ВЫДЕЛЕНИЕ ТЕКСТА НА СЛОЖНОМ ЦВЕТНОМ ФОНЕ

Роман Телятников, Иван Шумский, Ариф Мамедов, Анатолий Протосавицкий, Екатерина Матусевич, Екатерина Степанькова

***Аннотация**: В работе исследовано решение прикладной задачи выделения текста на сложном цветном фоне. Разработана эффективная процедура фильтрации фона от текста путем свертки цветного изображения в полутоновое. В основу процедуры фильтрации положена теория классификации. С целью повышения эффективности фильтрации предложен способ компенсации цветовых искажений, вызванных аппаратурой сканирования. Для практического применения процедуры выработаны конкретные рекомендации по выбору цветового пространства и ядра функции расстояния.*

***Ключевые слова**: фильтрация, цветовая модель, гистограмма, функция расстояния.*

***ACM Classification Keywords:** CCS - Computing methodologies - Computer graphics - Image manipulation - Image processing; CCS - Applied computing - Document management and text processing - Document capture - Optical character recognition.*

Введение

Существует огромный перечень прикладных задач, в которых на изображении требуется качественно выделить объект на сложном фоне. Если речь идет об обработке одного изображения, то для этой цели вполне пригодны профессиональные графические редакторы (Adobe Photoshop, Corel Photo-Paint, GIMP, и др. [Wikipedia, 2013]). В этих редакторах реализованы инструменты типа «Свободное выделение (Лассо)», «Умные ножницы» и «Выделение переднего плана», позволяющие вручную выполнить фильтрацию.

В случае необходимости обработки множества изображений, например, при обработке кадров видеоряда или при оцифровке сканов однотипных документов, требуется автоматизация процесса фильтрации. При этом допускается настройка параметров фильтра на одном изображении, но при обработке остальных изображений параметры должны автоматически подстраиваться под изменяющиеся цветовые характеристики объекта и фона.

Задача формулируется следующим образом: необходимо разработать процедуру фильтрации, увеличивающую расстояние между цветом объекта и цветом фона в RGB или другом цветовом пространстве.

В данной работе представлен анализ различных методов, которые могут быть использованы при построении процедуры фильтрации, а также описан итоговый алгоритм, позволяющий решать поставленную задачу с приемлемым уровнем качества.

Разработка велась для повышения эффективности чтения (OCR) цветного текста на изменяющемся цветном фоне в интересах прикладной задачи считывания данных с удостоверяющих личность документов (паспортов, идентификационных карт, водительских удостоверений и др.).

Исходные данные. Постановка задачи

В распоряжении имеем набор цветных RGB-изображений области с текстом, вырезанной из отсканированных изображений документов. Некоторые примеры из серии таких изображений представлены на рисунке 1.

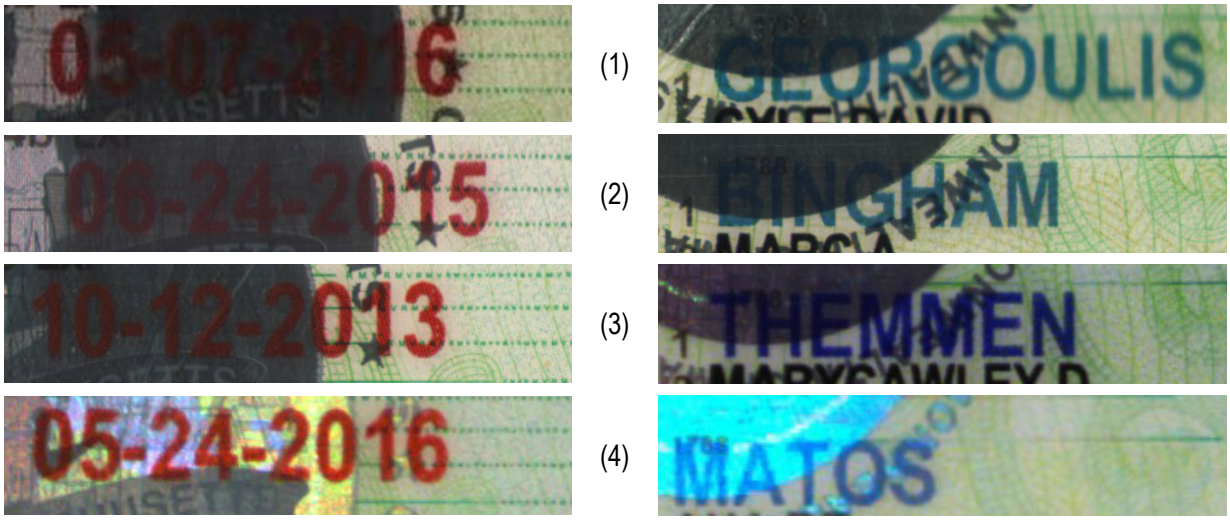

Рис 1. *Изображения текстовых полей, вырезанные из сканов идентификационных карт (а) и водительских удостоверений (б) штата Массачусетс (США).*

Процедура фильтрации предполагает предварительное назначение (вручную) точек текста и точек фона. Поэтому мы можем перенести рассматриваемую задачу в область теории распознавания образов [Журавлев, 1978]. В этом случае под множеством объектов X понимается множество всех пикселей изображения, которое необходимо разделить на подмножество пикселей текста X^{text} и подмножество пикселей фона X^{fond}. Каждый объект (пиксель) описывается 3-хмерным вектором признаков $x = (f_1,f_2,f_3)$. Значения компонент f_i вектора зависят от выбранной цветовой модели представления изображения, например в RGB-пространстве вектор признаков будет иметь вид x = (r,g,b), а для HLS модели x = (h,l,s). В качестве обучающей выборки будет выступать те пиксели, которые мы пометили как принадлежащие тексту $X^{m\ text} = \{x_1,x_2,\ldots x_m\}$ и пиксели, принадлежащих фону $X^{n\ fond} = \{x_1,x_2,\ldots x_n\}$.

Если для формализации понятия сходства между цветом произвольного пикселя изображения и цветом пикселей из обучающей выборки ввести функцию расстояния $\rho(x,x')$ то можно говорить о классической задаче метрической классификации [Айвазян, 1989]. Чем меньше значение этой функции, тем более схожи цвета двух сравниваемых пикселей x и x'.

К метрическим алгоритмам классификации относятся: метод ближайших соседей, метод потенциальных функций, метод парзеновского окна, алгоритм вычисления оценок и др. [Журавлев, 2006]. Фактически, выбор того или иного алгоритма классификации будет определять способ построения разделяющей классы поверхности.

Специфические особенности решаемой задачи накладывают ограничения на применимость тех или иных методов. Например, малый объем обучающей выборки $X^{m\ text}$ и $X^{n\ fond}$ (как правило, не более десяти точек для каждой подвыборки) не позволяет использовать методы байесовской классификации, в том числе метод парзеновского окна. Тот факт, что все включенные в обучающую выборку точки имеют одинаковую

важность, практически сводит метод потенциальных функций и алгоритм вычисления оценок к методу ближайших соседей. В дальнейшем мы проанализируем эффективность применения для решаемой задачи фильтрации метода ближайшего соседа (k=1), а также влияние на результат вида ядра функции расстояния $\rho(x,x')$.

Следующий важный аспект, вытекающий из анализа серии изображений, заключается в довольно большом разбросе цветовых характеристик как текста, так и фона. На изменчивость цвета фона и текста, оказывают влияние следующие факторы:

1) разброс параметров оптико-электронного тракта и параметров осветителей различных сканеров документов;
2) свечение голограммы (Рис.1, а(4) и б(4)) из-за изменения условий освещения или ориентации документа при сканировании;
3) изменение краски печати в различных партиях (редакциях) однотипных документов (Рис.1, б(1,2) и б(3,4));
4) зависимость цвета текста от цвета бланка, на котором он напечатан, обусловленная определенной степенью прозрачности краски, которой печатается текст.

Статистический анализ цветовых характеристик всего набора изображений позволяет сделать следующие выводы:

- в необработанных изображениях цвет текста сильно перемешан с цветом фона;
- факторы 2), 3) и 4) приводят к формированию кластеров цветовых характеристик текста с произвольным распределением точек внутри каждого кластера;
- фактор 1) приводит к размытию (расширению) кластеров в цветовом пространстве признаков;
- фактор 1) влияет глобально на всё изображение, т.е. почти синхронно смещает в цветовом пространстве не только цвет текста, но и цвет фона.

Для изображений текстовых областей (Рис. 1) данные выводы можно продемонстрировать следующей условной диаграммой. Для большей наглядности приведем цветовые характеристики текста в плоскости «цветовой тон - яркость» (Рис. 2).

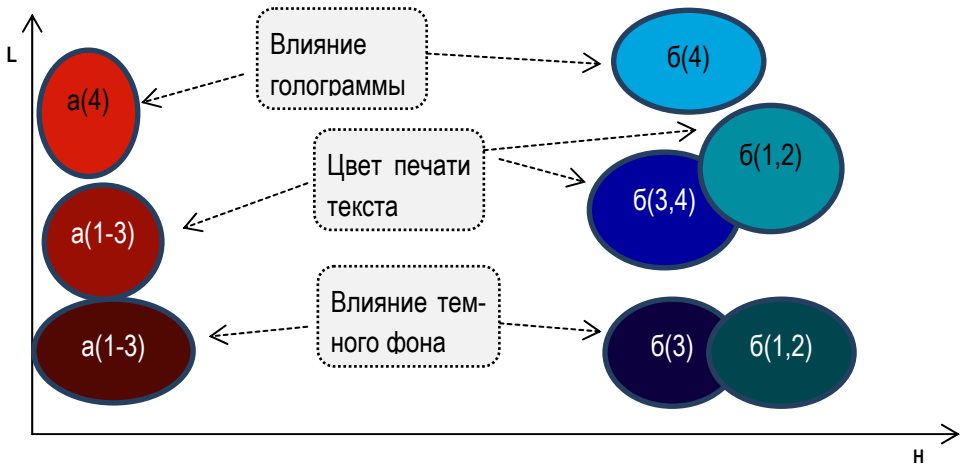

Рис. 2. Распределение цветовых характеристик текста в плоскости HL.

Очевидно, что перед фильтрацией необходимо попытаться устранить влияние фактора 1) с целью увеличения компактности кластеров цвета текста, а также стабилизации цветовых характеристик фона.

Таким образом, задача фильтрации цвета текста от цвета фона разделяется на две подзадачи: (I) компенсация аппаратных цветовых искажений изображения и (II) фильтрация – преобразование цвета из трехкомпонентного пространства в одномерное пространство полутонового изображения, где белым цветом будет обозначаться цвет фона, а черным – цвет текста.

Компенсация цветовых искажений

По условиям задачи для обработки предоставляются уже полученные с различных сканирующих устройств изображения. Поэтому нам недоступны такие процедуры нормализации изображений как калибровка видеотракта сканеров или проведение тестовых измерений их цветопередачи [Кровопусков, 2010]. Для разработки алгоритма компенсации цветовых искажений обратимся к методам представления цветовых характеристик, которые могут быть условно разделены на два класса: цветовые гистограммы [Гонсалес, 2006] и статистические модели представления цвета [Strieker, 1995].

Использование статистических моделей затруднительно в силу того, что

1) отсутствует нормальный закон распределения цветовых характеристик внутри каждого кластера (сложно для кластера определить «эталонное» значение цвета, к которому должны стягиваться цвета обрабатываемых изображений);
2) в то время как в обучающую выборку входят представители всех кластеров, то в отдельно взятом изображении присутствуют лишь некоторые из них, а иногда, что гораздо хуже, цвет фона может пересекаться с цветом имеющихся в выборке, но отсутствующих на изображении кластеров текста (в этом случае, без выполнения компенсации цвет фона будет принят за цвет текста);
3) для более качественного разделения текста и фона в обучающую выборку в качестве опорных могут включаться «граничные» цвета, множество которых на изображении, как правило, крайне мало.

Гистограммная форма описания изображения представляет собой информацию о распределении значений цветовых компонент в цветовых каналах. Для одного канала изображения I гистограмма является N-размерным вектором $H_I = (h_1, h_2, \ldots h_N)$, где h_i – отношение числа пикселей цвета i к общему числу пикселей в изображении I, а N – число квантов, например, градаций яркости.

В интересах компенсации аппаратных цветовых искажений мы применили процедуру пересчета значений цветовых компонент точек из обучающей выборки с учетом гистограмм обрабатываемого изображения и изображения, на котором данная точки была назначена. Перед обработкой (фильтрацией) изображения I для каждой m-ой точки из обучающей выборки значение k-ой цветовой компоненты f_k будет пересчитываться в соответствии с формулой:

$$f_k^{m\prime} = h_k^{I\,Min} + (f_k^m - h_k^{m\,Min}) \cdot \frac{(h_k^{I\,Max} - h_k^{I\,Min})}{(h_k^{m\,Max} - h_k^{m\,Min})}, \quad (1)$$

где $h_k^{I\,Min}$ и $h_k^{I\,Max}$ означают левую и правую границы гистограммы k-го канала изображения I, а $h_k^{m\,Min}$ и $h_k^{m\,Max}$ – границы гистограммы k-го канала того изображения, на котором была назначена m-я точка обучающей выборки.

Выражение (1) обеспечивает независимую для каждого канала подстройку значений опорных точек обучающей выборки под цветовые характеристики обрабатываемого изображения. Это очень важный момент, так как только в этом случае мы получаем возможность компенсировать различную цветопередачу сканирующих устройств. Процедуры типа автоконтрастирования для этих целей

непригодны, так как они выполняют пропорциональный для всех каналов пересчет значений. В результате, как правило, еще больше увеличивается расстояние между цветом опорных точек и цветом объекта, которому эти точки должны соответствовать (Рис. 3).

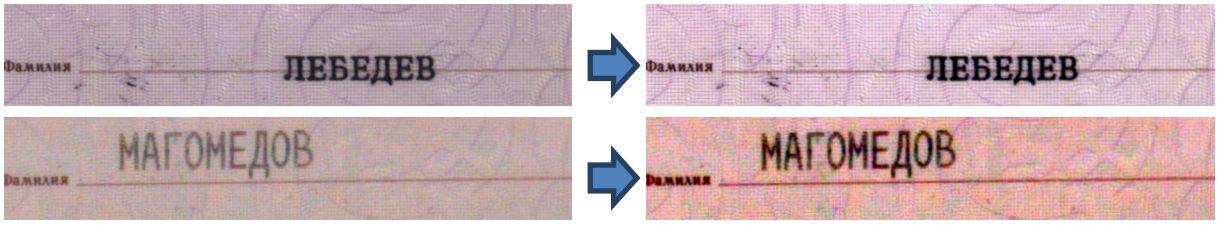

Рис. 3. Пример увеличения цветового рассогласования после применения процедуры автоконтрастирования.

Процедура фильтрации

Этот этап предполагает непосредственное преобразование цвета из трехкомпонентного пространства в одномерное пространство (полутоновое изображение). Значение яркости получаемого изображения должно быть пропорционально степени отличия цвета пикселя от цвета точек из множества $X^{m\ text}$ и степени сходства с цветом точек из множества $X^{n\ fond}$. Строго говоря, множество $X^{n\ fond}$ может быть и пустым. В большинстве случаев оказывается достаточным заполнить выборку только точками, задающими цвет текста.

Функция свертки трехкомпонентного значения цвета в значение яркости является не чем иным как функцией расстояния $\rho(x,x')$. При определении принципов работы функции расстояния мы исходили из следующих требований:

1) значение функции $\rho(x,x')$ должно зависеть от ближайшего соседа из выборок $X^{m\ text}$ и $X^{n\ fond}$;
2) функция $\rho(x,x')$ должна преобразовывать цветовое рассогласование $d(x,x')$ в значение яркости результирующего изображения.

Функция d(x,x') определяет ядро функции расстояния. В работе мы исследовали следующие ядра:

манхеттенское расстояние (норма L_1)

$$d(x,x') = \sum_{i=1}^{3} \left| x_i - x_i' \right|, \tag{2}$$

евклидово расстояние (норма L_2)

$$d(x,x') = \sum_{i=1}^{3} \left(x_i - x_i' \right)^2, \tag{3}$$

максимум модулей

$$d(x,x') = \max_i \left| x_i - x_i' \right|. \tag{4}$$

Преобразование отклика $d(x,x')$ в значение яркости осуществлялось посредством функции, имеющей следующий вид:

$$\rho(d) = \begin{cases} 0, \text{ если } d \le T/4 \\ \dfrac{255 \cdot (d - T/4)}{3 \cdot T/4}, \text{ если } T/4 < d < T \\ 255, \text{ если } d \ge T \end{cases}, \quad (5)$$

где T – параметр, определяющий толерантность или обобщающую способность функции расстояния.

В силу того, что значение функции d(x,x') вычисляется по расстоянию до ближайшего соседа, метод ближайшего соседа фактически реализует линейный классификатор, который определяет границу между классами как кусочно-линейную поверхность в цветовом пространстве признаков [Журавлев, 2006]. Вычисляемое в соответствии с (5) значение яркости можно интерпретировать как степень схожести цвета пикселя с заданным выборкой $X^{m\ text}$ цветом текста.

Функция расстояния (5) обеспечивает построение разделяющей поверхности даже при условии отсутствия в выборке точек фона, используя значение толерантности как максимально удаленную границу, отделяющую текст от фона. Если множество $X^{n\ fond}$ непустое и находится такое расстояние между x_i^{text} и x_j^{fond}, что $d(x_i^{text}, x_j^{fond}) < T$, то это означает, что j-ая опорная точка фона ограничивает область i-ой опорной точки текста. Для изображения б(4), Рис 1 сказанное можно проиллюстрировать следующей диаграммой (Рис. 4):

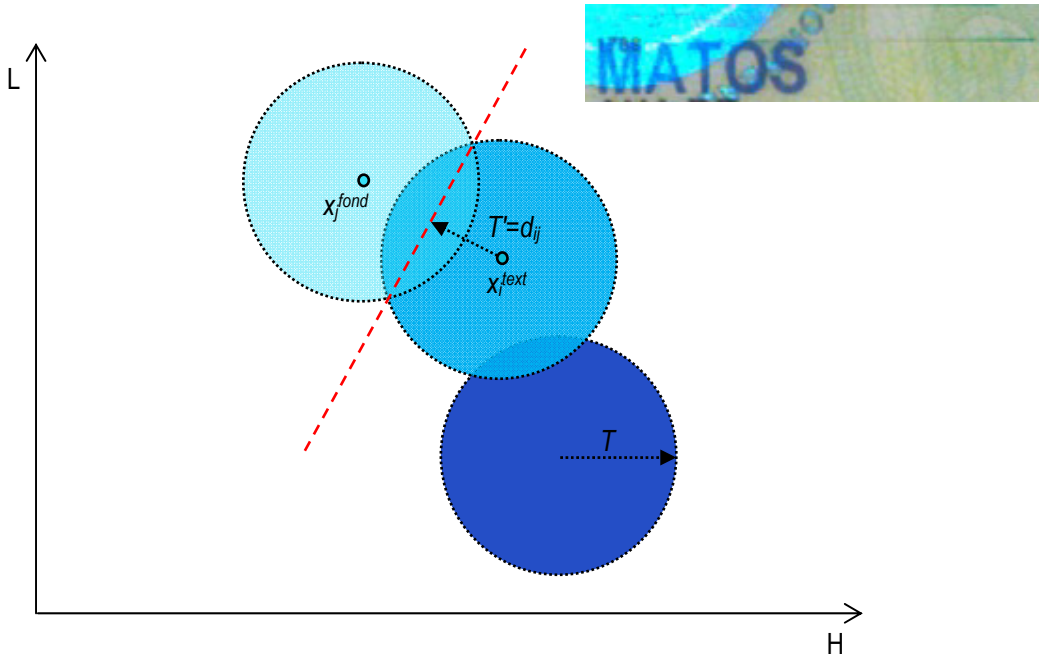

Рис. 4. Построение разделяющей поверхности в плоскости HL.

Из рисунка 4 и следует вывод, что нет необходимости задаваться опорными точками фона в случае, когда цвет фона отличается от цвета текста более чем на заданную величину толерантности T.

Выбор рабочей цветовой модели

Принятие решения о том, в какой цветовой модели осуществлять фильтрацию должно основываться на выполнении двух условий: во-первых, наилучшая цветовая модель должна увеличивать расстояние

между точками объекта и фона, и во-вторых, множество точек объекта должно располагаться как можно компактнее в цветовом пространстве выбранной модели.

Мы выработали следующие рекомендации:

1) если цвет текста или цвет фона можно отнести к категории черного, белого или серого, т.е. без явно выраженного цветового тона, то обработку следует осуществлять в RGB-пространстве;
2) если же цвет текста и фона можно назвать «цветными», то более эффективным будет использование перцепционной цветовой модели (HLS, HSV и др.).

Между HLS и HSV мы отдаем предпочтение первой, так как насыщенность в модели HLS всегда изменяется от полностью насыщенного цвета к эквивалентному серому цвету, в то время как в модели HSV при V=1 полностью насыщенный цвет переходит к белому.

Использование для обработки так называемых равноконтрастных цветовых моделей (Lab, LCH и др.) также возможно для ситуации 2), но будет оправдано только в тех случаях, когда необходимо алгоритмическую оценку цветоразличия привести в соответствие с человеческим цветовосприятием [Fairchild, 2005]. Платой за это станут повышенные вычислительные затраты. Использование равноконтрастных моделей имеет принципиальное значение, например, для кодирования изображений, как части графической системы, но не является обязательным в решаемой нами задаче.

В нашей работе мы использовали модифицированную HLS-модель: компоненту насыщенности S мы заменили на хрому C (Chroma) или ненормированную чистоту цвета, которая отражает степень приближения данного цвета к чистому спектральному цвету. В модели HLC хрома вычисляется по выражению

$$C = M - m,$$

где M = max(R,G,B) *и* m = min(R,G,B).

В отличие от насыщенности S значения хромы C не растянуты в диапазон [0,1]. Замена S на C преобразует цилиндрическое цветовое пространство HLS в форму трехмерного веретена. Хрома C максимальна только на средних уровнях яркости L, при увеличении или уменьшении яркости ощущение насыщенности падает (Рис. 5). Таким образом, при расчете цветового рассогласования, хрома будет вносить зависящий от уровня яркости вклад в итоговое расстояние.

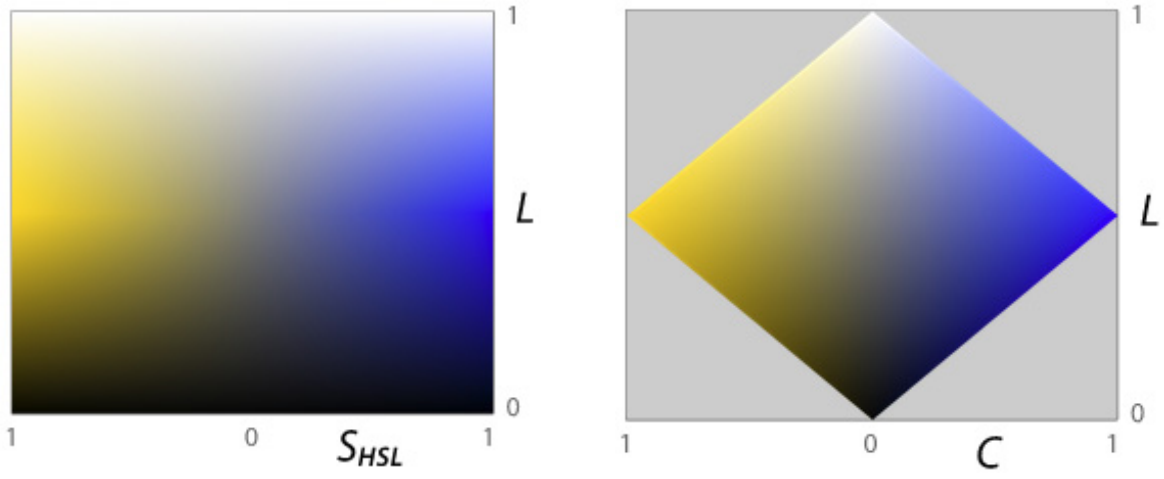

Рис. 5. Области определения насыщенности S в модели HLS и хромы C в модели HLC.

Результаты экспериментов

В качестве критерия эффективности процедуры фильтрации мы выбрали косвенный показатель – среднюю вероятность распознавания отфильтрованных изображений символов. Выбор данного критерия не противоречит здравому смыслу, так как обеспечивает оценку степени приближения формы символов к их эталонным изображениям. В тоже время такой способ оценки освобождает нас от весьма трудоемкого анализа правильности классификации каждого пикселя изображения.

Создание выборки опорных точек текста, а при необходимости и фона выполнялось вручную на нескольких изображениях с визуальной оценкой качества выполнения фильтрации. Здесь же выставлялось и значение толерантности T. Назначалось не более 8-ми точек в каждом подмножестве $X^{m\ text}$ и $X^{n\ fond}$. Затем процедура фильтрации запускалась на всем множестве изображений (от 500 до 1000) и подсчитывалась средняя вероятность распознавания символов отфильтрованного текста.

Настройка процедуры фильтрации, а также ее тестирование осуществлялись с применением различных ядер функции расстояния: dL1 (2), dL2 (3) и dmax (4). Кроме того, использовались два вида представления изображений: в цветовом пространстве RGB и в пространстве HLC. Примеры результатов фильтрации изображений (см. Рис. 1) представлены на рисунках 6 и 7.

Рис 6. Результат фильтрации текстового поля идентификационной карты штата Массачусетс (ID Mass).

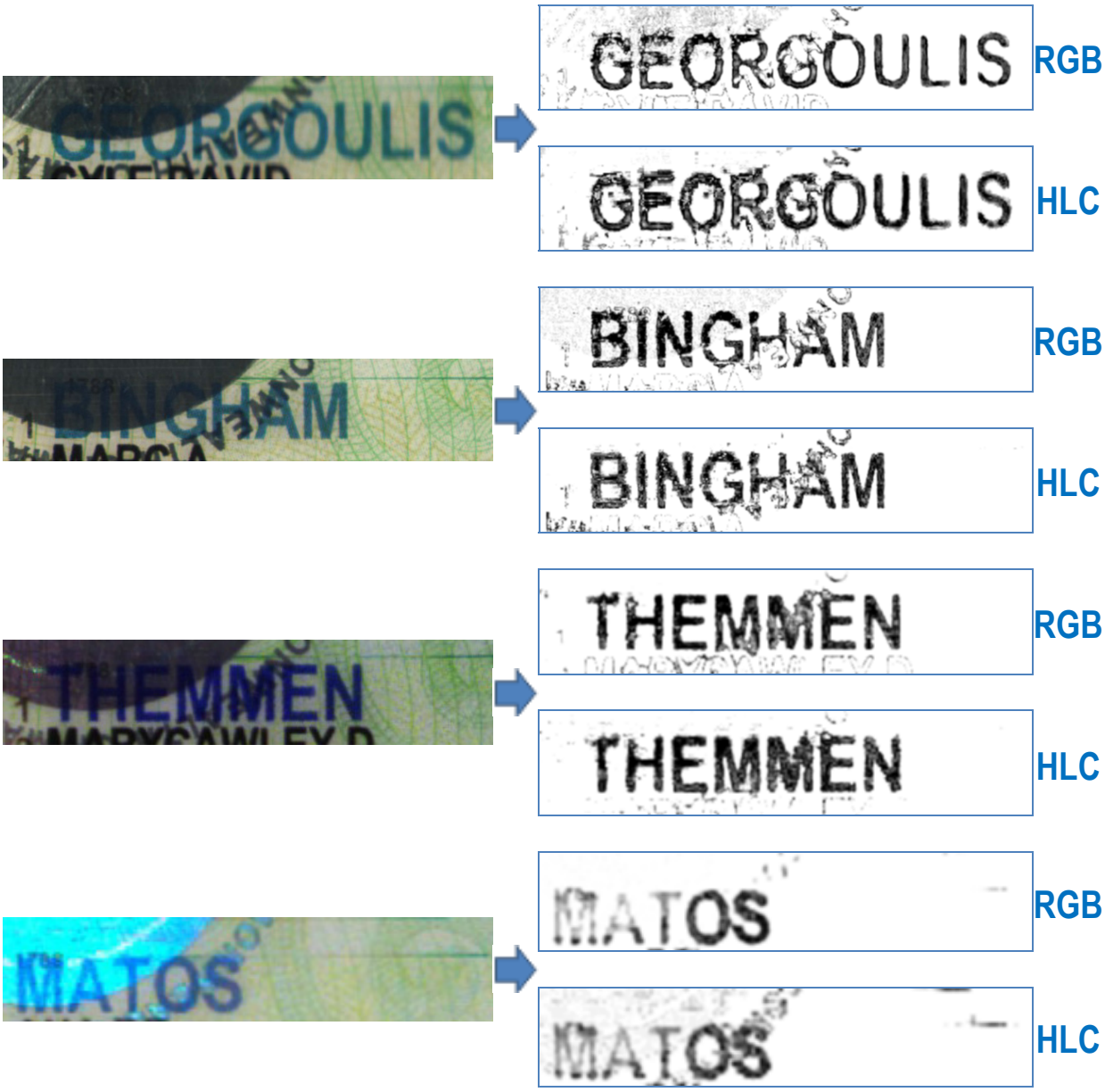

Рис 7. Результат фильтрации текстового поля водительского удостоверения штата Массачусетс (DL Mass).

Кроме указанных на рисунках 6 и 7 документах тестирование проводилось на текстовых полях еще трех видов документов. Примеры изображений текстовых полей приведены на Рис. 8.

Итоговые результаты тестирования процедуры фильтрации сведены в таблицы 1 и 2.

В последней строке таблицы под обозначением Def приводятся значения вероятностей распознавания изображений символов, обработанных с использованием таких инструментов, как выбор канала или смеси 2-х каналов, контрастирование и применение низкочастотных и высокочастотных фильтров.

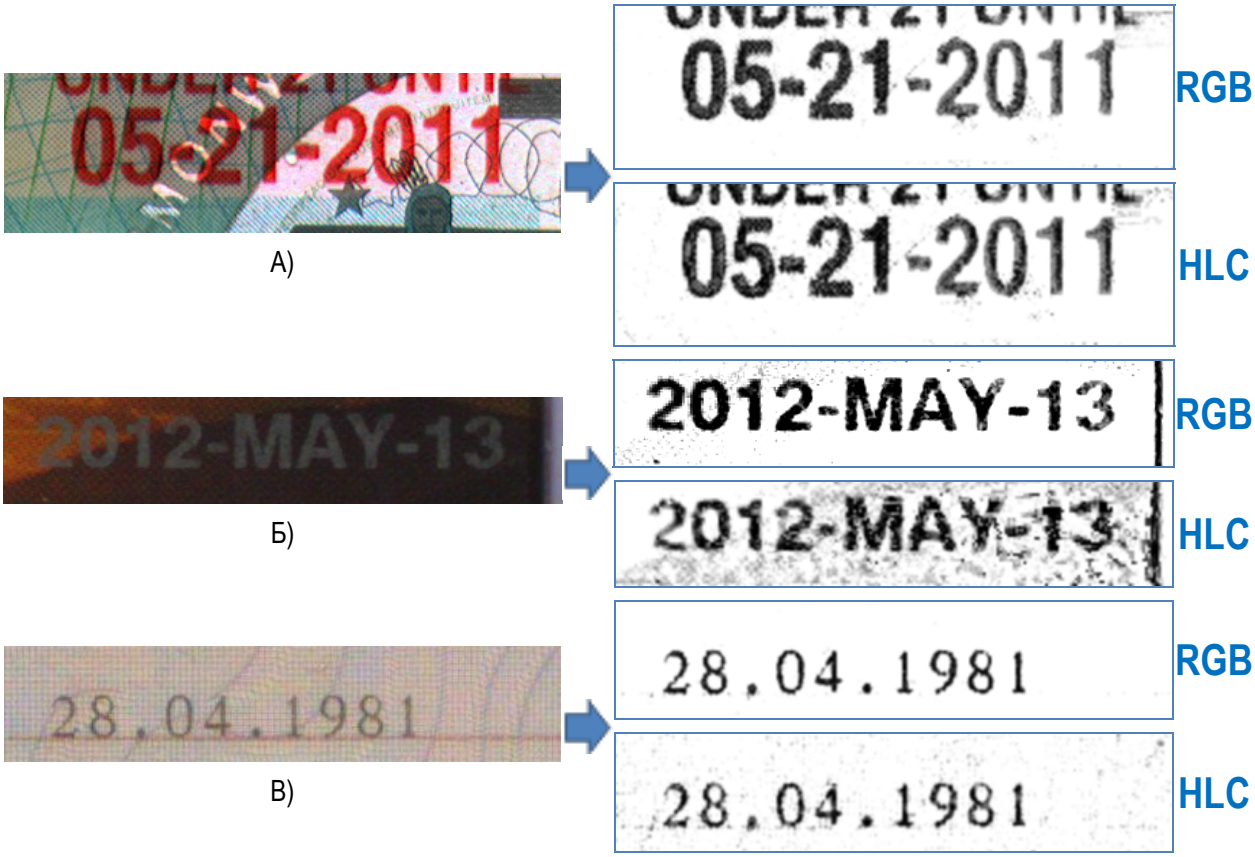

Рис 8. Изображения текстовых полей: а) водительское удостоверение для несовершеннолетних штата Массачусетс (DL Mass 18); б) водительское удостоверение штата Британская Колумбия, США (DL BC); в) Российский национальный паспорт (Pass RU).

Таблица 1. Средняя вероятность распознавания с применением RGB-модели.

Ядро функции сравнения	Документ 1 (ID Mass)	Документ 2 (DL Mass)	Документ 3 (DL Mass 18)	Документ 4 (DL BC)	Документ 5 (Pass RU)	Среднее по документам
d_{L1}	0,94	0,86	0,95	0,79	0,92	0,892
d_{L2}	0,93	0,89	0,96	0,82	0,92	0,904
d_{max}	0,90	0,88	0,89	0,71	0,90	0,856
Def	0,78	0,75	0,81	0,64	0,86	0,768

Таблица 2. Средняя вероятность распознавания с применением HLC-модели.

Ядро функции сравнения	Документ 1 (ID Mass)	Документ 2 (DL Mass)	Документ 3 (DL Mass 18)	Документ 4 (DL BC)	Документ 5 (Pass RU)	Среднее по документам
d_{L1}	0,90	0,85	0,97	0,70	0,89	0,862
d_{L2}	0,92	0,86	0,96	0,72	0,91	0,874
d_{max}	0,88	0,85	0,91	0,60	0,90	0,828
Def	0,75	0,71	0,85	0,61	0,83	0,75

Наилучшие результаты при обработке изображений продемонстрировала метрика L_2 (евклидово расстояние). Почти с такой же эффективностью работает и сравнение цвета по модулю рассогласования (метрика L_1). Худшие результаты, полученные на ядре d_{max} объясняются тем, что в сравнении участвует только одна цветовая компонента, причем на соседних пикселях нередко наблюдался эффект переключения процедуры сравнения с одного компонента на другой.

Как было изложено в предыдущем разделе, обработка изображения в цветовой модели HLC будет обоснована в тех случаях, когда цвет текста и цвет фона являются яркими и насыщенными. Только в этом случае (Документ 3 – DL Mass 18) эффективность фильтрации в цветовом пространстве HLC превысила эффективность фильтрации в пространстве RGB.

Заключение

В работе исследовано решение прикладной задачи выделения цветного текста на сложном цветном фоне. Для повышения эффективности решения задачи предложен способ компенсации цветовых искажений, вызванных сканирующей аппаратурой. На базе метода линейной классификации разработана процедура фильтрации фона от текста путем свертки цветного изображения в полутоновое изображение.

Для практического применения разработанной процедуры сделаны конкретные рекомендации по выбору рабочего цветового пространства и ядра функции расстояния.

Экспериментально получено подтверждение эффективности применения разработанного фильтра в задаче OCR. По сравнению с традиционно используемыми средствами предварительной обработки применение фильтра позволило повысить вероятность распознавания символов на 6-18%.

Благодарности

Статья публикуется при частичной поддержке проекта ITHEA XXI Международного научного общества ITHEA (www.ithea.org) и Ассоциации Создателей и Пользователей Интеллектуальных Систем ADUIS (www.aduis.com.ua).

Литература

[Fairchild, 2005] Mark D. Fairchild. Color Appearance Models, 2nd Edition. John Wiley & Sons, 2005. – 408 p.

[Strieker, 1995] Strieker M., Orengo M. Similarity of color images // Storage and Retrieval for Image and Video Databases (SPIE). – 1995. – P. 381-392.

[Wikipedia, 2013] http://en.wikipedia.org/wiki/Comparison_of_raster_graphics_editors.

[Айвазян, 1989] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989.

[Гонсалес, 2006] Гонсалес Р., Вудс Р. Цифровая Обработка Изображений. Техносфера.Москва, 2006 - 432с.

[Журавлев, 1978] Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания и классификации // Проблемы кибернетики. – М.: Наука, 1978, вып. 33. – С. 5-68.

[Журавлев, 2006] Журавлев Ю.И., Рязанов В.В., Сенько О.В. «Распознавание». Математические методы. Программная система. Практические применения. – М.: Фазис, 2006.

[Кровопусков, 2010] Кровопусков А.В., Крыловецкий А.А., Рындин А.А. Метод устранения цветового рассогласования в системе активного трехмерного сканирования. – Вестник ВГУ. сер.: Физика. Математика, 2010. №2, С. 247-251.

Информация об авторах

Телятников Роман *– к.т.н., научный руководитель разработками ПО, ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: raman.tsialiatnikau@regula.by*

Основные направления деятельности: распознавание образов, обработка изображений, нейрофизиология

Шумский Иван *– к.т.н., директор ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: ivan.shumsky@regula.by*

Основные направления деятельности: автоматизация анализа подлинности документов и банкнот: проектирование оборудования и программного обеспечения

Мамедов Ариф *– к.х.н., президент Regula Forensics, Inc., 1800 Alexander Bell Drive, Suite 400 Reston, VA 20191, USA, e-mail: arif.mamedov@regula.us*

Основные направления деятельности: подлинность документов и банкнот: маркетинговые исследования, автоматизация и проектирование

Протосавицкий Анатолий *– инженер-программист, ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: anatol.pratasavitski@regula.by*

Основные направления деятельности: обработка цветных изображений, элементы защиты и системы проверки подлинности документов

Матусевич Екатерина *– инженер-программист, ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: katsiaryna.harshkova@regula.by*

Основные направления деятельности: обработка цветных изображений, элементы защиты и системы проверки подлинности документов

Степанькова Екатерина *– инженер-программист, ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: KVayavodava@regula.by*

Основные направления деятельности: обработка изображений, статистический анализ, проверка подлинности документов