Синтез обучающей выборки в задаче распознавания текста в трехмерном пространстве

Д.П. Николаев, Д.В. Полевой, Н.А. Тарасова

Аннотация. В статье рассматривается проблема создания обучающей выборки для статистических методов обучения в контексте задачи распознавания моношрифтового текста. Исходные данные представляют собой изображения документов, содержащих машиносчитываемую зону, предъявленных видеокамере в произвольной ориентации. Геометрия сцены, освещение и источники изображений варьируются в широких пределах.

Ключевые слова: распознавание моношрифтового текста, синтез обучающей выборки, нейронные сети.

Введение

Каждый день человек сталкивается с огромным количеством всевозможной информации, значительную часть которой ему открывает зрение. Оно окружает его множеством образов, складывающихся в картину мира. Особое место среди этих образов занимают символы, с помощью которых осуществляются хранение и передача информации, что вызывает к ним особый интерес.

Ежедневно люди сталкиваются со множеством печатных документов, в которых им важна непосредственно информация, а не физический объект, в котором она содержится. Редактировать, хранить и распространять значительно проще электронные аналоги. Таким образом, процессы, связанные со множеством сопроводительных бумаг, можно было бы существенно упростить и ускорить за счет распознавания этих самых документов [1].

Одним из таких процессов является проверка пассажиров сотрудниками службы паспортного контроля в аэропортах. Для автоматизации данного процесса Международной организацией гражданской авиации (ИКАО) был разработан специальный стандарт документов, содержащих машиносчитываемую зону. Распространенные в настоящее время индустриальные системы распознавания машиносчитываемых зон, как правило, выполнены в виде специализированных сканеров, что существенно упрощает их работу, делая условия съемки более благоприятными и контролируемыми. Однако такой подход ведет к повышению стоимости системы в целом ввиду использования сложного оборудования. Это, в свою очередь, вызывает особый интерес к созданию систем, не имеющих особых требований ни к условиям съемки, ни к используемым регистрирующим устройствам. Выбор системы, удовлетворяющей данным требованиям, влечет за собой необходимость работы с изображениями, содержащими проективно искаженный, неравномерно освещенный документ на некотором фоне.

Как правило, при работе с подобным классом изображений в процессе распознавания

[1] Работа выполнена при финансовой поддержке РФФИ в рамках научных проектов № 13-01-12106, № 13-07-12172, № 13-07-12178

документа можно выделить этап, на котором отдельные символы, приведенные к некоторой нормализованной форме, подаются на вход классификатора. В качестве таких классификаторов часто используются нейронные сети и другие статистические обучаемые машины [2, 3]. При выборе данного подхода качество работы классификатора, главным образом, зависит от качества обучающей выборки: она должна быть максимально репрезентативной. Таким образом, на первый план выходит проблема получения такой выборки. Для решения данной проблемы необходимо учитывать особенности конкретной задачи распознавания.

В случае, когда речь идет о распознавании документов в неконтролируемых условиях, вариативность фона, взаимного расположения регистрирующего устройства и документа, условий освещения и многих других факторов оказывается чрезвычайно высокой. Отсутствие жестких требований к регистрирующему устройству лишь увеличивает число изменяемых параметров, поэтому задача составления репрезентативной обучающей выборки из реальных примеров является крайне сложной, требующей больших временных и человеческих затрат. Кроме того, ситуация осложняется необходимостью получения разрешения владельцев на использование их документов.

Одно из возможных решений описанной проблемы состоит в снижении вариативности параметров входных изображений, например, при помощи качественной бинаризации или же других методов обработки изображений. Однако в задачах, где есть сложный фон и всевозможные помехи в виде теней и бликов, подобная предобработка крайне сложна. В данной работе рассматривается другое решение проблемы получения репрезентативной обучающей выборки – синтез последней при помощи моделирования допустимых искажений.

В качестве основы для синтеза берутся реальные изображения высокого качества, затем к ним применяются преобразования, моделирующие искажения, вносимые системой в процессе ее работы на этапах, предшествующих классификации. Результирующие изображения образуют итоговую выборку. Таким образом, для получения обучающей выборки достаточного объема требуется сравнительно небольшое количество исходных данных и, следовательно, упрощается процесс их отбора и разметки.

Синтез обучающей выборки является широко используемым приемом, однако набор применяемых преобразований зависит от решаемой задачи. Так, в статьях [4, 5] для работы с бинаризованными рукописными символами применялись эластичные деформации, моделирующие особенности почерка и эффекты движения руки при письме. В работах [6, 7] предлагаются модели искажений, характерные для изображений, полученных со сканера и веб-камеры.

В данной работе предлагается итеративный подход к синтезу обучающей выборки на примере задачи распознавания машиносчитываемой зоны паспортов, предъявляемых для регистрации веб-камере. Далее будут описаны особенности рассматриваемой распознающей системы и итеративный подход к синтезу, а также приведены результаты экспериментов.

Особенности задачи

Рассматриваемая система предназначена для распознавания машиносчитываемой зоны на документах в реальном времени. Как регистрирующее устройство используется произвольная веб-камера, качество снимков которой достаточно для выполнения задачи распознавания человеком. Расстояние от документа до камеры, угол, образуемый его поверхностью с ее оптической осью, могут иметь произвольные значения, ограниченные требованием к читаемости текста на снимке человеком.

Машиносчитываемая зона удовлетворяет стандарту ICAO/ИКАО 9303 [8] и представляет собой от двух до трех строк текста, состоящего из заглавных букв латинского алфавита, цифр и знака “<” для заполнения пустых мест. Примеры документов, содержащих машиносчитываемую зону, представлены на Рис. 1.

Предварительная подготовка изображения, решаемая сторонними методами, заключается в детектировании документа, определении его положения на снимке и последующем приведении к прямоугольному виду, определении положения машиносчитываемой зоны и ее сегментации. Далее каждый из символов класси-

Рис.1. Примеры документа с машиносчитываемыми зонами: паспорт (слева) и идентификационная карта (справа)

фицируется заранее обученной нейронной сетью, результатом чего является распознанный текст.

При решении рассматриваемой задачи распознавания использовался итеративный подход. Сначала на небольшой обучающей выборке, составленной из высококачественных изображений, обучался классификатор. После этого проводился анализ нераспознанных элементов тестовой выборки, выделялись характерные особенности ошибочно классифицированных изображений, отсутствующие в обучающей базе, и подбирался набор преобразований, моделирующий эти особенности. Создавалась новая обучающая выборка, синтезированная из исходной с применением выбранных преобразований, после чего процедура повторялась вновь.

На первой итерации были обнаружены такие особенности, как дефокусировка и вариации контрастности, на второй - проективные искажения.

Применяемые преобразования

Для вариации контрастности изображения применялось контрастирование [9] со случайными параметрами в допустимом диапазоне (сохраняющем изображение читаемым для человека).

Изменение контраста для i точки j столбца растра производилось следующим образом:

$$I'(i,j) = \frac{I\ (i,j)}{\max\limits_{(i,j)}(I\ (i,j))} \cdot b\ ,\ b \in [0.5,1] \qquad (1)$$

где $I\ (i,j) \in [0,1]$ – значение яркости точки (i, j), b – фиксированное для каждого растра случайное значение, равномерно распределенное на отрезке [0.5, 1].

Для достижения эффекта дефокусировки использовалось гауссовское сглаживание [10], реализуемое при помощи свертки изображения с гауссовским окном. Среднеквадратичное отклонение определялось для каждого изображения из величины случайного значения b, равномерно распределенной на отрезке [0, 1] для этого изображения:

$$\sigma = c_1 \cdot b\ + c_2 \qquad (2)$$

где c_1 = 5.2 и c_2 = -2.2 – параметры, подобранные для сохранения изображения читаемым для человека. Размер окна свертки равен а $\max(6 \cdot \sigma + 1, 3)$, что связано со свойствами нормального распределения.

При синтезе примеров со сдвигами на исходном изображении $I \in \Omega_I = \{1,..,n\} \times \{1,..m\} \times \{1,..,k\}$ выделялась центральная область, содержащая символ $\Omega_{\text{int}} = \{h+1,..,n-h\} \times \{w+1,.m-w.\} \times \{1,..,k\}$ и "чистая" зона $\Omega_I \setminus \Omega_{\text{int}}$, представляющая собой фрагмент фона. В рассматриваемом случае $w = 2$, $h = 4$. Для получения величины сдвига dx, dy использовалось b – случайная величина, имеющая нормальное распределение со средним значением 0 и среднеквадратичным отклонением 1.

$$dx = [b \cdot (\Delta x + 1)/3] \qquad (3.1)$$

$$dy = [b \cdot (\Delta y + 1)/3] \qquad (3.2)$$

dx, dy подобраны таким образом, чтобы при сдвиге не происходило деформаций самого символа. Среднее фоновое значение $\hat{I}$ вычислялось следующим образом:

$$\hat{I} = \frac{\sum_{(i,j)\in\Omega_I \setminus \Omega_{\text{int}}} I(i,j)}{|\,\Omega_I \setminus \Omega_{\text{int}}\,|} \quad (4)$$

Синтезированный пример I' имел те же размеры, что и исходное изображение, и составлялся из фрагмента I и среднего фона в области сдвига:

$$I'(i,j) = \begin{cases} I(i-dn, j-dm), & \text{если } (i-dn, j-dm) \in \Omega_I \\ \hat{I}, & \text{если } (i-dn, j-dm) \notin \Omega_I \end{cases} \quad (5)$$

Растяжения производились с увеличением размера символа, так как в ожидаемых примерах он был не меньше, чем в идеальных. Коэффициенты *kx* и *ky* вычислялись по следующим формулам:

$$kn = (a*1/13 + 1); \quad (6.1)$$

$$km = (b*1/61 + 1); \quad (6.2)$$

где a и b – случайные величины с нормальным распределением, средним значением 0 и среднеквадратичным отклонением 1. Выбор коэффициентов производился с учетом допустимых в системе растяжений. Исходное изображение масштабировалось до размеров (n*kn)*(*m*km*), после чего выделялась его центральная часть, соответствующая исходным размерам n*m.

Структурный тензор

В качестве вектора признаков при обучении использовался центрированный структурный тензор. Структурным тензором полутонового растра I в точке p с окном w называется матрица 2*2:

$$S_w(p) = \begin{pmatrix} \sum_r w(r)(I_x(p-r))^2 & \sum_r w(r) I_x(p-r) I_y(p-r) \\ \sum_r w(r) I_x(p-r) I_y(p-r) & \sum_r w(r)(I_y(p-r))^2 \end{pmatrix}, \quad (7)$$

где суммирование ведется по r в некотором окне $r \in \{-m...m\} \times \{-m...m\}$, а w – весовая функция, задающая распределение весов суммирования в окне.

Эта матрица статистически устойчиво описывает двумерное распределение градиента изображения в окрестности точки p, а ее собственные значения соответствуют длинам полуосей эллипса инерции этого двумерного распределения.

Также возможно использование "центрированного" варианта структурного тензора.

$$S_w(p) = \begin{pmatrix} \overline{I_x^2}(p) - \overline{I_x(p)}^2 & \overline{I_x I_y}(p) - \overline{I_x(p)} \cdot \overline{I_y(p)} \\ \overline{I_x I_y}(p) - \overline{I_x(p)} \cdot \overline{I_y(p)} & \overline{I_y^2}(p) - \overline{I_y(p)}^2 \end{pmatrix}, \quad (7.1)$$

где верхней чертой обозначается усреднение в окне с весом w.

Данная конструкция описывает распределение границ по изображению символа.

Эксперименты по синтезу

Идеальная база Se, служащая основой для синтеза новых обучающих примеров, состояла из наборов по 50 изображений на символ алфавита, полученных со сканера, отнормированных до 54 пикселей по высоте и от 30 до 34 пикселей в ширину. При помощи описанных операций из Se были созданы новые выборки S_{100}, S_{200}, S_{400} и S_{800}, включающие в себя как эталонный, так и синтезированный наборы, содержащие в сумме 100, 200, 400 и 800 примеров для каждого символа соответственно. Растры, получаемые с изображений со сканера, веб-камеры и синтезированные из изображений со сканера, представлены на Рис. 2.

В качестве классификатора использовался многослойный персептрон с двумя скрытыми слоями, состоящими из 256 нейронов. Вектор признаков для каждого образца вычислялся на основе разреженного центрированного структурного тензора серого изображения символа, вычисляемого с дополнительным сглаживанием.

Тестовая выборка состояла из 19953 изображений символов, полученных с различных веб-камер. Распределение изображений по алфавиту в выборке не равномерно, существуют редкие символы, примерами которых могут служить Q, W, X, число которых не превышает 50. Данная особенность связана с трудностью набора достаточного числа различных доку-

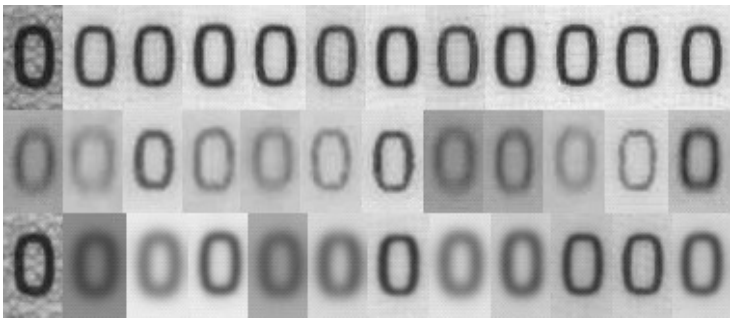

Рис. 2. Примеры изображений символа «0», полученных со сканера - верхняя строка, с веб-камеры – средняя строка, путем синтеза - нижняя строка

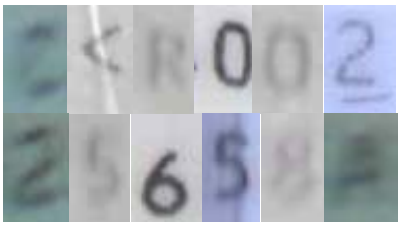

Рис. 3. Примеры ошибочно распознанных растров

ментов, значительной разницей частот вхождения различных букв в имена и фамилии.

Результаты тестирования обученных сетей представлены в Табл. 1. В столбцах расположены названия выборок, число изображений на символ, точность, которое показала сеть, обученная на данной выборке (процент правильно распознанных изображений к общему числу изображений). При увеличении объема обучающей выборки путем добавления синтезированных примеров до 400 изображений на символ качество распознавания возрастает, далее наблюдается падение. В нашем случае этот эффект может быть объяснен неточностью модели реальных данных, что может быть изменено с помощью параметров, используемых при синтезе. Из 800 изображений лишь 50 являются четкими и контрастными, то есть синтезированная выборка имеет более низкое качество, чем тестовая.

Несмотря на некоторую немонотонность уменьшения числа ошибок с ростом объема обучающей базы, результаты классификаторов, полученных на синтезированных наборах, значительно выше, чем на исходном.

Табл.1. Точности сетей, обученных на выборках S_e, S_{100}, S_{200}, S_{400} и S_{800}

Название	Объем выборки	Точность, %
$\mathbf{S_e}$	50	71,22
$\mathbf{S_{100}}$	100	96,86
$\mathbf{S_{200}}$	200	98,43
$\mathbf{S_{400}}$	400	**98,61**
$\mathbf{S_{800}}$	800	98,44

Стоит отметить, что при сравнении результатов искусственных выборок с эталонной речь идет о сравнении двух разных типов изображений. При этом стоит учитывать, что полученные классификаторы не настроены на работу с высококачественными примерами.

Для сравнения результатов распознавания классификаторов, обученных на естественной выборке с веб-камер и на синтезированных наборах, и оценки динамики качества их работы была проведена вторая серия экспериментов.

В процессе анализа ошибок сетей, обученных на предыдущем этапе, было обнаружено несколько особенностей неверно классифицированных символов: сдвиги символа относительно центра, растяжения, блики, мусор и искажения, изменяющие особенности шрифта. Примеры ошибочно распознанных изображений представлены на Рис. 3. Первые два типа ошибок представляют собой неточности системы при работе на предшествующих классификатору этапах. Так как подобные эффекты в той или иной степени присутствуют всегда, моделирующие их преобразования были добавлены в процедуру синтеза. При синтезе примеров сдвиги осуществлялись в трети, а растяжения – в половине случаев. Кроме того, было выявлено, что уменьшение размеров второго скрытого слоя в 2 раза не ведет к потере качества работы сети. Так как скорость обучения при этом возрастает, предпочтение было отдано новой архитектуре.

Были сформированы идеальные обучающие наборы по 5, 10 и 20 символов, содержащие только высококонтрастные изображения. Далее обучающие выборки расширялись описанными выше методами до объемов в 40, 80, 120, 160 и

Табл. 2. Точности сетей, обученных на естественных и синтезированных выборках

Объем	Естественные выборки		Синтезированные выборки		
	Со сканера	С веб-камер	Из 5 эталонов	Из 10 эталонов	Из 20 эталонов
40	74,24	94,39	94,59	95,22	96,17
80	75,98	95,39	97,34	96,00	98,33
120	78,71	97,19	98,45	98,31	98,45
160	83,66	98,48	98,21	98,87	98,86
200	88,16	98,48	98,42	98,78	**98,91**

200 изображений на символ с учетом удаления идеальных. Кроме того, было создано 2 дополнительных набора, состоящих из реальных примеров со сканера и веб-камеры тех же объемов.

В качестве классификатора использовался многослойный персептрон с двумя скрытыми слоями, размерами 256 и 128 нейронов. Точность работы обученных сетей оценивалась на той же тестовой выборке, что и в предыдущих экспериментах.

Точности сетей, обученных на исследуемых наборах, полученные на тестовой выборке, состоящей из изображений с веб-камер, представлены в Табл. 2; график зависимости десятичного логарифма процента ошибок классификации от объема выборки изображен на Рис. 4. Прослеживается общая тенденция роста точности при увеличении обучающей выборки. Результаты обученных на изображениях со сканера классификаторов существенно проигрывают. Этот эффект объясняется тем, что такие обучающие наборы плохо моделируют реальное распределение примеров. Быстрый рост качества с увеличением объема обучающих выборок со сканера связан с накоплением их “неидеальности”: сдвигов, разных фонов и контрастности.

Классификаторы, обученные на синтезированных данных, показывают результаты, близкие к обученным на реальных изображениях с веб-камер. Лучший результат 98.91 верно распознанных символов достигается при выборе в качестве тренировочных данных набора, синтезированного из 20 эталонных, то есть максимального числа идеальных примеров, при этом точность сети, обученной на 200 реальных растрах, полученных с веб-камер, заметно ниже – 98.42.

Таким образом, возможен синтез обучающей выборки, не содержащей чистых реальных данных, позволяющей добиться более высокого качества, чем при обучении на тех же объемах реальных данных.

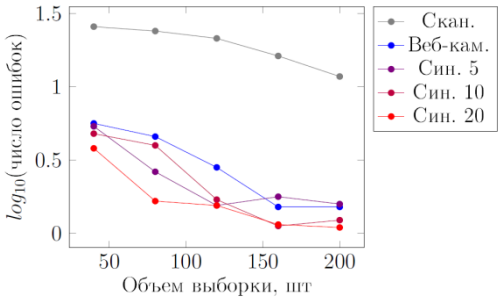

Рис.4. Зависимость десятичного логарифма числа ошибок от объема синтезированной выборки (шт.) при различных размерах эталонной базы изображений

Заключение

В задачах распознавания моношрифтового текста получение хорошей обучающей выборки часто является проблематичным. Это может быть связано с особенностями конкретной системы, осуществляющей распознавание. Для рассмотренной системы с вариативностью освещения, геометрии, фонов документов и регистрационных устройств расширение обучающей выборки при помощи синтеза примеров на основании реальных изображений дает положительный результат.

Применение дефокусировки гауссовского сглаживания и контрастирования при моделировании изображений веб-камер позволяет повысить качество распознавания при использовании малого объема исходных данных. Полезными могут быть сдвиги и растяжения, моделирующие неидеальность работы системы, предшествующей этапу распознавания символов.

Возможность повышения качества распознавания с помощью моделирования особенностей конкретной задачи является предпосылкой для дальнейшего анализа входных данных и поиска новых преобразований для синтеза. Разбор неверно классифицируемых символов позволяет выделить некоторые классы ошибок изображения, содержащие символ, повернутый на небольшой угол, и изображения с бликами. Таким образом, моделирование этих эффектов может дать дополнительное повышение качества распознавания.

Литература

1. Арлазаров В. Л., Емельянов Н. Е. Документооборот как информационная база накопления знаний // Труды ИСА РАН «Информационно-аналитические аспекты в задачах управления». М.: URSS, 2007. Т. 29 (ISBN 978–5–382–00486–0).
2. Haykin S. Neural Networks - A Comprehensive Foundation. - Prentice Hall, 1999, 842 p.
3. LeCun, Y., Jackel, L.D., Bottou, L., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Muller, U.A., Sackinger, E., Simard, P. and Vapnik, V. Learning Algorithms for Classification: A Comparison on Handwritten Digit Recognition - Neural Networks: The Statistical Mechanics Perspective, Oh, J. H., Kwon, C. and Cho, S. (Ed.), World Scientific, (1995), pp. 261-276.
4. Солдатова О.П Гаршин А.А. Применение сверточной нейронной сети для распознавания рукописных цифр // Компьютерная оптика , т. 34, 2010
5. Simard P.Y., Steinkraus D., Plat J.C. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis, Microsoft Research, One Microsoft Way, Redmond WA 98052
6. Li Y., Lopresti D., Nagy G., Tomkins A.- Validation of Image Defect Models for Optical Character Recognition - IEEE Transactions on Pattern Analysis and Machine Intelligence, February 1996, v. 18, pp. 99-108
7. Жуковский А.Е., Усилин С.А., Тарасова Н.А., Николаев Д.П. Синтез обучающей выборки на основе реальных данных в задачах распознавания изображений // Конференция ИТИС-2012, стр. 377-382.
8. Welcome to the ICAO Machine Readable Travel Documents Programme. URL: www.icao.int/Security/mrtd/
9. L.G. Shapiro and G. C. Stockman, Computer Vision, Prentice Hall, 2001.
10. В. А. Сойфер, Методы компьютерной обработки изображений, ФИЗМАТЛИТ, 2003.

Николаев Дмитрий Петрович. Заведующий сектором ИППИ РАН. Окончил МГУ им. М.В. Ломоносова в 2000 году. Кандидат физико-математических наук. Автор 116 печатных работ. Область научных интересов: быстрые алгоритмы обработки изображений. E-mail: dimonstr@iitp.ru

Полевой Дмитрий Валерьевич. Старший научный сотрудник ИСА РАН. Окончил. МФТИ (государственный университет) в 2004 году. Кандидат технических наук. Автор 10 печатных работ. Область научных интересов: методы искусственного интеллекта, оптическое распознавание, системы обработки документов. E-mail: dvpsun@gmail.com

Тарасова Наталья Андреевна. Программист ООО «Смарт Энджинс Рус». Окончила МФТИ (государственный университет) в 2013 году. Автор одной печатной работы. Область научных интересов: машинное обучение, нейронные сети. E-mail: nilsonii.nt@gmail.com