Синтез обучающей выборки на основе реальных данных в задачах распознавания изображений

Александр Жуковский
Московский физико-технический институт (государственный университет)
zhukovsky@phystech.edu

Наталья Тарасова
Московский физико-технический институт (государственный университет)
nilsonii.nt@gmail.com

Сергей Усилин
Московский физико-технический институт (государственный университет)
usilin.sergey@gmail.com

Дмитрий Николаев
Институт проблем передачи информации имени А. А. Харкевича РАН
dimonstr@iitp.ru

Аннотация

Работа посвящена проблемам построения обучающей выборки для алгоритмов обучения машин. Основное внимание уделено использованию особенностей физической модели формирования изображения для порождения релевантных синтетических примеров. Рассмотрены три модельные задачи распознавания изображений: детектирование логотипов кредитных карт в условиях изменчивой освещенности, детектирование лиц на фотографиях произвольно повернутых документов и распознавание печатного текста на изображениях низкого качества. К каждой задаче предложен оригинальный подход по синтезу обучающей выборки на основе реальных данных и приведено сравнение качества работы алгоритмов на исходном и расширенном обучающих наборах.

1. Введение

Недавние научные исследования привели к взрывному росту количества цифровых данных в таких областях как интернет, безопасность, финансы, медицина и многих других. Это придало огромный импульс изучению методов обработки больших объемов данных и извлечения из них знаний. Разработанные методы находят применение во многих областях современного мира - от повседневной жизни до крупномасштабных промышленных систем.

Особое место в списке активно исследуемых областей заслуживает машинное зрение. В круг рассматриваемых задач входят распознавание текста, поиск и идентификация объектов, отслеживание перемещения и многие другие.

Для решения большинства задач распознавания образов и поиска зависимостей между данными зачастую используются статистические алгоритмы. Одной из наиболее часто встречающихся на практике задач является задача классификации. Она заключается в определении класса, к которому принадлежит рассматриваемый объект на основе обучающего набора объектов, для которых принадлежности классам известны. Причем размер обучающего набора должен быть достаточно большим для того, чтобы алгоритм смог построить хорошую обобщающую гипотезу[1], и набор должен быть репрезентативен относительно генеральной совокупности объектов[2].

Большинство методов классификации предполагают, что априорные вероятности принадлежности рассматриваемого примера тому или иному классу и стоимости ошибки классификации равны. Однако в реальном мире такой случай является довольно редким. Одним из подходов к решению данной проблемы являются механизмы построения сбалансированной выборки на основе исходных данных [3]. Самыми простыми методами балансировки являются случайное сокращение элементов большего класса и случайное дублирование элементов меньшего класса [4,5]. Кроме того, для таких алгоритмов

обучения машин, как метод ближайших соседей и бустинг были разработаны модификации, синтезирующие искусственные вектора признаков на основе распределения реальных данных[6-8].

Для задачи распознавания рукописных цифр было предложено несколько методов расширения обучающего набора на базе набора MNIST[9-11]. По большей части они основаны на упругих искривлениях, моделирующих неконтролируемые колебания руки в процессе письма и аффинных преобразованиях, таких как сдвиг, масштабирование и поворот исходного изображения. Используемые в данной задаче в качестве классификатора нейронные сети, обученные на порожденном такими преобразованиями тренировочном наборе, показывают результаты существенно лучшие, чем при обучении на исходном наборе.

Однако в задачах машинного зрения проблема получения качественных обучающих наборов стоит намного острее. Большую роль в машинном зрении играет физическая модель формирования изображения.

В работе рассмотрены 3 задачи распознавания изображений, полученных с камеры. Первая посвящена поиску логотипов кредитных карт. Данная задача помимо очевидной трудности (подготовить довольно обширный для обучения набор логотипов настоящих кредитных карт достаточно сложно в связи с конфиденциальностью распознаваемого типа изображения) сопровождается дополнительными проблемами, привносимыми изменчивостью освещения. Чтобы улучшить качество детектирования в работе предложен способ расширения обучающей выборки с помощью гамма-коррекции (см. раздел 2).

Следующей задачей является детектирование лиц на изображениях документов. Несмотря на то, что в такой формулировке задача кажется уже решенной [12], при обработке изображений документов, полученных с камеры, качество работы детекторов сильно падает. Дело в том, что исходно прямоугольный документ под воздействием центрально-проективного преобразования камерой [14] может принять форму произвольного выпуклого четырехугольника (в зависимости от ракурса, под которым происходит съемка). При этом параметры проективного преобразования варьируются в довольно широком диапазоне для того, чтобы можно было подготовить обучающую выборку естественным образом. Однако по одной фотографии возможно с малыми потерями смоделировать снимок, сделанный с других ракурсов. Таким образом появляется возможность расширить обучающую выборку примерами лиц, которые могли бы быть на документах, снятых с разных ракурсов (см. раздел 3).

Последняя из рассмотренных задач заключается в распознавании текста на изображениях низкого качества. Дело в том, что повседневно используемые камеры, предназначенные для съемки малоразмерных изображений, ввиду малой разрешающей способности и слабой оптики, сильно искажают изображение. Это ведет к размытию символов на изображении и, следовательно, трудностям при его распознавании. Данный эффект хорошо моделируется гауссовским сглаживанием, примененным к высококачественным изображениям, например, полученным со сканера (см. раздел 4).

2. Моделирование освещения

Частным случаем распознавания изображений является детектирование объектов. Для решения этой задачи часто используют метод Виолы и Джонса [12]. Несмотря на то, что данный метод был изначально разработан для детектирования лиц, он также отлично подходит для поиска других объектов, обладающих достаточно жесткой геометрией. Метод Виолы и Джонса получил столь широкое применение благодаря робастности к незначительным геометрическим искажениям объекта и высокой скорости работы. Однако при всех достоинствах данный алгоритм обладает существенным недостатком: используемые в процессе вычисления Хаар-подобные признаки не инвариантны к различным условиям освещения. Поэтому часто используются различные модификации Хаар-подобных признаков, не зависящие так жестко от параметров освещения. Но иногда даже контрастно-устойчивые Хаар-подобные признаки не позволяют решить проблему переменчивой освещенности.

Рассмотрим задачу детектирования логотипа на изображении кредитной карты, полученной с помощью веб-камеры (см. рисунок 1). Задача возникает при распознавании кредитной карты: идентификация типа карты (который однозначно определяется найденным логотипом) позволяет настроить последующие шаги распознавания, существенно уменьшив тем самым общее количество ошибок. Обычно кредитная карта изготовлена из глянцевого материала. Следовательно, увеличение освещенности не приводит к пропорциональному увеличению яркости получаемого изображения, в результате чего контрастно-устойчивые Хаар-подобные признаки оказываются бесполезны.

Повысить качество детектирования логотипов в данном случае можно только лишь с помощью

а) б)

Рисунок 1. Пример работы обученного на синтезированной базе детектора в условиях а) низкой освещенности и б) засветки

Таблица 1. Качество детекторов логотипов, обученных на различных выборках

Обучающая выборка	Качество	
	Visa	MasterCard
Исходная (900 образцов VISA и 1319 образцов MasterCard)	85.0%	83.8%
Синтезированная (4500 образцов VISA и 3957 образцов MasterCard)	98.2%	97.6%

существенного расширения обучающей выборки, включающей образцы, полученные при разной освещенности. Однако из-за высокой конфиденциальности данного типа документа значительное расширение обучающей выборки естественным образом (собирая большое количество образцов) оказывается невозможным.

Для синтезирования новых образцов можно воспользоваться гамма-коррекцией [13] (коррекцией яркости цифрового изображения с помощью степенной функции):

$$I = I_0^{\gamma}, \qquad (1)$$

где I_0 - исходная яркость изображения, γ - степенной коэффициент, а I - результат коррекции. Использование различных коэффициентов γ позволяют моделировать различные параметры освещенности.

Описанный выше подход был опробован в процессе обучения детекторов логотипов двух основных в России видов кредитных карт (VISA и MasterCard). Первая серия детекторов обучалась на исходной выборке (состоящей из 900 изображений логотипов VISA и 1319 изображений логотипов MasterCard), тогда как вторая партия детекторов обучалась на синтезированной выборке. При синтезе обучающей выборки логотипов VISA использовалась гамма-коррекция с коэффициентами 0.5, 0.66, 1.0, 1.5, 2.0 (итого 4500 обучающих примера). Обучающая выборка для логотипов MasterCard была получена применением гамма-коррекции с коэффициентами 0.66, 1.0, 1.5 (итого 3957 примеров).

Оценка качества детекторов поводилась на тестовой выборке, состоящей из 3562 изображений кредитной карты MasterCard и 2721 изображений карты VISA. Качество работы детекторов проиллюстрировано в таблице 1.

Из таблицы 1 видно, что детектор, обученный на синтезированной выборке, значительно опережает по качеству детектор, обученный классическим образом.

Стоит также отметить, что дальнейшее расширение не приводило к улучшениям, а только наоборот понижало качество работы классификаторов. Дело в том, что при больших (меньших) коэффициентах изображения логотипов вырождались в практически полностью черные (белые) изображения, не позволяя тем самым классификатору выделить характерные особенности логотипов.

3. Моделирование перспективных искажений

Следующим применением, рассмотренным в данной работе, является поиск лиц на изображении. Задача возникла при распознавании сфотографированных водительских прав. Поиск прямоугольника документа сам по себе является сложной задачей ввиду произвольности положения документа и фона. В таком случае локализация лица, как объекта инвариантного, на исходном изображении и на проективно исправленном помогает уточнить правильность определения положения документа. В этой задаче, прежде всего, важно отсутствие ложных срабатываний детектора лица, что достигается увеличением порогового значения и, как следствие, уменьшением доли найденных объектов. В свою очередь, при подтверждении корректности определенного прямоугольника документа найденное на проективно исправленном изображении лицо может быть использовано для дальнейшей идентификации личности. Однако, повышение порогового значения, сильно уменьшает и так невысокую долю найденных объектов на проективно искаженных документах. Для улучшения работы классификатора предлагается обучать его на тренировочной выборке, смоделированной с учетом возможных проективных искажений.

При фотографировании происходит проецирование плоскости документа из 3-х мерного реального мира на 2-х мерную плоскость изображения. Данное преобразование наиболее точно воспроизводит преобразование центральной проекции:

$$x_i = \frac{a_i \cdot X + b_i \cdot Y + c_i \cdot Z + e_i}{d_X \cdot X + d_Y \cdot Y + d_Z \cdot Z + 1} \qquad (2)$$

Это преобразование несколько упрощается в случае плоских объектов. 8 параметров проективного преобразования плоскости в 3-х мерном пространстве задают трансформацию любого плоского документа в

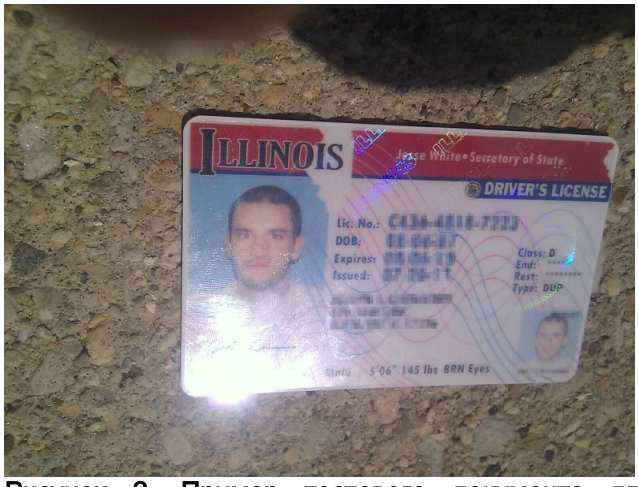

Рисунок 2. Пример тестового документа для распознавания лиц

Таблица 2. Качество детекторов лиц, обученных на различных выборках

Обучающая выборка	Кол-во сраб-ний	Качество
Исходная (218 образцов лиц)	87	44.39%
Синтезированная (2398 образцов лиц)	128	65.31%

его изображение, связывая координаты сцены с координатами изображения:

$$x_i = \frac{a_i \cdot X + b_i \cdot Y + c_i}{d_X \cdot X + d_Y \cdot Y + 1} \quad (3)$$

После такого преобразования изначально прямоугольный документ может принять форму произвольного выпуклого четырехугольника. Соответственно и лицо, находящееся на нем подвергается подобному искажению.

При построении качественной выборки для обучения поиску проективно искаженных лиц требуется большое количество примеров проекций фотографий на разные проективные плоскости. Для этого можно расширить базу, разметив множество документов, сфотографированных с различных ракурсов. Однако такой подход потребует огромного количества человеческих усилий и все равно не покроет всего пространства проективно искаженных лиц.

С другой стороны, имея точное изображение документа (взятое, например, со сканера) мы можем смоделировать его проекцию с произвольного ракурса. Для этого, зная координаты углов документа на исходном и конечном изображениях, мы можем вычислить матрицу проективного преобразования плоскости. Применив это преобразование к каждой точке исходного изображения, мы получим изображение документа, снятого с другого ракурса. Для суб-пиксельной точности при преобразовании используется билинейное сглаживание. Таким образом мы можем получить семейство снимков, соответствующее снимкам данного документа с разных ракурсов

В качестве базового алгоритма для обучения лиц использовался упомянутый выше метод Виолы и Джонса [12], основывающийся на построении классификатора лица как линейной композиции слабых классификаторов, регистрирующих контрастностные особенности лица. В качестве исходных данных использовалось 218 фотографий с заграничных паспортов. На отсканированных паспортах были размечены прямоугольники, обрамляющие каждое лицо. Для расширения обучающего набора координаты угловых точек документа подвергались случайному, небольшому сдвигу и вычислялись 8 параметров проективного преобразования плоскости, задающих трансформацию исходного изображения в конечное, по которым моделировалось новое изображение документа. Для каждого исходного документа независимо было сгенерирован 10 новых снимков.

Для оценки качества детектирования использовалась тестовая выборка, состоящая из 196 изображений водительских прав, сфотографированных в разных условиях (см. рисунок 2). Примеры для тестирования выбирались из условия максимального проективного искажения при сохранении распознаваемости лица. Пороговое значение классификатора завышалось для уменьшения количества ложных срабатываний. Качество работы детекторов проиллюстрировано в таблице 2.

Из таблицы видно, что добавление к исходной выборке синтезированных примеров существенно сказывается на результатах детекции – качество детектора, обученного на искусственных данных выше на 20%. С учетом того факта, что количество ложных срабатываний нулевое, итоговое качество можно считать удовлетворительным для поставленной задачи.

4. Моделирование неидеальной оптической системы

Задача распознавания печатного текста на изображениях занимает особое место в области компьютерного зрения. Исторически OCR-системы использовались для анализа изображений, полученных путем высококачественного сканирования. Однако, в настоящее время все большим спросом пользуются системы, способные распознавать текст на изображениях низкого качества (например, полученных с помощью веб-камеры или камеры мобильного телефона).

Одним из практических применений является распознавание текста на документах с целью идентификации личности владельца. Для

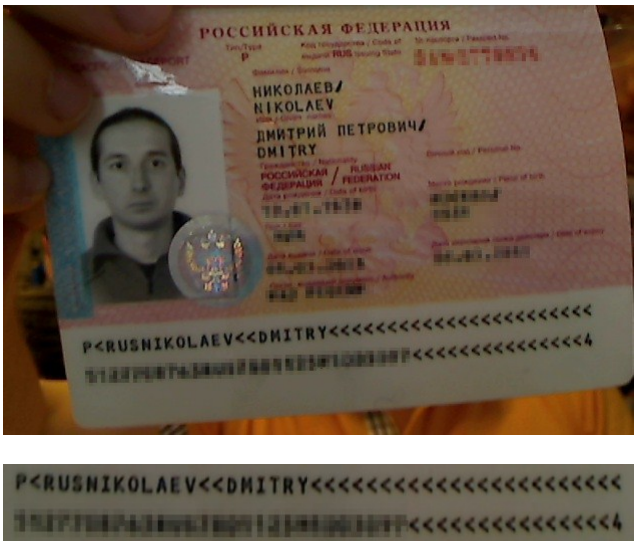

Рисунок 3. Машиночитаемая зона заграничного паспорта

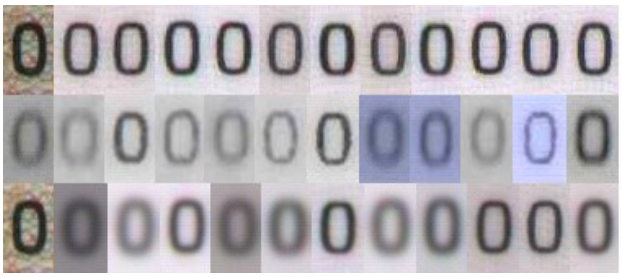

Рис. 4 Примеры изображений символа «0», полученных: Верхняя строка: со сканера, средняя: с веб-камеры, нижняя: путем синтеза.

автоматизации и ускорения данного процесса было принято соглашение о введении машиночитаемой зоны (см. рисунок 3), удовлетворяющей стандарту ICAO/ ИКАО 9303 [15]. Она представляет собой от одной до трех строк текста, состоящего из заглавных букв латинского алфавита, цифр и знака "<" для заполнения пустых мест, и содержит информацию как о владельце (фамилию, имя, дату рождения) так и о самом документе (тип документа, срок истечения действия). На распознавании машиночитаемой зоны документа построены основные системы безопасности и массового обслуживания. Большинство из них представляет собой комплекс, состоящий из регистрирующего устройства (камеры или сканера) и программного обеспечения, осуществляющего распознавание изображения.

В случае же, когда источником изображения документа является веб-камера или камера мобильного телефона, обладающая слабыми оптическими характеристиками, задача усложняется. Используемые камеры имеют индивидуальные технические характеристики, съемка производится в различных условиях освещения, документ по-разному расположен относительно камеры, кроме того, возможно движение камеры относительно документа. Все это ведет к тому, что изображения, полученные с веб-камер, обладают низким качеством: объекты могут быть нечеткими и низкоконтрастными, что сильно усложняет задачу распознавания.

Предварительная подготовка изображения к распознаванию машиночитаемой зоны, решаемая сторонними методами, заключается в детектировании документа, определении его положения на снимке, и последующем приведении к правильному виду, определении положения машиночитаемой зоны и ее сегментации на символы. Далее каждый из символов классифицируется заранее обученной нейронной сетью, результатом чего является распознанный текст. Заключительный этап распознавания заключается в исправлении ошибок классификации на основе известного формата машиночитаемой зоны.

Основная проблема состоит в получении качественной нейронной сети, способной правильно распознавать символы. Для универсальности классификатора, данные для обучения должны быть получены со всевозможных камер и учитывая фактические условия съемки, что, ввиду ограниченности ресурсов, не представляется возможным. Одним из вариантов получения необходимой тренировочной выборки является синтез новых обучающих примеров на основе реальных данных высокого качества моделированием искажений оптической системы веб-камеры.

Для получения новых обучающих наборов к эталонным изображениям применялись контрастирование [13] и гауссовское сглаживание [16] со случайными параметрами в допустимом диапазоне (сохраняющем изображение читаемым для человека) для достижения эффекта дефокусировки.

Изменение контраста для точки (x,y) i-го изображения производилось следующим образом:

$$I_i'(x,y) = \frac{I_i(x,y)}{\max\limits_{(x,y)}(I_i(x,y))} \cdot b_i, b_i \in R[0.5,1] \quad (4)$$

где $I_i(x,y) \in [0,1]$ – значение яркости точки (x,y) i-го изображения; b_i – фиксированное для каждого изображения случайное значение из отрезка [0.5,1].

Сглаживание проводилось при помощи свертки изображения с гауссовским окном. Дисперсия гауссовского фильтра определяется для каждого изображения из величины случайного значения b_i для этого изображения:

$$\sigma = c_1 \cdot b_i + c_2 \quad (5)$$

Где $c_1=5.2$ и $c_2 = -2.2$ – параметры, подобранные для сохранения изображения читаемым для человека. Размер окна свертки равен $a = 6 \cdot \sigma + 1$, но не меньше 3-х пикселей.

Исходная эталонная база обучающих примеров S_e для обучения состояла из высококачественных изображений, полученных со сканера. На каждый символ (кроме редко встречающихся) приходилось по 50 тренировочных изображений, отнормированных по высоте до 54 пикселей и в ширину от 30 до 34 пикселей. При помощи описанных операций из S_e были синтезированы новые выборки S_{100}, S_{200}, S_{400} и S_{800}, включающие в себя эталонный набор и содержащие в сумме 100, 200, 400 и 800 примеров для каждого символа соответственно (см. рисунок 4).

В качестве распознавателя использовался многослойный персептрон с 2 скрытыми слоями,

Таблица 3. Зависимость результата распознавания от обучающей выборки

Обучающая выборка	Качество
50 растров на символ (исходная)	71,22%
100 растров на символ (расширенная)	96,86%
200 растров на символ (расширенная)	98,43%
400 растров на символ (расширенная)	98,61%
800 растров на символ (расширенная)	98,44%
400 растров на символ (расширенная, с добавлением)	99.46%

состоящими из 256 нейронов[17]. Вектор признаков для каждого образца вычислялся на основе разреженного структурного тензора изображения символа.

Для определения качества распознавания обученных нейронных сетей использовалась тестовая выборка, состоящая из 19953 растров символов, «нарезанных» с изображений документов, отснятых на веб-камеры. После обучения и проверки качества распознавателей на тестовой выборке была отобрана лучшая сеть, и ее обучающая выборка была дополнена частью ошибочно распознанных примеров из тестовой базы(55 изображений, то есть менее 0.28% тестового набора). На полученной тренировочной выборке S_{400+} была обучена новая нейронная сеть. Качество ее работы тестировалось на тех же данных. Результаты приведены в таблице 3.

Из таблицы видно, что при достижении некоторого предела качество распознавания перестает улучшаться (можно заметить, что нейронная сеть, обученная на 800 растрах на символ, распознает несколько хуже, чем сеть, обученная на 400 растрах на символ), что объясняется переобучением. Тем не менее, расширение обучающей выборки до некоторых значений синтезированием новых примеров путем гауссовского размытия и контрастирования эталонных образцов приводит к повышению качества работы классификатора и возможно дальнейшее улучшение качества за счет дополнения выборки ошибочно классифицированными примерами.

5. Заключение

Для достижения высокого качества распознавания методами обучения машин необходимо, чтобы обучающая выборка была репрезентативна относительно генеральной совокупности, чего сложно добиться естественным образом (коллекционируя различные образцы). В данной работе показано, что при известной модели формирования изображения допустимо расширение обучающей выборки искусственным образом. В качестве примера были рассмотрены три задачи распознавания образов на изображениях, полученных с камер: поиск логотипов кредитных карт, поиск лиц на изображениях документов и распознавание печатного текста. Было показано, что расширение обучающих выборок за счет синтезированных примеров позволяет существенно улучшить качество распознавания.

Список литературы

[1] Russell S. J., Norvig P. - *Artificial Intelligence: A Modern Approach, 3rd edition*, Prentice Hall, 2010.

[2] Вапник В.Н., Червоненкис А.Я. *Теория распознавания образов*, М.: Наука, 1974.

[3] Haibo He, *Learning from Imbalanced Data,* IEEE Transactions on Knowledge and Data Engineering, 2009.

[4] R.C. Holte, L. Acker, B.W. Porter, *Concept Learning and the Problem of Small Disjuncts*, Proc. Int'l J. Conf. Artificial Intelligence, 1989.

[5] C. Drummond and R.C. Holte, *C4.5, Class Imbalance, and Cost-Sensitivity: Why Under Sampling Beats Over-Sampling,* Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II, 2003.

[6] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, *SMOTE: Synthetic Minority Over-Sampling Technique*, J. Artificial Intelligence Research, vol. 16, 2002.

[7] H. Guo, H.L. Viktor, Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost IM Approach, ACM SIGKDD Explorations Newsletter, vol. 6, 2004.

[8] H. Guo, H.L. Viktor, *Boosting with Data Generation: Improving the Classification of Hard to Learn Examples,* Proc. Int'l Conf. Innovations Applied Artificial Intelligence, 2004.

[9] Larry Yaeger, Richard Lyon , Brandyn Webb, *Effective Training of a Neural Network Character Classifier for Word Recognition*, NIPS, 1996.

[10] D.C. Ciresan, U. Meier, L. M. Gambardella, J. Schmidhuber, *Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition*, Neural Computation, Vol. 22, Num. 12, 2010.

[11] P. Y. Simard, D. Steinkraus, J. C. Platt, *Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis*, Int'l Conf. Document Analysis and Recognition, 2003.

[12] P. Viola and M. Jones, *Robust Real-time Object Detection*, *Int'l Journal of Computer Vision*, 2001.

[13] В. А. Сойфер, *Методы компьютерной обработки изображений*, ФИЗМАТЛИТ, 2003.

[14] D. A. Forsyth, J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, 2002.

[15] www.icao.int/Security/mrtd/

[16] L. G. Shapiro and G. C. Stockman, *Computer Vision*, Prentice Hall, 2001.

[17] S. Haykin, *Neural Networks - A Comprehensive Foundation*, 3nd Edition, Prentice Hall, 2008.