УДК 004.93

ЭТАЛОННЫЙ КЛАССИФИКАТОР НА БИНАРНЫХ МАТРИЦАХ В ЗАДАЧЕ РАСПОЗНАВАНИЯ РУКОПИСНЫХ СИМВОЛОВ КЫРГЫЗСКОГО АЛФАВИТА

Бийбосунов Б. И., Халмуратов Б. О.
КГУ им. И. Арабаева

Рассматривается задача распознавания рукописных символов кыргызского алфавита, специфика которой определяется тремя буквами — Ө, Ң и Ү, отличающимися от близких по начертанию О, Н и У лишь малым локальным элементом. Формализован эталонный классификатор, работающий на бинарных матрицах фиксированного размера. Показано, что для бинарных векторов решающее правило по минимуму евклидовой нормы совпадает с правилом по минимуму расстояния Хэмминга, что позволяет заменить арифметику побитовыми операциями и снизить трудоёмкость сравнения с одного эталона до нескольких машинных слов. Введено условие устойчивой различимости близких по начертанию пар и получена модельная оценка площади различающего элемента как функции разрешения сетки. Расчёт показывает, что при сетке 10×10 запас различимости пары «О — Ө» мал, а приемлемое отношение достигается начиная с сетки 28×28. Описана двухкаскадная схема, в которой посимвольное решение уточняется словарной проверкой собранного слова, и архитектура веб-прототипа на Python, Django и PostgreSQL. Работа носит методический характер: приводятся формальное описание метода, оценки сложности и различимости, а также программа измерительного эксперимента, необходимого для количественной проверки полученных оценок.

Ключевые слова: распознавание образов, рукописный текст, кыргызский алфавит, эталонный метод, расстояние Хэмминга, бинарные признаки, вычислительная сложность.

КЫРГЫЗ АЛФАВИТИНИН КОЛ ЖАЗМА СИМВОЛДОРУН ТААНУУ МАСЕЛЕСИНДЕ БИНАРДЫК МАТРИЦАЛАРДАГЫ ЭТАЛОНДУК КЛАССИФИКАТОР

Бийбосунов Б. И., Халмуратов Б. О.
И. Арабаев атындагы Кыргыз мамлекеттик университети

Кыргыз алфавитинин кол жазма символдорун таануу маселеси каралат; анын өзгөчөлүгү Ө, Ң жана Ү тамгалары менен аныкталат, алар жазылышы жагынан жакын О, Н жана У тамгаларынан кичинекей жергиликтүү элемент менен гана айырмаланат. Белгиленген өлчөмдөгү бинардык матрицаларда иштеген эталондук классификатор формалдаштырылган. Бинардык векторлор үчүн эң аз евклиддик норма боюнча чечим эрежеси Хэмминг аралыгынын минимуму боюнча эреже менен дал келери көрсөтүлгөн; бул арифметиканы биттик операциялар менен алмаштырууга жана бир эталон менен салыштыруунун эмгек сыйымдуулугун бир нече машина сөзүнө чейин төмөндөтүүгө мүмкүндүк берет. Жазылышы жакын жуптарды туруктуу ажыратуу шарты киргизилип, айырмалоочу элементтин аянты тордун чечилиши боюнча функция катары бааланган. Эсептөө көрсөткөндөй, 10×10 тордо «О — Ө» жубунун ажыратылыш запасы аз, ал эми алгылыктуу катыш 28×28 тордон баштап жетишилет. Символ боюнча чечим чогултулган сөздү сөздүк боюнча текшерүү менен тактала турган эки каскаддуу схема жана Python, Django, PostgreSQL негизиндеги веб-прототиптин архитектурасы сүрөттөлгөн. Иш методикалык мүнөзгө ээ: методдун формалдык баяндамасы, татаалдык жана ажыратылыш баалоолору, ошондой эле алынган баалоолорду сандык текшерүү үчүн зарыл болгон өлчөө экспериментинин программасы келтирилет.

Баштапкы сөздөр: образдарды таануу, кол жазма текст, кыргыз алфавити, эталондук метод, Хэмминг аралыгы, бинардык белгилер, эсептөө татаалдыгы.

A TEMPLATE CLASSIFIER ON BINARY MATRICES FOR HANDWRITTEN CHARACTER RECOGNITION OF THE KYRGYZ ALPHABET

Biibosunov B. I., Khalmuratov B. O.
Kyrgyz State University named after I. Arabaev

The paper addresses handwritten character recognition for the Kyrgyz alphabet, whose specific difficulty is caused by three letters — Ө, Ң and Ү — that differ from the graphically close О, Н and У only by a small local element. A template classifier operating on binary matrices of fixed size is formalised. It is shown that for binary vectors the decision rule based on the minimum Euclidean norm coincides with the rule based on the minimum Hamming distance, which allows arithmetic to be replaced by bitwise operations and reduces the cost of a single template comparison to a few machine words. A condition of stable discriminability for graphically close pairs is introduced, and a model estimate of the area of the distinguishing element is obtained as a function of grid resolution. The calculation shows that on a 10×10 grid the discriminability margin of the pair О — Ө is small, while an acceptable ratio is reached starting from a 28×28 grid. A two-stage scheme is described in which the per-character decision is refined by a dictionary check of the assembled

word, together with the architecture of a web prototype based on Python, Django and PostgreSQL. The work is methodological: it provides a formal description of the method, complexity and discriminability estimates, and a programme of the measurement experiment required to verify these estimates quantitatively.

Keywords: pattern recognition, handwritten text, Kyrgyz alphabet, template method, Hamming distance, binary features, computational complexity.

Введение

Автоматическое чтение рукописного текста остаётся одной из немногих задач распознавания образов, где переход к цифровым документам всё ещё упирается в ручной ввод. В Кыргызской Республике эта задача имеет дополнительную сторону: массово доступные системы оптического распознавания настроены на латиницу и на русскую кириллицу, тогда как кыргызский алфавит содержит три буквы, отсутствующие в русском, — Ө, Ң и Ү. Все три образованы добавлением малого элемента к уже существующей букве: перекладина внутри овала О даёт Ө, хвостик у правой стойки Н даёт Ң, петля в основании У даёт Ү. Для системы, обученной на русской кириллице, такой элемент оказывается ниже порога значимости, и Ө читается как О, Ң — как Н.

Классические подходы к распознаванию образов подробно изложены в работах [1—5]. Развитие признаковых описаний привело к дескрипторам, устойчивым к сдвигу и освещённости, — гистограммам направленных градиентов [6] и локальным бинарным шаблонам [7]. Свёрточные сети [8] сегодня дают наилучшие результаты на массовых наборах рукописных символов, а практика их применения к документам сложилась ещё в начале двухтысячных [9]. Для кириллицы существует размеченный набор Cyrillic-MNIST [10], однако он покрывает русский алфавит и не содержит кыргызских букв.

У нейросетевого пути есть цена: он требует размеченной выборки достаточного объёма, вычислительных ресурсов для обучения и не даёт простого объяснения отдельного решения. Для учебных и вспомогательных задач, где алфавит невелик, а вычислительные ресурсы ограничены, сохраняет смысл эталонный метод — прямое сравнение изображения символа с набором заранее сохранённых образцов. Его достоинства известны: отсутствие этапа обучения, прозрачность решения и возможность пополнять базу эталонов, не переобучая модель [2, 3].

Цель настоящей работы — дать формальное описание эталонного классификатора на бинарных матрицах применительно к кыргызскому

алфавиту, оценить его вычислительную трудоёмкость и, главное, установить, при каком разрешении растровой сетки малый различающий элемент букв Ө, Ң и Ү остаётся устойчиво различимым. Последний вопрос в литературе по эталонным методам обычно не ставится, между тем именно он определяет применимость подхода к кыргызскому письму.

Формальная постановка

Пусть изображение символа приведено к квадратной растровой сетке k × k и бинаризовано. Результат представляется вектором $x \in \{0, 1\}^n$, где $n = k^2$, а единица соответствует закрашенной клетке. База эталонов содержит M образцов $e_1, \ldots, e_m$, каждый из которых снабжён меткой класса $c(e_i) \in \{1, \ldots, K\}$, где K — мощность алфавита. Один класс представлен несколькими эталонами, отражающими варианты начертания.

Решающее правило состоит в выборе эталона, ближайшего к предъявленному вектору в смысле евклидовой нормы разности:

$$\hat{c}(x) = c(e_j),\ j = \arg\min_i \|x - e_i\|_2,\ i = 1, \ldots, M. \qquad (1)$$

Правило (1) не требует обучения: настройка классификатора сводится к наполнению базы эталонов. Именно это свойство делает его удобным для алфавита, который дополняется новыми начертаниями по мере накопления образцов почерка.

Эквивалентность евклидовой и хэмминговой метрик

Для бинарных векторов правило (1) допускает существенное упрощение. Поскольку каждая координата принимает лишь значения 0 и 1, разность $x_i - e_i$ принимает значения −1, 0 или 1, и её квадрат равен модулю:

$$\|x - e\|_2^2 = \sum_i (x_i - e_i)^2 = \sum_i |x_i - e_i| = d_h(x, e). \qquad (2)$$

Правая часть (2) есть расстояние Хэмминга, то есть число несовпадающих разрядов. Поскольку квадратный корень — монотонно возрастающая функция, порядок эталонов по евклидовой норме и по расстоянию Хэмминга совпадает, а значит правило (1) эквивалентно правилу

$$j = \arg\min_i d_h(x, e_i). \qquad (3)$$

Практическое следствие состоит в том, что вычисление корней и умножений не требуется вовсе. Расстояние Хэмминга между двумя битовыми строками равно числу единиц в их поразрядной сумме по модулю два, то есть вычисляется как популяционный счёт результата операции «исключающее ИЛИ» [11]. Современные процессоры выполняют такой счёт одной инструкцией над машинным словом, поэтому сравнение с одним эталоном сводится к $[n / w]$ парам операций,

где w — разрядность слова. Соответствующие оценки приведены в таблице 1.

Таблица 1. — Трудоёмкость сравнения с одним эталоном при разной сетке (w = 64)

Сетка k × k	Размерность n	Операций при поэлементном сравнении	Машинных слов при битовом представлении
10 × 10	100	100	2
16 × 16	256	256	4
20 × 20	400	400	7
28 × 28	784	784	13

Полный перебор базы требует M сравнений, то есть трудоёмкость распознавания одного символа составляет O(M·⌈n / w⌉) операций и линейно растёт с объёмом базы. Это и есть главное ограничение эталонного подхода: точность повышается наполнением базы, а цена решения растёт вместе с ней. Для алфавита из K классов при равномерном представлении классов M = K·r, где r — число вариантов начертания на класс, и трудоёмкость линейна по r.

Предел различимости близких по начертанию пар

Специфика кыргызского алфавита ставит перед эталонным методом вопрос, который для русской кириллицы не возникает. Буквы каждой из пар О — Ө, Н — Ң, У — Ү совпадают почти полностью и различаются только малым элементом. Обозначим через S число клеток сетки, занимаемых этим элементом. Тогда для идеально написанных образцов расстояние между эталонами пары равно $d_h = S$.

Реальный почерк отклоняется от эталона: линия смещается, толщина штриха меняется, овал вытягивается. Обозначим через δ типичное число клеток, меняющих своё значение из-за такой вариативности. Символ будет устойчиво отнесён к своему классу, а не к парному, если различающий элемент вносит в расстояние больший вклад, чем случайное отклонение почерка:

$$S > 2\delta. \qquad (4)$$

Множитель два учитывает, что отклонение действует в обе стороны — приближая образец к чужому эталону и одновременно удаляя от своего. Условие (4) даёт прямой критерий выбора разрешения сетки.

Оценим обе величины как функции k. Примем, что символ занимает по ширине около 0,6k клеток, а толщина штриха составляет max(1; 0,06k) клеток. Перекладина буквы Ө пересекает овал по ширине, поэтому её

площадь S ≈ 0,6k · max(1; 0,06k). Вариативность δ пропорциональна длине контура: при смещении линии на одну клетку меняется число клеток порядка длины контура, которая для овала О составляет около π · 0,6k. Результаты расчёта сведены в таблицу 2.

Таблица 2 — Модельная оценка различимости пары «О — Ѳ» при разной сетке

Сетка k × k	**Длина перекладины, клеток**	**Толщина штриха, клеток**	**Площадь элемента S, клеток**	**Длина контура овала, клеток**	**Отношение S к длине контура**
10 × 10	6,0	1,0	6	19	0,32
16 × 16	9,6	1,0	10	30	0,32
20 × 20	12,0	1,2	14	38	0,38
28 × 28	16,8	1,7	28	53	0,53

Расчёт показывает, что до сетки 16 × 16 отношение площади различающего элемента к длине контура остаётся постоянным: и элемент, и вариативность растут линейно по k, поэтому увеличение разрешения само по себе преимущества не даёт. Положение меняется, когда толщина штриха начинает превышать одну клетку: с этого момента площадь перекладины растёт квадратично, а длина контура — по-прежнему линейно, и отношение улучшается. При сетке 28 × 28 оно составляет 0,53 против 0,32 при сетке 10 × 10, то есть запас различимости возрастает более чем в полтора раза.

Отсюда следуют два вывода. Первый: сетка 10 × 10, привлекательная своей вычислительной дешевизной, для пар О — Ѳ, Н — Ң и У — Ү находится вблизи границы применимости, и ошибки на этих парах ожидаемо составят основную долю всех ошибок. Второй: если увеличивать разрешение нежелательно, условие (4) можно выполнить иначе — дополнив бинарный вектор несколькими целевыми признаками, отвечающими именно за наличие различающего элемента, например признаком пересечения внутренней области овала горизонтальным штрихом. Такой признак вносит в расстояние постоянный вклад независимо от разрешения и потому эффективнее наращивания сетки.

Словарная постобработка

Посимвольное решение по правилу (3) не использует того, что символы образуют слово. Второй каскад схемы устраняет этот пробел. Последовательность распознанных символов сопоставляется со словарём целевого языка; если слово найдено, оно принимается, если нет — рассматриваются варианты, полученные заменой символов на ближайшие по расстоянию Хэмминга альтернативы из правила (3).

Именно на парах О — Ө, Н — Ң и У — Ү словарная проверка даёт наибольший выигрыш, поскольку эти пары почти никогда не порождают двух осмысленных слов одновременно: замена Ө на О в кыргызском слове, как правило, приводит к несуществующей словоформе. Тем самым второй каскад компенсирует ровно ту слабость первого, которая выявлена условием (4). Ограничение схемы очевидно: она применима к словам, входящим в словарь, и бесполезна для имён собственных и числовых последовательностей.

Архитектура прототипа

Изложенный метод реализован в виде веб-приложения. Серверная часть построена на языке Python с использованием каркаса Django, база эталонов размещена в системе управления базами данных PostgreSQL. Клиентская часть содержит область рисования на элементе canvas, куда пользователь вводит символ или слово указателем, и кнопку запуска распознавания. Введённое изображение приводится к бинарной матрице, разделяется на отдельные символы и передаётся на сервер, где выполняется правило (3) и словарная проверка. Результат возвращается с цветовой индикацией: найденное в словаре слово выделяется одним цветом, не найденное — другим, что даёт пользователю немедленный сигнал о вероятной ошибке распознавания.

Хранение эталонов в реляционной базе, а не в файлах, выбрано из соображений пополняемости: добавление нового варианта начертания сводится к вставке записи и не требует перестроения системы. Это же свойство позволяет накапливать образцы почерка в ходе эксплуатации и постепенно наращивать r — число вариантов на класс.

Обсуждение и программа измерений

Полученные результаты носят аналитический характер. Соотношение (2) доказано и верно для любых бинарных векторов; оценки трудоёмкости в таблице 1 следуют из него непосредственно. Оценки в таблице 2 являются модельными: они опираются на принятые пропорции символа и толщины штриха и указывают порядок величины и характер зависимости, но не заменяют измерения на реальных образцах почерка.

Никаких экспериментальных значений точности распознавания в настоящей работе не приводится, и утверждать превосходство описанной схемы над существующими системами на основании изложенного нельзя.

Для количественной проверки предлагается следующая программа. Формируется размеченный набор образцов всех K классов кыргызского алфавита, собранный не менее чем от тридцати пишущих, с фиксацией числа образцов на класс и разделением на обучающую и контрольную части без пересечения по пишущим. На этом наборе измеряются: доля правильных ответов в целом и по классам; матрица ошибок с выделением подматрицы пар О — Ө, Н — Ң, У — Ү; зависимость доли ошибок от разрешения сетки k при $k \in \{10, 16, 20, 28\}$, что позволит проверить расчёт таблицы 2; зависимость доли ошибок и времени распознавания от числа эталонов на класс r; прирост доли правильных ответов, вносимый словарным каскадом. В качестве базы сравнения целесообразно взять общедоступную систему оптического распознавания и компактную свёрточную сеть, обученную на том же наборе.

Заключение

Формализован эталонный классификатор на бинарных матрицах для распознавания рукописных символов кыргызского алфавита. Показано, что на бинарных векторах решающее правило по минимуму евклидовой нормы тождественно правилу по минимуму расстояния Хэмминга, вследствие чего сравнение с эталоном выполняется побитовыми операциями и требует $\lceil n / w \rceil$ машинных слов вместо n арифметических действий; при сетке 10 × 10 и разрядности 64 это два слова на эталон.

Введено условие устойчивой различимости близких по начертанию пар, связывающее площадь различающего элемента с вариативностью почерка. Модельный расчёт показал, что при сетках до 16 × 16 увеличение разрешения не улучшает различимость пары О — Ө, поскольку обе величины растут линейно, а выигрыш появляется лишь тогда, когда толщина штриха превышает одну клетку сетки. Отсюда получен практический вывод: для кыргызского алфавита разумнее не наращивать разрешение, а дополнить описание целевыми признаками наличия различающего элемента.

Дальнейшая работа состоит в проведении описанного измерительного эксперимента, построении набора образцов кыргызского рукописного письма, отсутствующего сегодня в открытом доступе, и в сопоставлении эталонной схемы с компактной свёрточной сетью по точности, времени отклика и требованиям к вычислительным ресурсам.

ЛИТЕРАТУРА

1. Журавлёв, Ю. И. Распознавание. Классификация. Прогноз. Математические методы и их применение / Ю. И. Журавлёв.– Москва: Наука, 1999. – Вып. 2.
2. Фисенко, В. Т. Компьютерная обработка и распознавание изображений: учебное пособие / В.Т.Фисенко. – Санкт-Петербург : СПбГУ ИТМО, 2018. – 192 с.
3. Фу, К. С. Лингвистический подход к распознаванию образов / К. С. Фу. – Москва : Мир, 2010.
4. Bishop, C. M. Pattern Recognition and Machine Learning [Text] / C. M. Bishop. – New York : Springer, 2006. – 738 p.
5. Duda, R. O. Pattern Classification [Text] / R. O. Duda, P. E. Hart, D. G. Stork. – 2nd ed. – New York: Wiley, 2001.
6. Dalal, N. Histograms of Oriented Gradients for Human Detection [Text] / N. Dalal, B. Triggs // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2005. – Vol. 1. – P. 886–893.
7. Ojala, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns [Text] / T. Ojala, M. Pietikäinen, T.Mäenpää// IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – Vol. 24, No. 7. – P. 971–987.
8. LeCun, Y. Deep learning [Text] / Y. LeCun, Y. Bengio, G. Hinton // Nature. – 2015. – Vol. 521, No. 7553. – P. 436–444.
9. Simard, P. Y. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis [Text] / P. Y. Simard, D. Steinkraus, J. C. Platt // Proceedings of the Seventh International Conference on Document Analysis and Recognition. – 2003. – Vol. 3. – P. 958–962.
10. Cyrillic-MNIST: a Cyrillic Version of the MNIST Dataset [Text] / B. Tleubayev, A. Khassanov, R. Mussabayev [et al.] // Proceedings of the Thirteenth Language Resources and Evaluation Conference. – Marseille: European Language Resources Association, 2022.
11. Warren, H. S. Hacker’s Delight [Text] / H. S. Warren. – 2nd ed. – Boston: Addison-Wesley, 2012. – 512 p.