифровые изображения. Преимущества хранения изображений документов в цифровом виде очевидны: они не портятся, их легко найти (при надлежащей структуре хранения), к ним можно обратиться из любого места, например, по сети Интернет [1, 2].

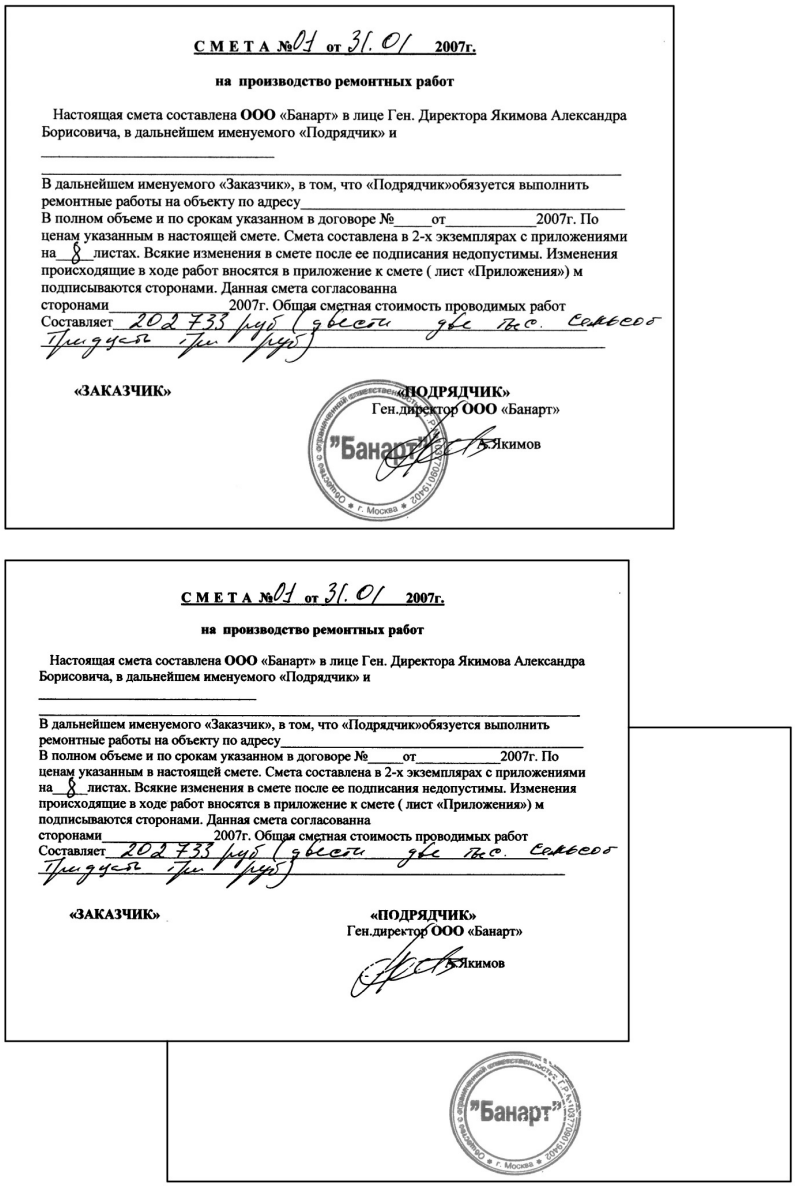
Для эффективного сжатия изображений печатных документов используется модель смешанного растрового содержимого (MRC) [3], в основе которой лежит разбиение изображения на непересекающиеся информационные слои. Каждый информационный слой содержит объекты определенного типа (текст, графика, фон) и независимо сжимается оптимальным образом. Модель MRC реализована в таких популярных форматах, как DjVu [4] и PDF/A [5].

Автоматическая цветовая сегментация (расслоение) изображения на информационные слои (рис. 1) является важнейшим этапом работы основанных на модели MRC систем сжатия [6]. Из-за визуальной природы результата, оценка качества проводится с помощью экспертных оценок, что неприемлемо для эффективного контроля качества и автоматической настройки параметров. В работе предлагается метод автоматической оценки качества цветовой сегментации изображений печатных документов.

1. Общая идея метода

Информационные слои изображения печатного документа содержат качественно различную информацию, поэтому проводится их независимая оценка с последующим комбинированием результатов. Обобщенная функция оценки качества цветовой сег-

* Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта №13–07–12173.



СМЕТА №01 от 31.01 2007г.

на производство ремонтных работ

Настоящая смета составлена ООО «Банарт» в лице Ген. Директора Якимова Александра Борисовича, в дальнейшем именуемого «Подрядчик» и

_____

В дальнейшем именуемого «Заказчик», в том, что «Подрядчик»обязуется выполнить ремонтные работы на объекту по адресу_____
В полном объеме и по срокам указанном в договоре №_____от_____2007г. По ценам указанным в настоящей смете. Смета составлена в 2-х экземплярах с приложениями на 8 листах. Всякие изменения в смете после ее подписания недопустимы. Изменения происходящие в ходе работ вносятся в приложение к смете (лист «Приложения») м подписываются сторонами. Данная смета согласованна сторонами_____2007г. Общая сметная стоимость проводимых работ Составляет 202 733 руб (двести две тыс. семьсот тридцать три руб)

«ЗАКАЗЧИК»

«ПОДРЯДЧИК»
Ген.директор ООО «Банарт»
Якимов

"Банарт"

Рис. 1. Пример цветовой сегментации изображения печатного документа

ментации изображения документа представлена в формуле (1):

$$Q = \sum_i w_i Q_i, \qquad (1)$$

где Q_i — функция оценки качества i типа информационного слоя;

w_i — задаваемый экспертом весовой коэффициент значимости i слоя.

Основной идеей предлагаемых функций оценок информационных слоев является использование заранее подготовленной идеальной разметки, задающей модели качества информационных слоев с учетом специфики принадлежащих к ним объектов. В работе рассматривается оценка качества текстового и графического типа информационного слоя.

2. Оценка качества текстового слоя с помощью OCR-систем

Текстовый слой имеет важные отличительные особенности: он должен содержать только текстовую информацию, а его изображение удобно представлять монохромным (бинарным, однобитовым). Эти особенности позволяют проверять качество выделения и сжатия изображения в текстовом слое путем попытки его распознавания автоматической системой оптического распознавания символов (OCR).

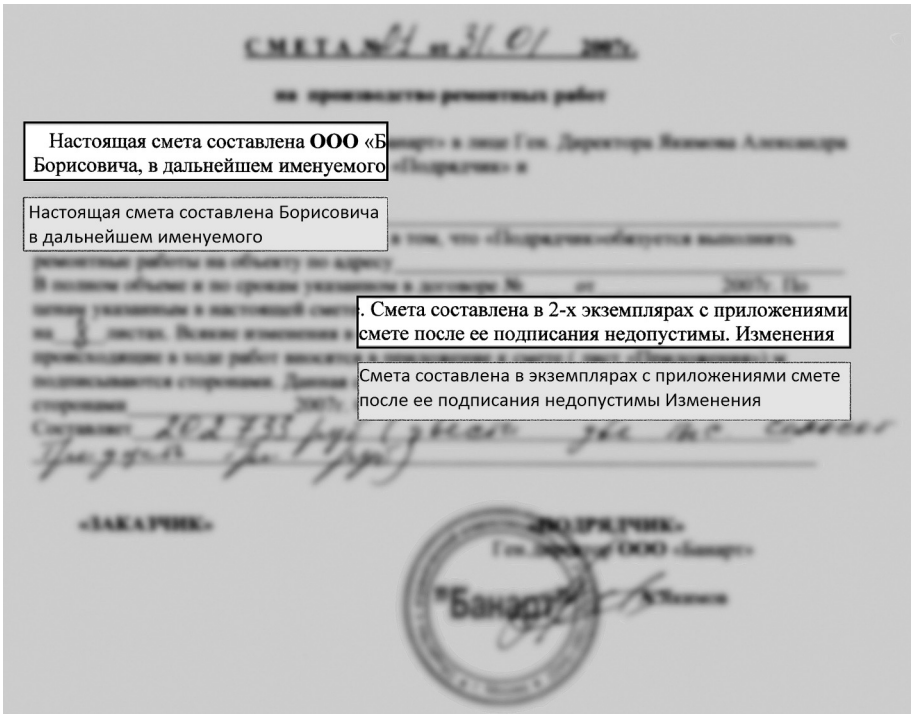


Рис. 2. Пример разметки интересующих для OCR областей

Для оценки качества бинаризации подобный подход описан в работе [7].

Оценка с помощью OCR-систем интуитивно ясна: чем больше точность распознавания (отношение количества верно распознанных символов к общему количеству символов в разметке документа), тем более высокую оценку качества получит изображение. Проблема такого метода в том, что необходимы большие трудозатраты на предварительную посимвольную разметку текста на изображениях из тестовой выборки.

Предлагается модифицировать метод, и проверять не посимвольное качество распознавания, а точность распознавания слов в некоторых областях изображений (рис. 2). Для каждого изображения из тестовой выборки размечается набор прямоугольных областей, для которых указан набор значимых для распознавания OCR-системой слов.

Для подсчета оценки используется следующая формула:

$$Q_{OCR} = \frac{\sum_{w_i} |w_i| \cdot recognized(w_i)}{\sum\limits_{w_i} |w_i|} \cdot 100\%, \qquad (2)$$

$$recognized(w_i) = \begin{cases} 1, & w_i \text{ распознано полностью,} \\ 0, & \text{в противном случае} \end{cases} \qquad (3)$$

где w_i — слово из разметки, $|w_i|$ — его длина (количество символов).

Такой подход нормализует результаты по отношению к длинам слов, поскольку полностью верно распознанное длинное слово должно давать больший вклад, чем короткое. С другой стороны, неизбежны ошибки в одном символе в длинных словах, которые огрубляют оценку, но подобные ошибки выровняются среди областей всех изображений тестовой выборки. Важно, что предложенная статистика считается глобально по всем областям разметки изображения.

В качестве внешней OCR-системы для проверки метода была выбрана популярная кросс-платформенная система Tesseract OCR [8].

3. Оценка качества графического слоя

Графический слой изображения печатного документа должен содержать цветные информативные участки изображения, такие как печати, штампы, подписи и т. п. Наличие в маске слоя не принадлежащих ему по смыслу пикселей ухудшает качество работы алгоритма: размер выходного файла увеличивается в силу применения менее эффективных алгоритмов сжатия; объекты, подлежащие сжатию без потерь, получают различные артефакты. Попадание информативных пикселей графического слоя

Рис. 3. Исходное изображение (слева) и его карта насыщенности

в текстовый слой означает их последующую бинаризацию, что сильно скажется на визуальном качестве выходного документа.

Требуется разработать метод, оценивающий качество построенной маски слоя с точки зрения отклонения ее от некоторой заранее известной идеальной маски. Классической функцией оценки такого отклонения является среднеквадратичное отклонение (*MSE*) для случая двух монохромных изображений x (построенной маски) и y (идеальной маски) размера N:

$$MSE(x, y) = \frac{1}{N}\sum_{i=1}^{N} b(x_i, y_i),$$
$$b(x_i, y_i) = \begin{cases} 1, & x_i \neq y_i, \\ 0, & x_i = y_i, \end{cases} \qquad (4)$$

Проблема такого подхода в высокой трудоемкости попиксельной разметки идеальной маски графического слоя для тестовой выборки изображений за разумное время [9].

В работе [6] описывается схема расслоения изображения печатного документа технологии Cognitive PDF/A. Она основывается на построении гистограммы насыщенности изображения и последующей бинаризации карты насыщенности. Картой насыщенности (рис. 3) называется полутоновое изображение, в котором каждому пикселю (R, G, B) исходного изображения соответствует значение его насыщенности S, вычисляемое по формуле (5) (при работе с 8-битными изображениями).

$$S = 255 - \frac{3 \cdot \min(R, G, B)}{R + G + B}. \qquad (5)$$

Для упрощения вычислений можно использовать формулу:

$$S = \max(R, G, B) - \min(R, G, B). \qquad (6)$$

С целью повышения удобства работы для пользователя, полученная карта насыщенности инвертируется: $S_{2,i} = 255 - S_{1,i}$ (рис. 4, а).

Следующим шагом схемы расслоения является глобальная пороговая бинаризация (метод Оцу [10]), т. е. вычисляется некоторое единственное пороговое значение, в соответствии с которым пиксели после вычисления насыщенности попадают в текстовый или графический слой. Идею пороговой бинаризации удобно использовать для эффективной разметки идеальной маски: при ручном изменении порога насыщенности происходит монотонное добавление (удаление) в разметку графического слоя всех удовлетворяющих (не удовлетворяющих) этому порогу пикселей.

Для разметки инвертированной карты насыщенности экспертом выделяется набор непересекающихся прямоугольных областей, и для области устанавливается оптимальный с точки зрения визуального восприятия порог бинаризации пикселей, принадлежащих этой области (рис. 4, б).

Для непосредственной оценки качества предлагается две метрики, основанные на описанном подходе. Первой из них является вычисление суммарного MSE для размеченных областей:

$$Q_{BINMSE}(X, Y) = \sum_{i=1}^{N_A} MSE(X_i, Y_i), \qquad (7)$$

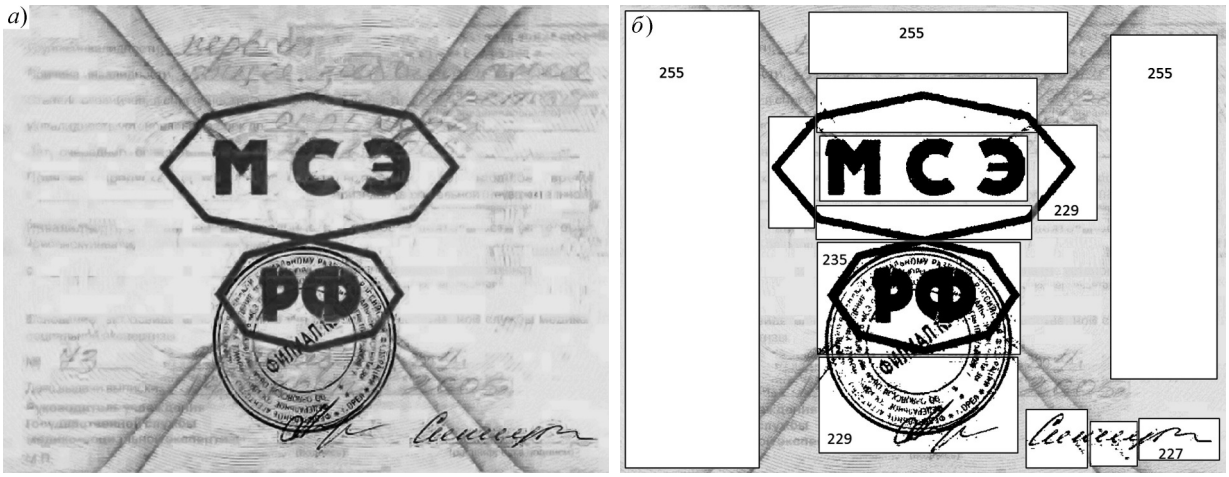


Рис. 4. Инвертированная карта насыщенности (слева) и разметка порогов бинаризации для ее областей

где X — исходное изображение (инвертированная карта насыщенности);

Y – бинаризованное изображение (полученная маска графического слоя);

X_i – соответствующая i области часть разметки карты насыщенности;

Y_i – соответствующая i области часть графического слоя;

N_A – количество областей разметки.

У данной метрики есть недостаток — вообще говоря, *идеальную* разметку расслоения составить невозможно: от присутствия или отсутствия в маске слоя некоторых пикселей (чаще всего, на границе объектов) визуальное восприятие качества может остаться прежним. Предлагается следующий метод оценки: вместо расчета попиксельного отклонения проводить расчет отклонений плотностей пикселей для целых областей.

Введем величину ρ плотности пикселей для области разметки A размерами $W_A \times H_A$:

$$\rho(A) = \frac{N_1(A)}{N_0(A) + N_1(A)} = \frac{N_1(A)}{W_A \cdot H_A}, \qquad (8)$$

где $N_k(A)$ — количество пикселей в области разметки Aсо значением $k \in \{0, 1\}$.

Тогда, второй функцией оценки будет:

$$Q_{BINCNT}(X, Y) = \frac{\sum\limits_{k=1}^{N_A} (\rho(X_k) - \rho(Y_k))^2 \cdot (W_{A_k} \cdot H_{A_k})}{\sum_{k=1}^{N_A} W_{A_k} \cdot H_{A_k}}. \qquad (9)$$

Плотность пикселей — относительная величина, поэтому отклонения плотностей взвешиваются по площадям областей.

4. Анализ пригодности предложенных методов

В качестве меры пригодности метода оценки слоя предлагается использование коэффициента линейной корреляции (Пирсона) между оценками метода и средней экспертной оценкой (MOS) на сегментированной фиксированным алгоритмом тестовой выборке изображений. Подобный подход использовался в работе [11]. В качестве экспериментальной системы расслоения взята реализация технологии Cognitive PDF/A [6].

Экспертное оценивание результатов цветовой сегментации собранной и размеченной тестовой выборки из 15 изображений документов проводилось среди 5 экспертов и показало высокую их согла-

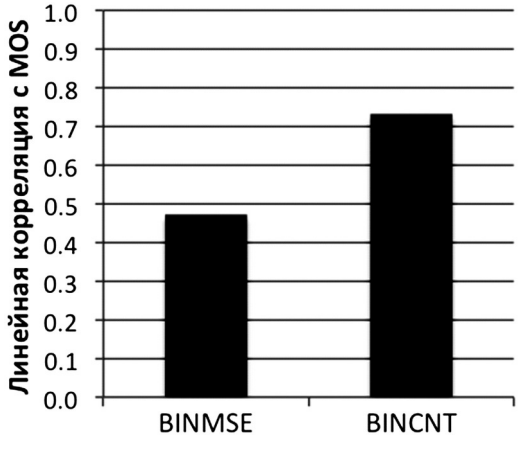


Рис. 5. Вычисленные значения линейной корреляции между MOS и оценками предложенных методов на тестовой выборке изображений

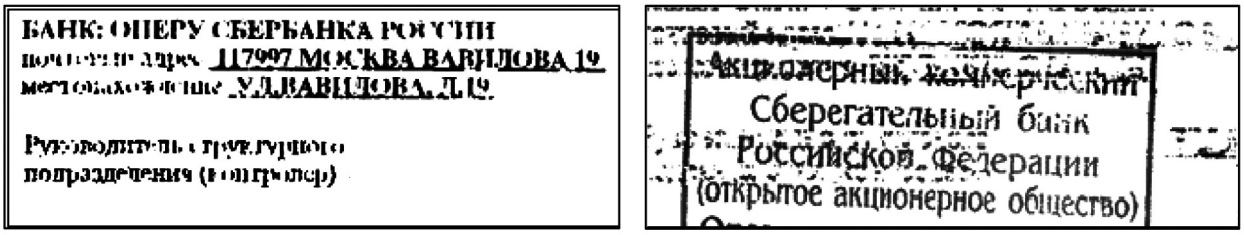


Рис. 6. Пример проблемных фрагментов текстового и графического слоя до проведения оптимизации на тестовой выборке изображений

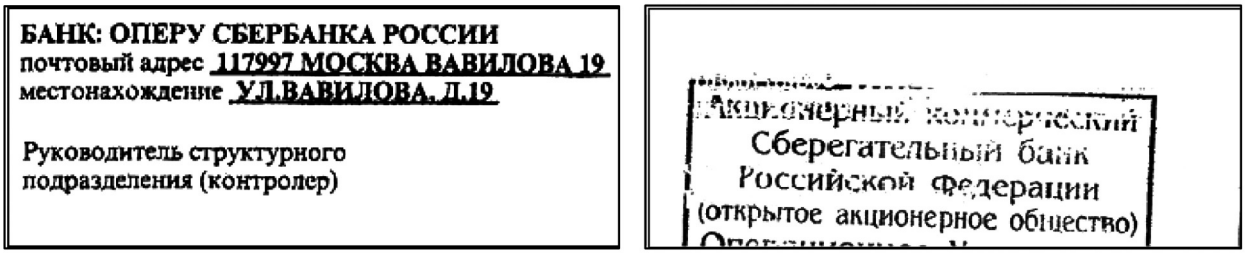


Рис. 7. Пример оптимизации качества проблемных фрагментов текстового и графического слоя

сованность. В опросе использовалась 20-балльная шкала. На рис. 5 представлены вычисленные значения корреляции, отмечающие, что для оценки качества графического слоя в самом деле выгоднее использовать метод оценки отклонений плотностей пикселей.

Итоговая функция оценки качества определяется как комбинация наилучших оценок слоев:

$$Q = w_1 \cdot Q_{BINCNT} + w_2 \cdot Q_{OCR}. \qquad (10)$$

Веса w_1 и w_2 выставляются экспертом для регулировки значимости графического или текстового слоя.

5. Автоматическая оптимизация качества цветовой сегментации

При наличии автоматического метода оценки качества цветовой сегментации становится возможной автоматическая оптимизация качества. По причине сложной структуры как системы цветовой сегментации, так и системы оценки качества, строится система типа «черный ящик». Входами являются некоторые внутренние параметры, влияющие на результат цветовой сегментации и размеченная тестовая выборка изображений документов, а выходом — значение оценки качества для результата расслоения с данными параметрами.

Был проведен эксперимент с 4 переменными (целыми, с диапазоном от 0 до 255) в качестве некоторых пороговых параметров цветовой сегментации. Для автоматической оптимизации использовалось программное обеспечение NOMAD [12]. На рис. 6 приведены примеры фрагментов слоев проблемного изображения тестовой выборки, полученные в результате работы системы при исходных параметрах (начальном приближении, заданном разработчиком). На рис. 7 изображены фрагменты того же изображения, полученные при автоматически подобранных оптимальных параметрах.

До оптимизации часть маски текстового слоя попадала в графический, тем самым делая нечитаемым текст, а также создавая сильный шум в области печати на графическом слое. Введенные в работе методы оценки качества учитывают оба этих фактора, поэтому оптимизация прошла успешно с точки зрения визуального восприятия человека. Про другие изображения выборки можно сказать, что качество проблемных (при исходных параметрах алгоритма цветовой сегментации) изображений также визуально улучшилось, а изначально удачных — осталось прежним.

Заключение

В работе был предложен новый подход к автоматической оценке качества цветовой сегментации в задаче упаковки изображений документов. Основной идеей подхода является независимая оценка информационных слоев. Оценка текстового слоя производится с помощью OCR-систем, а графического — путем разметки идеальной маски слоя на карте насыщенности изображения. Для проверки пригодности предложенных методов был проведен экс-

пертный опрос и вычислены значения корреляции между показателями методов и результатами экспертных оценок. Итоговый метод оценки был успешно опробован для автоматической оптимизации качества цветовой сегментации.

Литература

1. *Арлазаров В. Л.*, *Славин О. А.* Алгоритмы распознавания и технологии ввода текстов в ЭВМ // Информационные технологии и вычислительные системы. 1996. Т. 6. Вып. 1. С. 48–54.
2. *Арлазаров В. В.*, *Постников В. В.*, *Шоломов Д. Л.* Cognitive Forms — система массового ввода структурированных документов // Управление информационными потоками М.: URSS, 2002. С. 37–49.
3. *Queiroz R.* De, *Buckley R.*, *Xu M.* Mixed Raster Content (MRC) Model for Compound Image Compression. 2000. P. 1–12.
4. *Haffner P. et al.* A general segmentation scheme for DjVu document compression // ISMM. 2002.
5. *I*SO 19005–1:2005. Document management — Elecronic document file format for long-term preservation — Part 1: Use of PDF 1.4 (PDF/A-1). 2005.
6. *Усилин С. А.*, *Николаев Д. П.*, *Постников В. В.* Cognitive PDF/A — технология оцифровки текстовых документов для публикации в Интернет и долговременного архивного хранения // Труды Института системного анализа РАН. Технологии программирования и хранения данных / Под ред.: Арлазаров В. Л., Емельянов Н. Е. М.: Ленанд/URSS, 2009. С. 159–173.
7. *Rangoni Y.*, *Shafait F.*, *Breuel T.* Ocr based thresholding // Conference on Machine Vision. 2009. P. 1–4.
8. *T*esseract OCR [Online], [Электронный ресурс], http://code.google.com/p/tesseract-ocr/
9. *Самойлов О. С.*, *Полевой Д. В.* Оценка качества сегментации изображений печатных документов на примере системы оптического распознавания текстов OPEN-OCR // Труды ИСА РАН. 2010. Т. 58. С. 164–171.
10. *Otsu N.* A Threshold Selection Method from Gray-Level Histograms // IEEE Transactions on Systems, Man, and Cybernetics. 1979. Vol. 9. № 1. P. 62–66.
11. *Grgi S.*, *Mrak M.* Reliability of Objective Picture Quality Measures. 2004. Vol. 55. № 1. P. 3–10.
12. *N*OMAD Black Box Optimization [Online], [Электронный ресурс], http://www.gerad.ca/nomad

Николаев Дмитрий Петрович. Зав. сектором ИППИ РАН. К.ф.-м. н. Окончил в 2000 г. МГУ им. М. В. Ломоносова. Кол-во печатных работ: 103. Область научных интересов: быстрые алгоритмы обработки изображений. E-mail: dimonstr@iitp.ru

Полевой Дмитрий Валерьевич. С. н. с. ИСА РАН. К. т. н. Окончил в 2004 г. МФТИ. Кол-во печатных работ: 9. Область научных интересов: методы искусственного интеллекта, оптическое распознавание, системы обработки документов. E-mail: dvpsun@gmail.com

Чернов Тимофей Сергеевич. Аспирант НИТУ МИСиС. Окончил в 2013 г. НИТУ МИСИС. Область научных интересов: обработка и анализ изображений и сигналов, системный анализ, системное программирование. E-mail: chernov.tim@gmail.com