ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 2 (8)

УДК 004.94 DOI: https://doi.org/10.30837/2522-9818.2019.8.044

А. В. Ищенко

РАЗРАБОТКА МОДУЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ОБРАБОТКИ ОТСКАНИРОВАННЫХ ДОКУМЕНТОВ НА БАЗЕ КОМБИНИРОВАННОГО МЕТОДА СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Предметом исследования в статье является модуль сегментации, созданный на базе комбинированного метода сегментации изображений, и внедренный в интеллектуальную систему обработки отсканированных документов, используемую в Одесской полиграфической компании "Студия "Печать". **Целью** работы является разработка модуля сегментации изображений для повышения оперативности интеллектуальной системы обработки отсканированных документов полиграфической компании "Студия "Печать". Для этого используется комбинированный метод сегментации изображений отсканированных документов, который позволяет сократить время обработки изображения. В статье решаются следующие **задачи**: анализ существующих методов сегментации изображений, которые используются в интеллектуальных системах обработки отсканированных документов; разработка процедур для модуля сегментации на основе комбинированного метода сегментации изображений для интеллектуальной системы обработки отсканированных документов. В работе используются **методы**: методы цифровой обработки изображений, методы фильтрации и морфологического анализа изображений, методы математического анализа, нейронные сети. Получены следующие **результаты**: результаты обработки изображений с помощью интеллектуальной системы обработки отсканированных документов на базе предложенного модуля сегментации подтверждают работоспособность процедур модуля сегментации изображений. Среднее время обработки изображений отсканированных документов составило 5,3 с по сравнению с ранее полученным – 42 с, что позволяет сделать вывод об увеличении оперативности исследуемой интеллектуальной системы обработки отсканированных документов. **Выводы:** Внедрение разработанного модуля сегментации изображений в интеллектуальную систему обработки отсканированных документов полиграфической компании "Студия "Печать" позволило сократить время обработки изображений отсканированных документов в 8 раз при сохранении достаточного качества, благодаря чему увеличилась оперативность данной интеллектуальной системы.

Ключевые слова: сегментация изображений; отсканированные документы; обработка документов; интеллектуальная система.

Введение

Современные предприятия все чаще переходят на использование электронных архивов документов, т. к. работа с ними имеет преимущества: облегчается поиск и хранение информации в больших базах данных, облегчается передача документов по цифровым каналам связи, качество документов со временем не ухудшается, в связи с чем увеличивается надежность их хранения. В зависимости от задач и целей предприятий обработка таких документов выполняется с помощью различных интеллектуальных систем обработки отсканированных документов (ИСООД).

В связи с научно-техническим прогрессом объем документации на предприятиях имеет тенденцию к постоянному росту. Часто приходится сканировать печатные документы с целью последующей пересылки по электронной почте или архивного хранения, поэтому возникает задача обработки большого количества отсканированных документов. В связи с этим возрастают требования к оперативности обработки печатных документов, следовательно, и к оперативности интеллектуальных систем. Недостаточная оперативность ИСООД приводит к: 1) длительному времени обработки документов; 2) невозможности предоставления документов в установленные сроки; 3) увеличению времени на поиск документов в больших БД и их подготовки при аудиторских проверках.

Таким образом, актуальным является усовершенствование существующих ИСООД для повышения их оперативности.

Анализ проблемы и существующих методов

Существует большое разнообразие ИСООД, используемых на различных предприятиях, например, наиболее распространенные – FineReader, ABBYY FormReader, OCR CuneiForm, Readiris Pro7, OmniPage 11 и др. Они отличаются оперативностью и качеством обработки документов.

Предполагается, что недостаточная оперативность существующих ИСООД во многом может определяться недостаточной оперативностью блока сегментации, который используется в таких системах. Сегментация является важным этапом в работе интеллектуальных систем обработки документов. При качественном сканировании изображения документов часто имеют достаточно большой размер, использование которых неэффективно в электронных архивах. Сегментация предназначена для уменьшения обрабатываемой информации, и чем меньше времени уходит на обработку изображения, тем выше оперативность интеллектуальной системы. Качество обработки документов интеллектуальной системой определяется качеством сегментации изображения, которое должно быть достаточным для конечного пользователя.

Определим качество сегментации как процент пикселей изображения, которые правильно сегментированы как определенный класс: текст, фото, графика и фон. Будем считать его достаточным, если вероятность сходства изображения,

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 2 (8) ISSN 2524-2296 (online)

сегментированного экспертом, и изображения, сегментированного определенным методом, составляет не менее 85%. Такое числовое значение выбрано исходя из предварительных исследований оценки качества сегментации изображения исследователями.

Сегментация – важный этап обработки документа в ИСООД, целью которого является разделение изображения на однородные области по определенному признаку. Такие области могут содержать, например, только текст, или таблицы, или графику, или фотоизображения и др. Сегментация изображений отсканированных документов широко используется в системах оптического распознавания символов (OCR) [1]. В большинстве случаев качество распознавания символов в системах OCR зависит от результата сегментации. В системах, использующих модель растрового контента (MRC) [2, 3], изображение представляется в виде слоев, каждый из которых содержит объекты определенного типа и независимо сжимается выбранными кодерами. Использование MRC-модели для представления изображений предназначено для определенных типов документов и иногда приводит к значительному искажению изображения документа. В [4] сегментация применяется для извлечения графики и текста из изображений, содержащих чертежи, в работе также решается проблема пересечения областей графики и текста. В [5] – сегментация используется для извлечения текста, фото и линий разделителей из статей, рекламных буклетов, визиток. Сегментация с помощью классификации связных компонент изображений на текстовые области и нетекстовые используется в [6].

Для решения задачи сегментации изображений отсканированных документов с целью повышения оперативности ИСООД необходимы методы, которые удовлетворяют быстрой обработке изображения. Для эффективности работы ИСООД качество сегментации таких документов должно быть достаточным для цели обработки качества.

Целью данной статьи является повышение оперативности ИСООД, которая используется в полиграфической компании "Студия "Печать" (г. Одесса) путем сокращения времени обработки изображения. Для сокращения времени обработки изображения отсканированного документа разрабатывается модуль, который для сегментации изображения использует комбинированный метод сегментации изображения отсканированного документа [7].

Постановка задачи исследования

Одним из направлений работы полиграфической компании полиграфической компании "Студия "Печать" (г. Одесса) является обработка изображений отсканированных документов с целью упорядочивания и хранения информации.

Как уже было замечено, хранение документов в электронном виде удобно для быстрого поиска необходимой информации, передачи их по электронной почте и много другого. Поэтому ЧП "Студия "Печать" предоставляет сервис по переводу документов в цифровую форму для создания и наполнения электронных архивов и каталогов предприятий, а также наполнений систем документооборота библиотек, коммерческих и некоммерческих компаний с помощью ИСООД Cognitive PDF/A (рис. 1) [1].

Документ, поступивший на сканирование в компанию "Студия "Печать", проходит несколько этапов обработки: определяется объем документа – количество страниц, формат, параметры сканирования; этап сканирования; изображения отсканированных документов вводятся в компьютер, и, согласно схеме (рис. 1), обрабатываются для распознавания и архивируются для дальнейшего хранения; обработанные изображения отсканированных документов копируются на удобный для клиента носитель.

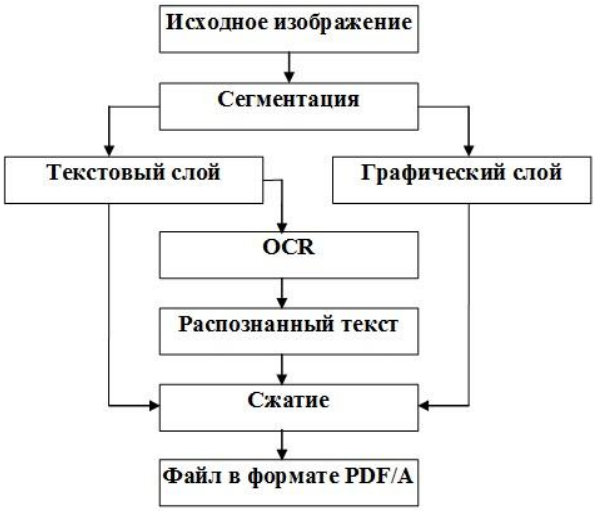

Рис. 1. Общая схема ИСООД Cognitive PDF/A [1]

Сканирование документов на предприятии происходит с помощью поточных сканеров Fujitsu-fi-5950, которые предназначены для сканирования больших объемов документов. Скорость сканирования на таком оборудовании составляет 125 стр./мин. Также данный вид сканера позволяет выполнять двусторонне сканирование, что влияет на скорость сканирования. Это обеспечивает быструю подготовку пакета документов, например, для аудита. Также на предприятии "Студия "Печать" предоставляется услуга по сканированию чертежей, планов, схем больших размеров (до 1000 мм). Для этого используются сканеры Colortrac (Smart LF Ci 40).

Перед компанией "Студия "Печать" клиенты ставят следующие задачи: скорость обработки большого количества документов в периоды аудита при достаточном качестве сегментации изображений отсканированных документов на текст, фото и графику.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 2 (8)

Однако система, представленная на рис. 1, не обеспечивала достаточно быструю обработку изображений отсканированных документов. Предположительно, это связано с тем, что модуль сегментации изображений не удовлетворял требованиям одновременно к быстрой обработке и достаточному качеству сегментации изображений отсканированных документов.

Разработка модуля сегментации изображений для ИСООД, используемой в компании "Студия "Печать"

Учитывая перечисленные задачи в предыдущем подразделе, которые ставятся перед компанией "Студия "Печать", предлагается использовать модуль сегментации на основе комбинированного метода сегментации изображения отсканированного документа вместо модуля сегментации ИСООД Cognitive PDF/A, которая используется в компании "Студия "Печать".

В системе оцифровки документов Cognitive PDF/A компании "Студия "Печать" реализован модуль сегментации для различных типов документов, например, финансовых документов, журнальных статей и других документов, которые содержат текст, графику, фото.

На рис. 2 приведена схема ИСООД Cognitive PDF/A с внедренным разработанным модулем сегментации (выделено серым). Согласно данной схеме в первом модуле происходит сбор данных, в котором печатные документы из внешней среды поступают на сканер. Во втором модуле отсканированные документы сохраняются в базе данных используемых изображений документов. В третьем модуле изображение отсканированного документа проходит предварительную обработку, и затем сегментируется.

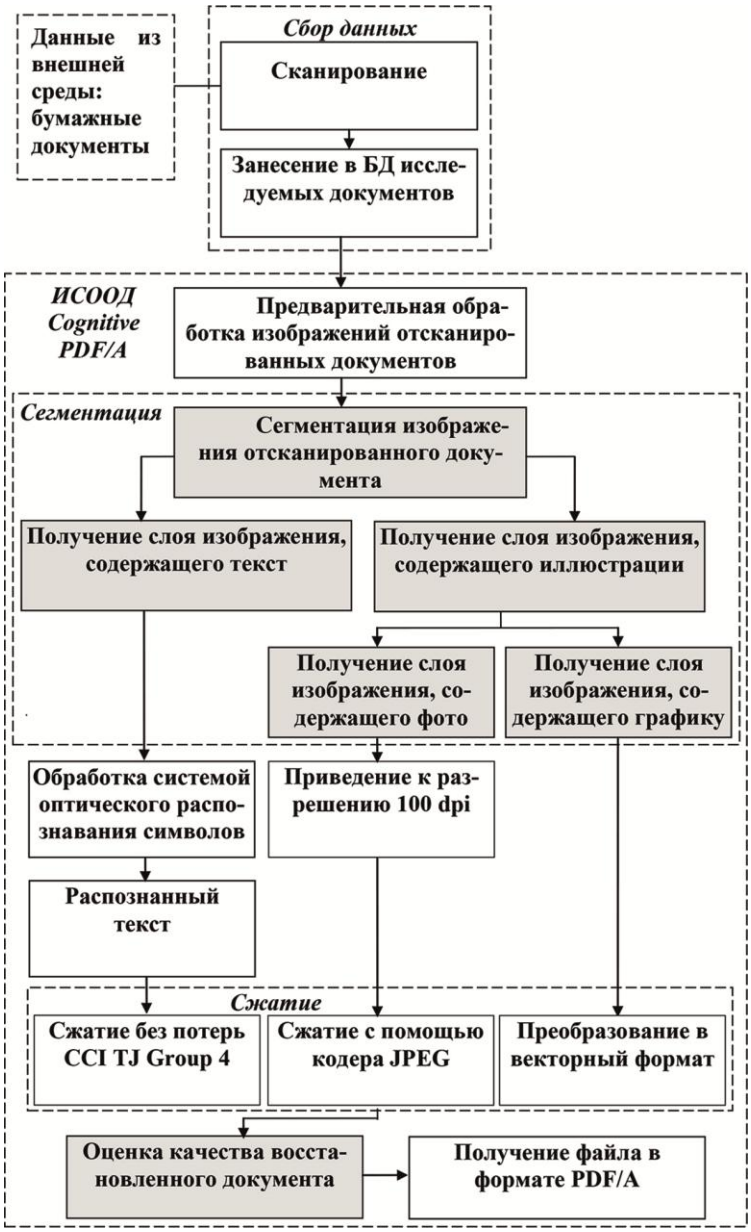

Рис. 2. Схема усовершенствованной ИСООД Cognitive PDF/A

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 2 (8) ISSN 2524-2296 (online)

В результате сегментации изображение отсканированного документа представляется в виде трех изображений, или слоев. Первый слой содержит области исходного изображения отсканированного документа, соответствующее текстовой информации, второе изображение слоя содержит графику, третье – фото. Такое представление изображения позволяет распознать текстовые области с помощью модуля оптического распознавания символов.

В качестве модуля распознавания символов в системе Cognitive PDF/A используется OCR CuneiForm с открытым исходным кодом. Заключительным модулем системы является архивирование полученных изображений слоев и распознанного текста в формате PDF/A. Данный формат базируется на описании стандарта PDF версии 1.4 от Adobe Systems Inc. и предназначен специально для долговременного архивного хранения электронных документов.

Существующая в компании "Студия "Печать" ИСООД не отвечает требованиям оперативности, т. к. имеет модуль сегментации, не удовлетворяющий требованиям к достаточному качеству сегментации и времени обработки изображений отсканированных документов. Поэтому для повышения оперативности ИСООД вместо существующего модуля сегментации предлагается использовать модуль, процедуры которого реализуют разработанный в [7] комбинированный метод сегментации и идентификации однородных областей изображений отсканированных документов на базе разработанной модели [8] представления этих изображений.

Этапы комбинированного метода сегментации.

1. Сегментация изображения отсканированного документа на области, содержащие иллюстрации (фото и графику), и текстовые области, и фон, используя метод с усредняющей фильтрацией [9]. Данный метод основан на методе сегментации Блумберга с заполнением отверстий [10]. Метод выделения областей иллюстраций Блумберга с заполнением отверстий использует операцию уменьшения изображения, которая используется для сокращения обрабатываемой информации. Рассмотрим операцию порогового уменьшения изображения 4×1 более подробно.

Пусть имеется бинарное изображение, в котором значения фоновых пикселей равны 1, а значения остальных пикселей равны 0. Изображение разбивается на блоки размерности 2×2. Каждый блок пикселей изображения размерностью 2×2 заменяется одним пикселем. Значение пикселя, на которое заменяется блок 2×2 равно либо 0, либо 1 в зависимости от выбранного порога, который может принимать значения от 1 до 4. Значение пикселя, на которое заменяется блок пикселей 2×2 равно 1, если сумма значений четырех пикселей блока больше либо равна выбранному порогу, в противном случае значение пикселя равно 0. После однократной операции уменьшения 4×1 количество пикселей изображения уменьшается с 2^n до 2^{n-2}, где n – размерность блока (рис. 3).

В отличие от метода Блумберга вместо операции уменьшения изображения, которая ухудшает качество сегментации, используется усредняющая фильтрация с пороговым преобразованием. Это позволяет сохранить достаточное качество сегментации изображения на текстовые области и области, содержащие иллюстрации.

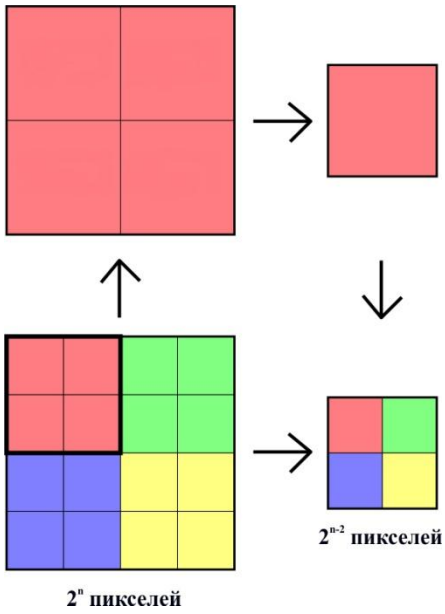

Рис. 3. Операция порогового уменьшения изображения 4×1

2. Идентификация областей фото и графики. Для этого используется метод идентификации областей графики и фото [7] с вычислением статистическо-геометрических признаков, а именно: оценки математического ожидания высоты перепада интенсивности на границах областей однородной интенсивности – признак $f_1(i, j)$ и соотношения размеров объектов – признак $f_2(i, j)$. При этом используется блочная обработка для сокращения времени обработки изображения. Полученные значения признаков для каждого блока l нормировались по формуле:

$$z_k(i, j) = \frac{p_k(i, j)}{p_{k\max}}, \; k = 1, 2, \qquad (1)$$

где $z_k(i, j)$ – нормированное значение признака k-го признака в блоке с индексами i, j, $p_k(i, j)$ – исходное ненормированное значение k-го признака блока с индексами i, j; $p_{k\max}$ – максимальное значение k-го признака по всем блокам областей, содержащих графику/фото для всех изображений.

На рис. 4 изображены примеры обучающей выборки в пространстве двух данных признаков z_1 и z_2 для 5 изображений. Маркеры "×", "□" отображают примеры обучающих выборок, которые имеют значения выходов "графика" и "фото" соответственно.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 2 (8)

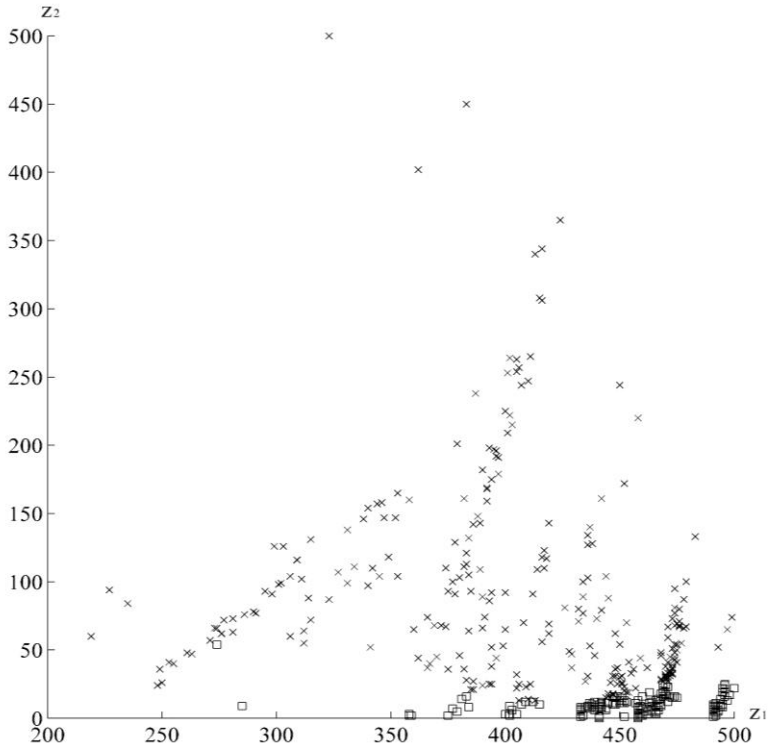

Рис. 4. Примеры обучающей выборки в пространстве двух признаков

3. Классификация областей графики и фото полиномиальной машиной опорных векторов. Выбор классификации анализируемых областей машиной опорных векторов обоснован тем, что для классификации методом опорных векторов, в отличие от большинства других методов, достаточно небольшого набора данных, что делает его наиболее быстрым для нахождения решающих правил. Работа данного метода сводится к решению задачи квадратичного программирования, которая всегда имеет единственное решение, и находит разделяющую гиперплоскость максимальной ширины, что позволяет осуществлять наиболее точную классификацию.

Машина опорных векторов в работе предназначена для классификации объектов на 2 класса. На вход каждой машины опорных векторов подается объект, который ранее не классифицировался. Далее рассчитывается расстояние от неизвестного объекта до границы каждого из классов. Затем неизвестный объект относится к тому классу, расстояние до границы которого наименьшее.

В работе с помощью машины опорных векторов [12] классифицируются области, содержащие фотоизображения, и области графики. На этапе 1 комбинированного метода сегментации на каждом изображении выделяются области, содержащие графику и фото. Затем эти области разбиваются на блоки размерности $N \times N$, и для каждого блока вычисляется вектор признаков $p^{train}(i, j) = \left(p_1(i, j), p_2(i, j)\right)$, который нормируется по формуле (1). Обучающую выборку составляют нормированные вектора признаков $z^{train}(i, j) = \left(z_1(i, j), z_2(i, j)\right)$ для каждого блока изображения. В нее также входят целевые значения $y(i, j)$ для каждого блока. В качестве целевых значений принимается мода, т.е. наиболее часто встречающееся значение всех меток пикселей для данного блока, т. к. она показывает наиболее вероятную принадлежность к определенному классу.

4. Выделение текстовых фрагментов из фона, используя обработку в окрестности каждого пикселя, а именно: сначала используется низкочастотная фильтрация, которая сглаживает значения интенсивности изображения внутри однородных областей текста и фона, затем применяется пороговое преобразование.

Для эксперимента были отобраны документы трех различных компаний-клиентов "Студия "Печать": аудиторской компании, медицинского центра и учебного заведения. Размеры изображений отсканированных документов составляли 2550×3506 со средним размером файла 2,44 Мб. Сканирование этих документов производилось с разрешением в 300 точек/дюйм.

Результаты обработки изображений (рис. 5) с помощью усовершенствованной ИСООД путем предложенного модуля сегментации подтверждают его работоспособность. Среднее время обработки изображений отсканированных документов составило 5,8 с по сравнению с использованием настоящего модуля сегментации – 42 с. Сокращение время обработки документов позволяет сделать вывод об увеличении оперативности усовершенствованной ИСООД.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 2 (8) ISSN 2524-2296 (online)

Рис. 5. Результаты обработки изображений отсканированных документов разработанным модулем сегментации:
а – исходные изображения; *б* – результат экспертной разметки; *в* – результат сегментации.
(светло-серым обозначена графика, серым – текст, темно-серым – фото)

Методы сжатия, используемые в интеллектуальной системе Cognitive PDF/A

Для увеличения коэффициента сжатия слой изображения, содержащий графику и фото, и слой, содержащий текст, сжимаются разными способами. Этот модуль в системе обработки документов не менялся. Рассмотрим его более подробно.

Слой, содержащий фотоизображение, преобразуется к разрешению 100 dpi, а затем кодируется методом сжатия JPEG. Метод сжатия JPEG основан на дискретном косинусном преобразовании (ДКП), в результате которого происходит потеря информации, но обеспечивается высокий коэффициент сжатия при минимальной потере данных для графики и фотоизображений, в которых различие между соседними пикселями незначительное. Основные этапы метода сжатия JPEG представлены на рис. 6.

Рис. 6. Этапы работы кодера JPEG

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 2 (8)

1. Предварительная обработка. Исходное изображение переводится из цветового пространства RGB, с компонентами, отвечающими за красную (Red), зеленую (Green) и синюю (Blue) составляющие цвета пикселей, в цветовое пространство *YCrCb* (YUV), где *Y* — яркостная составляющая, а *Cr* и *Cb* — компоненты, отвечающие за хроматический красный и хроматический синий цвет соответственно. Затем устраняются группы пикселей за счет уменьшения количества пикселей для каналов цветности. При этом уменьшается разрешение в цветоразностных каналах и сохраняются разрешения в канале яркости.

2. Яркостная компонента *Y* и отвечающие за цвет компоненты *Cb* и *Cr* разбиваются на блоки 8×8 пикселей, содержащие информацию о пикселях, и к каждому из них применяется ДКП. В результате получается матрица 8×8 амплитудных значений, которые отражают скорости изменения цвета в изображении.

3. Квантование — этап, на котором происходит округление высокочастотных ДКП-коэффициентов для фильтрации амплитуд, которые незначительно влияют на общий вид изображения. Высокочастотная информация не воспринимается человеческим глазом, поэтому коэффициенты, отвечающие за высокие частоты, можно хранить с меньшей точностью. На этом этапе происходит основная потеря информации.

4.Кодирование полученных данных стандартными методами (кодирование серий и кодирование по Хаффману).

Человеческий глаз восприимчив к изменению яркости, а незначительные изменения цвета он замечает хуже, поэтому при сжатии JPEG запоминается больше информации о разнице между яркостями пикселей и меньше — о разнице между их цветами. Различия в цвете соседних пикселей мала, и человеческим глазом практически не заметна, поэтому изображение после восстановления выглядит почти неизменным. Поэтому можно архивировать массивы для *Cr* и *Cb* компонент с большими потерями и, соответственно, с большими коэффициентами сжатия.

Слой, содержащий графику, для удобства представляется в векторном виде. Это позволяет значительно сократить дисковое пространство для хранения документов.

Слой, содержащий текстовые области, заключает в себе основную информацию документа, поэтому он сохраняется в исходном разрешении, а для его кодирования используется метод сжатия без потерь CCITT (International Telegraph and Telephone Committie) Group 4. Метод CCITT Group 4 предназначен для сжатия монохромных изображений с высокой степенью сжатия и используется для сжатия черно-белых изображений, в которых преобладают большие одноцветные области. В основе этого метода сжатия лежит поиск и исключение из исходного изображения последовательностей данных, которые дублируются. При обработке отдельных рядов пикселей выполняется замена последовательности подряд идущих черных и белых пикселей числом, равным их количеству. Затем этот ряд кодируется по Хаффману. Достоинством метода сжатия CCITT Group 4 является его быстродействие и простота реализации.

Оценка качества восстановленных изображений отсканированных документов

Для создания БД документов предприятий, изображения отсканированных документов после сегментации сжимаются. Классификация областей иллюстраций на фото и графику расширяет возможности базовой ИСООД Cognitive PDF/A: слой изображения, содержащий фото, сжимается отдельно с помощью кодера JPEG, а слой графики представляется в векторном формате. Поэтому в работе разработан модуль для оценки качества восстановленного документа после сжатия.

Лучше всего качество изображения может оценить человеческий глаз, даже несмотря на то, что такая оценка является субъективной. При исследовании качества восстановленного изображения отсканированного документа использовалась процедура поиска медианных консенсусных ранжирований по Кемени-Снеллу и Куку-Сейфорду, описанная в [11]. Согласно данному методу в экспертизе принимала участие группа из 10 экспертов, задачей которых было оценить качество восстановленных после сжатия изображений отсканированных документов путем голосования. В качестве экспертов были приняты клиенты и сотрудники компании "Студия "Печать". Эксперты должны были визуально оценить, насколько им подходит качество восстановленного изображения отсканированного документа соответствует исходному изображению в интервале от 1 до 10 баллов. Т. е. эксперты сравнивали изображения отсканированных документов до обработки его в ИСООД и после. На основании оценок изображений формировались индивидуальные ранжирования. Далее формировалась представительная статистическая выборка псевдослучайных наборов индивидуальных ранжирований, для которой рассчитывались медианы Кемени-Снелла и Кука-Сейфорда. Затем вычислялась степень согласованности индивидуальных и экспертных ранжирований, которая оценивалась с помощью коэффициента конкордации Кэнделла. Оценка качества восстановленного изображения отсканированного документа, вычисленная по этому методу, получилась близкой к оценке экспертов.

Существует множество методов кодирования изображений, которые предназначены для сжатия изображений любого типа. Но в работе был предложен комбинированный метод сегментации, предназначенный для работы с определенным типом изображений отсканированных документов, реализованный с помощью процедур для модуля сегментации ИСООД, что увеличивает коэффициент сжатия изображений отсканированных документов при сохранении высокого их качества восстановления по сравнению, например, с таким методом кодирования как JPEG.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 2 (8) ISSN 2524-2296 (online)

Выводы

В полиграфической компании "Студия "Печать" с используемой ИСООД существующий модуль сегментации был заменен разработанным модулем для сегментации изображений отсканированных документов. Данная система используется для автоматизации перевода документов в цифровую форму для создания электронных архивов и каталогов документов предприятий.

Был проведен вычислительный эксперимент, который показал, что степень сжатия изображения после обработки их в усовершенствованной ИСООД увеличилась на 42% при качестве восстановления по процедуре поиска медианных консенсусных ранжирований 8 из 10 баллов по сравнению с ИССОД, используемой компанией "Студия "Печать", что удовлетворяет задаче хранения архивных документов. Полученный результат может говорить о достаточном качестве сегментации изображений на текст, графику, фото и фон, каждый из которых сжимается определенным кодером.

Разработанный модуль сегментации, внедренный в ИСООД компании "Студия "Печать", позволяет сократить время обработки изображений отсканированных документов в 8 раз при достаточном качестве сегментации, которое определили пользователи системой обработки изображений.

Список літератури

1. 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, 2009. Т. 45. С. 159–173.
2. Rajeswari, N., Rathnapriya, S., Nijandan, S. (2014), "Test Segmentation of MRC Document Compression and Decompression by Using MATLAB", *International Conference on Engineering Technology and Science-(ICETS'14), Tamilnadu, India,* Vol. 3, Special Issue 1, P. 914–919.
3. Antonacopoulos, A., Pletschacher, S., Bridson, D. and Papadopoulos, C. (2009), "ICDAR2009 Page Segmentation Competition", *10th International Conference on Document Analysis and Recognition, Barcelona, Spain*, P. 1371–1374. DOI: https://doi.org/10.1109/ICDAR.2009.27
4. Thai, V. H., Tabbone, S. (2010), "Text Extraction from Graphical Document Images Using Sparse Representation", *ACM International Conference Proceeding Series: International Workshop on Document Analysis Systems - DAS'2010, Jun 2010, Boston, United States, ACM*, P. 143–150.
5. Erkilinc, S., Saber, E., Jaber, M. (2012), "Text, photo, and line extraction in scanned documents", *Journal of Electronic Imaging*, Vol. 21 (3), P. 033006-1–033006-18.
6. Bukhari, S. S., Azawi, M. A, Shafait, F, Breuel, T. (2010), "Document image segmentation using discriminative learning over connected components", *The 9th IAPR International Workshop DAS 2010 (Document Analysis Systems). Boston, Massachusetts. USA*, P. 183–190.
7. Polyakova, M., Ishchenko, A., Volkova, N., Pavlov, O. (2018), "The combining segmentation method of the scanned documents images with sequential division of the photo, graphics, and the text areas", *Eastern-European Journal of Enterprise Technologies*, No. 5/2 (95), P. 6–16. DOI: https://doi.org/10.15587/1729-4061.2018.142735
8. Ishchenko, A., Polyakova, M., Kuvaieva, V., Nesteryuk, A. (2018), "Elaboration of structural representation of regions of scanned document images for MRC model", *Eastern-European Journal of Enterprise Technologies*, No. 6/2 (96), P. 32–38. DOI: https://doi.org/10.15587/1729-4061.2018.147671
9. Polyakova, M., Ishchenko, A., Huliaieva, N. (2018), "Document image segmentation using averaging filtering and mathematical morphology", *14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). Lviv-Slavske, Ukraine*, P. 966–969. Doi: 10.1109/TCSET.2018.8336354
10. Bloomberg, D. S. (1992), "Multiresolution Morphological Approach to Document Image", *Visual Communications and Image Processing, Boston, MA, United States, SPIE*, Vol. 1818, P. 648–663.
11. Boltenkov, V., Kuvaieva, V., Galchonkov, O., Ishchenko, A. (2018), "Application of the assignment problem in the calcucation of median consensus rankings", *Eastern-European Journal of Enterprise Technologies*, No. 4 (94), P. 27–35.
12. Chu, W., Keerthi, S. (2002), "A general formulation for support vector machines", *C. J. Ong.In Proc. of the 9th Int. Conf. on Neural Information Processing (ICONIP '02)*, Singapore.
13. Гонсалес Р. С., Вудс Р. Е., Эддинс С. Л. *Цифровая обработка изображений в среде MATLAB*. М. : Техносфера, 2006. 616 с.
14. Харалик Р. Статистический и структурный походы к описанию текстур. ТИИЭР, 1979. Т. 67, № 5. С. 98–120.
15. Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002), "Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, Vol. 46, No. 1-3, P. 389–422.
16. Wang, H., Khoshgoftaar, T., Napolitano, A. (2011), "An Empirical Study of Software Metrics Selection Using Support Vector Machine", *Proceedings of the 23rd International Conference on Software Engineering & Knowledge Engineering (SEKE'2011)*, Eden Roc Renaissance, Miami Beach, USA, P. 83–88.
17. Сойфер В. А. Методы компьютерной обработки изображений : под ред. В. А. Сойфера. М. : Физматлит, 2003. 784 с.
18. Бологова Н. М., Рубан І. В. Дослідження моделей та методів обробки зображень та шляхи вдосконалення технологій розпізнавання маркерів в системах доповненої реальності. *Сучасний стан наукових досліджень та технологій в промисловості.* 2019. № 1 (7). С. 25–33. DOI: https://doi.org/10.30837/2522-9818.2019.7.025

References

1. Usylyn, S. A., Nykolaev, D. P., Postnykov, V. V. (2009), "Cognitive PDF / A - the technology of digitizing text documents for publication in the Internet and long-term archiving" ["Cognitive PDF/A – texnologyya ocyfrovky tekstovih dokumentov dlya publykacyy v Ynternet y dolgovremennogo arhyvnogo hranenyya"], *Trudi Ynstytuta systemnogo analyza RAN. Texnologyy programmyrovanyya yhranenyya dannih / pod red. Arlazarov V.L., Emelyanov N.E,* Moscow : LENAND, Vol. 45. P. 159–173.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2019. No. 2 (8)

2. Rajeswari, N., Rathnapriya, S., Nijandan, S. (2014), "Test Segmentation of MRC Document Compression and Decompression by Using MATLAB", *International Conference on Engineering Technology and Science-(ICETS'14), Tamilnadu, India,* Vol. 3, Special Issue 1, P. 914–919.
3. Antonacopoulos, A., Pletschacher, S., Bridson, D. and Papadopoulos, C. (2009), "ICDAR2009 Page Segmentation Competition", *10th International Conference on Document Analysis and Recognition, Barcelona, Spain*, P. 1371–1374. DOI: https://doi.org/10.1109/ICDAR.2009.27
4. Thai, V. H., Tabbone, S. (2010), "Text Extraction from Graphical Document Images Using Sparse Representation", *ACM International Conference Proceeding Series: International Workshop on Document Analysis Systems - DAS'2010, Jun 2010, Boston, United States, ACM*, P. 143–150.
5. Erkilinc, S., Saber, E., Jaber, M. (2012), "Text, photo, and line extraction in scanned documents", *Journal of Electronic Imaging*, Vol. 21 (3), P. 033006-1–033006-18.
6. Bukhari, S. S., Azawi, M. A, Shafait, F, Breuel, T. (2010), "Document image segmentation using discriminative learning over connected components", *The 9th IAPR International Workshop DAS 2010 (Document Analysis Systems). Boston, Massachusetts. USA*, P. 183–190.
7. Polyakova, M., Ishchenko, A., Volkova, N., Pavlov, O. (2018), "The combining segmentation method of the scanned documents images with sequential division of the photo, graphics, and the text areas", *Eastern-European Journal of Enterprise Technologies*, No. 5/2 (95), P. 6–16. DOI: https://doi.org/10.15587/1729-4061.2018.142735
8. Ishchenko, A., Polyakova, M., Kuvaieva, V., Nesteryuk, A. (2018), "Elaboration of structural representation of regions of scanned document images for MRC model", *Eastern-European Journal of Enterprise Technologies*, No. 6/2 (96), P. 32–38. DOI: https://doi.org/10.15587/1729-4061.2018.147671
9. Polyakova, M., Ishchenko, A., Huliaieva, N. (2018), "Document image segmentation using averaging filtering and mathematical morphology", *14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). Lviv-Slavske, Ukraine*, P. 966–969. DOI: https://doi.org/10.1109/TCSET.2018.8336354
10. Bloomberg, D. S. (1992), "Multiresolution Morphological Approach to Document Image", *Visual Communications and Image Processing, Boston, MA, United States, SPIE*, Vol. 1818, P. 648–663.
11. Boltenkov, V., Kuvaieva, V., Galchonkov, O., Ishchenko, A. (2018), "Application of the assignment problem in the calcucation of median consensus rankings", *Eastern-European Journal of Enterprise Technologies*, No. 4 (94), P. 27–35.
12. Chu, W., Keerthi, S. (2002), "A general formulation for support vector machines", *C. J. Ong.In Proc. of the 9th Int. Conf. on Neural Information Processing (ICONIP '02)*, Singapore.
13. Gonsales, R. S., Vuds, R. E., Eddins, S. L. (2006), *Digital image processing in MATLAB* [*Cyfrovaya obrabotka yzobrazhenyj v srede MATLAB*], Moscow : Tehnosfera, 616 p.
14. Haralyk R. (1979), "Statistical and structural approaches to the description of textures" ["Statisticheskiy i strukturnyy podkhody k opisaniyu tekstur"], TYYER, Vol. 67, No. 5, P. 98–120.
15. Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002), "Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, Vol. 46, No. 1-3, P. 389–422.
16. Wang, H., Khoshgoftaar, T., Napolitano, A. (2011), "An Empirical Study of Software Metrics Selection Using Support Vector Machine", *Proceedings of the 23rd International Conference on Software Engineering & Knowledge Engineering (SEKE'2011)*, Eden Roc Renaissance, Miami Beach, USA, P. 83–88.
17. Sojfer, V. A. (2003), *Computer image processing methods* [*Metodi kompyuternoj obrabotky yzobrazhenyj*] : pod red. V. A. Sojfera, Moscow : Fyzmatlyt, 784 p.
18. Bolohova, N., Ruban, I. (2019), "Image processing models and methods research and ways of improving marker recognition technologies in added reality systems", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (7), P. 25–33. DOI: https://doi.org/10.30837/2522-9818.2019.7.025

Поступила (Received) 31.05.2019

Відомості про авторів / Сведения об авторах / About the Authors

Іщенко Олеся Володимирівна – Одеський національний політехнічний університет, старший викладач кафедри прикладної математики та інформаційних технологій інституту комп'ютерних систем, Одеса, Україна; e-mail: alesya.ishchenko@gmail.com; ORCID: http://orcid.org/0000-0002-7882-4718.

Ищенко Алеся Владимировна – Одесский национальный политехнический университет, старший преподаватель кафедры прикладной математики и информационных технологий института компьютерных систем, Одесса, Украина.

Ishchenko Alesya – Odessa National Polytechnic University, Senior Lecturer of the Department of Applied Mathematics and Information Technologies, Odessa, Ukraine.

РОЗРОБКА МОДУЛЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ОБРОБКИ ВІДСКАНОВАНИХ ДОКУМЕНТІВ НА БАЗІ КОМБІНОВАНОГО МЕТОДУ СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ

Предметом дослідження в статті є модуль сегментації, створений на базі комбінованого методу сегментації зображень, і впроваджений в інтелектуальну систему обробки відсканованих документів, яка використовується в Одеській поліграфічній компанії "Студія "Друк". **Метою** роботи є розробка модуля сегментації зображень для підвищення оперативності інтелектуальної системи обробки відсканованих документів поліграфічної компанії "Студія "Друк". Для цього використовується комбінований метод сегментації зображень відсканованих документів, який дозволяє скоротити час обробки зображення. У статті вирішуються наступні **завдання**: аналіз існуючих методів сегментації зображень, які

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2019. № 2 (8) ISSN 2524-2296 (online)

використовуються в інтелектуальних системах обробки відсканованих документів; розробка процедур для модуля сегментації на основі комбінованого методу сегментації зображень для інтелектуальної системи обробки відсканованих документів. В роботі використовуються **методи**: методи цифрової обробки зображень, методи фільтрації і морфо-логічного аналізу зображень, методи математичного аналізу, нейронні мережі. Отримані наступні **результати**: Результати обробки зображень за допомогою інтелектуальної системи обробки відсканованих документів на базі запропонованого модуля сегментації підтверджують працездатність процедур модуля сегментації зображень. Середній час обробки зображень відсканованих документів становив 5,3 с в порівнянні з раніше отриманим - 42 с, що дозволяє зробити висновок про збільшення оперативності інтелектуальної системи обробки відсканованих документів, яка досліджується. **Висновки**: Впровадження розробленого модуля сегментації зображень в інтелектуальну систему обробки відсканованих документів поліграфічної компанії "Студія "Друк" дозволило скоротити час обробки зображень відсканованих документів в 8 разів при збереженні достатньої якості сегментації, завдяки чому збільшилася оперативність даної інтелектуальної системи.

Ключові слова: сегментація зображень; відскановані документи; обробка документів; інтелектуальна система.

DEVELOPMENT OF AN INTELLIGENT PROCESSING SYSTEM MODULE FOR SCANNED DOCUMENTS BASED ON THE COMBINED IMAGE SEGMENTATION METHOD

The **subject** of research in the article is a segmentation module, created on the basis of a combined method of image segmentation, and embedded in an intelligent processing system for scanned documents used in the Odessa printing company “Studio “Print”. The **aim** of the work is to develop a module of image segmentation to improve the efficiency of the intellectual system of processing scanned documents at the printing company "Studio "Print". The combined method of image segmentation of scanned documents, which reduces the processing time of the image is used with this purpose. The article solves the following **problems**: analysis of existing methods of image segmentation, which are used in intelligent systems for processing scanned documents; development of procedures for the segmentation module based on the combined image segmentation method for an intelligent system for processing scanned documents. The work uses the following **methods**: methods of digital image processing, methods of filtering and morphological image analysis, methods of mathematical analysis, neural networks. The following **results** were obtained: The results of image processing using an intelligent system for processing scanned documents based on the proposed segmentation module confirm the operability of the procedures of the image segmentation module. The average processing time for images of scanned documents was 5.3 seconds compared to the previously obtained - 42 seconds, which allows to conclude that the efficiency of the investigated intellectual system for processing scanned documents is increased. **Conclusions**: The introduction of the developed image segmentation module into the intellectual processing system of scanned documents of the printing company “Print Studio” reduced the processing time of images of scanned documents by 8 times while maintaining sufficient quality, which increased the efficiency of this intelligent system.

Keywords: image segmentation; scanned documents; document processing; intelligent system

Бібліографічні описи / Bibliographic descriptions

Іщенко О. В Розробка модуля інтелектуальної системи обробки відсканованих документів на базі комбінованого методу сегментації зображень. *Сучасний стан наукових досліджень та технологій в промисловості*. 2019. № 2 (8). С. 44–53. DOI: https://doi.org/10.30837/2522-9818.2019.8.044.

Ishchenko, A. (2019), "Development of an intelligent processing system module for scanned documents based on the combined image segmentation method", *Innovative Technologies and Scientific Solutions for Industries*, No. 2 (8), P. 44–53. DOI: https://doi.org/10.30837/2522-9818.2019.8.044.