О некоторых подходах к распознаванию оптических образов текстов

А.Б. Фролов, И.Д. Четрафилов

Изучаются некоторые подходы к распознаванию при наличии искажений оптических образов текстов, содержащих фрагменты на разных языках. Процесс распознавания организуется на сети принятия решений, в узлах которой реализуются различные решающие правила для идентификации классов изображений. При этом допускается неопределенность упорядочения и количества признаков, характеризующих распознаваемое изображение. Рассматриваются алгоритмы распознавания упорядоченных объектов, методы назначения и проекций, используемые на этапе первичного распознавания формы изображений символов, и методы постобработки для контекстного анализа с целью правильного алфавитного кодирования.

1 Введение

Одной из первых задач кибернетики оказалась задача распознавания образов. Еще в 1947 г. У.С.Мак-Каллохом была сформулирована проблема создания аппарата, позволяющего слепому воспринимать письменный текст на слух [1, 2].

Задача распознавания решается на некотором множестве объектов, определенном для конкретной ее постановки. Например, для упомянутой выше задачи объектами являются изображения текстов на том или ином языке или их фрагментов. Ее решение — образ текста, т.е. последовательность символов алфавита этого языка. Применительно к распознаванию оптических образов текстов перспективными являются методы комбинаторно-логического направления в распознавании образов [3].

Распознаванию предшествует измерение объекта. Последующей обработкой получают вектор характерных параметров изображения, по которым и осуществляется распознавание. Особенностью параметризации оптических образов изображений при этом является неопределенность упорядочения параметров и разнообразие их количества в зависимости от условий сканирования, полиграфических особенностей и искажений.

Задачей настоящей работы является описание и исследование некоторых подходов к распознаванию при наличии искажений оптических образов текстов, содержащих фрагменты на разных языках. Процесс распознавания организуется на сети принятия решений, в узлах которой реализуются различные решающие правила для идентификации классов изображений. При этом допускается неопределенность упорядочения и количества признаков, характеризующих распознаваемое изображение. Рассматриваются алгоритмы распознавания упорядоченных объектов, методы назначения и проекций, используемые на этапе первичного распознавания формы изображений символов, и методы постобработки для контекстного анализа с целью правильного алфавитного кодирования.

При таком подходе структурные методы распознавания [4] применяются лишь на заключительных этапах процесса, им предшествует применение комбинаторно-логических [3, 6, 7] и метрических [1] методов.

2 Сходство и упорядочение классов шейпов. Распознавание упорядоченных объектов

Базовой задачей распознавания изображения текста является распознавание образа символа. В простейшем случае четко различимых уникальных символов известного алфавита она решается однократно, в других случаях, например, при наличии существенных искажений и при распознавании текстов с фрагментами на разных языках решение разделяется на этап распознавания формы символа и этап определения его алфавитного кода в зависимости от контекста.

Определив, что картинка имеет вид эллипса, мы еще не можем указать соответствующий ей код, так как с равным правом это может быть код цифры 0, латинской, русской или греческой буквы "o"или "O". Поэтому удобно ввести понятие шейпа символа как инварианта формы его

изображения.

Заметим, что есть похожие шейпы, имеющие лишь тонкие различия, например, шейпы "3"и "Э", "h"и "b", "0"и "Q"и другие. Если отказаться от точного распознавания шейпа, то есть распознавать шейпы лишь с точностью до такого рода различий, то получим задачу распознавания групп похожих шейпов. Естественно в этом случае имеется в виду, что результат распознавания шейпа далее будет уточняться.

Полезно ввести в связи с этим отношение толерантности на множестве шейпов: толерантными являются шейпы похожих символов, различия которых могут исчезать при искажениях. Оно позволяет рассматривать задачу распознавания шейпа как задачу многоэтапного сужения класса толерантности.

При таком подходе возникает возможность использования на начальных этапах грубых моделей и методов, допускающих быстрое принятие решений при распознавании группы возможно искаженных оптических образов. Тогда на заключительных этапах применяются более тонкие методы, но не к оптическому образу в целом, а к его фрагментам обладающим особенностями, по которым и различаются шейпы данного класса толерантности. В последующих разделах данной работы рассматриваются метод назначения и метод проекций, применимые соответственно на начальных и заключительных этапах распознавания.

Заметим, что шейпы символов, изображения которых различаются добавлением или исключением существенно отличимых фрагментов относятся к непересекающимся классам толерантности. Это позволяет на множестве хорошо различимых объектов определить отношение частичного порядка. Пусть имеется полное описание этого отношения на множестве классов шейпов. Упорядочим классы линейно, согласованно с этим отношением.

Примечание. Всякое конечное частично упорядоченное множество $(A, \leq)$ может быть преобразовано в линейно упорядоченное множество $(A, \prec)$ таким образом, что $a \leq b$ влечет $a \prec b$ [5].)

Каждую группу представим некоторой совокупностью эталонов — моделей символов, соответствующих шейпам данной группы. Факт вложения по отношению частичного порядка распознаваемого оптического образа в один из этих эталонов и соответствует признанию того, что он

относится к данной группе шейпов. Перебор эталонов классов при этом подчиняется линейному упорядочению классов.

Эталоны формируются по изображению реального оптического образа из обучающей выборки как вектор, получаемый по результатам первичного измерения (например, как бинарная матрица масштабированного изображения, набор числовых характеристик, полученных по тому или иному методу параметризации). При этом векторы различных изображений могут иметь разную длину.

Механизм принятия решений может быть реализован в виде алгоритмов распознавания упорядоченных объектов [6, 7, 8, 9].

Применительно к распознаванию оптических образов текстов функция f, используемая в качестве обучающей выборки, является существенно недоопределенной. Алгоритмы распознавания данного класса реализует некоторое ее расширение. Неопределенность этого расширения для анализируемого объекта соответствует отказу от распознавания.

Применение размытых отношений частичного порядка позволяет уменьшить число отказов от распознавания, характерных для классически определяемых отношений.

Однин из приемов "размытия"заключается в использовании вместо реальных объектов, т.е. непосредственно оптических образов символов их представлений, которые находятся в отношении частичного порядка с исходными, но имеют (в зависимости от выбора направления упорядочения) существенно более широкий верхний или нижний конус, в который, однако, не должны попадать большие (или меньшие) по отношению частичного порядка эталоны других объектов. "Размытие"эталона осуществляется в процессе обучения автоматически или в интерактивном режиме и повышает его эффективность как следствие сокращения области недоопределенности формируемой решающей функции.

При другом подходе для каждой пары объектов определяется несимметричная количественная мера их упорядоченности.

3 Метод назначения

Метод назначения основан на использовании структурной модели оптического образа как совокупности определенным образом связанных фрагментов. В описание модели наряду с перечислением связей включается ряд количественных характеристик фрагментов и модели в целом.

Распознавание осуществляется на основе установления меры соответствия модели оптического образа и эталонных моделей. При этом используются описания структурных связей фрагментов. В условиях искажений возможны нарушения связей фрагментов модели распознаваемого оптического образа. Возможно также изменение упорядоченности фрагментов в описании модели. Тогда для установления степени сходства моделей применяется классическая задача о назначении [10]. Исходные данные при этом образуются по количественным характеристикам моделей.

Например, при построении скелета изображения формируются фрагменты в виде замкнутых или разомкнутых отрезков кривых, каждый из которых характеризуется координатами концов, площадью описанного прямоугольника, типом кривой, ее длиной и другими параметрами. По таким наборам $(a_1, a_2, ..., a_n)$, и $(b_1, b_2, ..., b_m)$, $n \leq m$, относящимся к изображениям двух символов, первое из которых принимается за эталонное изображение можно составить матрицу $W = (w_{ij})$ размера $m \times m$ весовых коэффициентов, $w_{st} = \max\limits_{(i \leq n, j \leq m)} m_{i,j}$, если $s > n$, характеризующих степень различия количественных характеристик i-го фрагмента первого изображения и j-го фрагмента второго изображения. При этом для эталона могут быть выбраны весовые коэффициенты, определяющие степень участия числовых характеристик его фрагментов в формировании матрицы W, то есть меру соответствия элементам неизвестного изображения в зависимости от их числовых и качественных параметров.

Степень соответствия количественных характеристик двух изображений может быть получена в виде суммы $w = \sum_{i=1}^{m} w_{i,\pi(i)}$, где π- подстановка, соответствующая решению задачи о назначении.

Сумма

$$w = \sum_{i=1}^{n} w_{i,\pi(i)} \tag{1}$$

характеризует степень невыполнимости отношения частичного порядка для двух изображений (степень "невложенности"первого изображения во второе).

Если $n < m$, то в результате решения задачи о назначении часть фрагментов изображения оказывается неиспользованной. В этом случае осуществляется итерационный процесс объединения таких фрагментов с использованными. Для этого по тем же правилам формируется новая матрица весовых коэффициентов и решается задача о назначении с

тем, чтобы максимально использовать назначенные фрагменты. Попытка объединения признается успешной, если суммарный вес (1), полученный при данной итерации не превышает результата предыдущей.

Если попытка объединения оказалась неуспешной, то формируется итоговая оценка соответствия шаблона изображению, в которой наряду с последней минимальной суммой весовых коэффициентов учитываются неназначенные фрагменты изображения.

Решение соответствует эталону, для которого получена наилучшая оценка сходства с моделью распознаваемого оптического образа. Перебор эталонов при этом подчиняется линейному порядку, описанному в разд. 2.

Рассмотренный метод предполагает использование максимально упрощенных (не содержащих нехарактерных для данного символа деталей) структурных моделей изображений символов в качестве эталонов. Кроме того эффект "размытия"усиливается путем введения достаточно широких интервалов изменения количественных характеристик. Этим достигается высокая степень обучаемости системы распознавания, поскольку такая модель представляет достаточно большое число реальных изображений. При этом появляется возможность автоматизации выбора границ интервалов при обучении и реализации общей стратегии распознавания частично упорядоченных объектов.

Экспериментально подтверждается устойчивая работа алгоритма распознавания методом назначений в условиях значительного перекоса строк, разрывов изображений символов и включений в виде точечных фрагментов.

Примечание. В реализованной версии алгоритма для определения основных характеристик фрагментов *как вариант* используется авторская версия алгоритма "степной огонь"для построения скелета изображения символа. Его принцип действия схож с получением внутреннего несгораемого каркаса некоторой конструкции после равномерного сжигания внешнего слоя. В результате получается скелет, отражающий структуру символа. Основное требование к процессу выявления скелета это недопущение сопряжения в одной точке более чем трех фрагментов. В результате работы алгоритма символ представляется следующими характеристиками: количество фрагментов, связность фрагментов (количество связей и собственно связи), горизонтальный и вертикальный размер. Фрагменты описываются параметрами: координаты начала и конца, площадь описанного прямоугольника, длина кривой фрагмента и ее характер.

4 Метод проекций

Как отмечено в разделе 2, процесс распознавания формы оптического образа завершается уточнением результата путем анализа возможных тонких топологических особенностей. Наиболее полно они выражены в характерных участках внешнего контура изображения символа и могут быть не замечены при анализе отфильтрованного изображения по методу назначения. Рассматриваемый ниже метод проекций позволяет формировать модели эталонов, отражающие также в обобщенной форме топологические особенности участков внешнего контура. В то же время он устойчив к искажениям локального характера и инвариантен к шрифтовым и другим стилистическим особенностям изображений символов.

Локализация анализируемого участка осуществляется с учетом результата предварительных этапов распознавания.

Анализируются проекции фрагментов по одному из четырех направлений. При этом в оптическом образе выделяется система вложенных прямоугольных фрагментов, расширяющихся в одном из этих направлений.

Примечание. В более общем случае используется восемь или более направлений проектирования и рассматриваются фрагменты более сложной формы.

Динамика изменения проекций соответствует степени выраженности топологической особенности и оценивается рядом количественных характеристик.

На основе подобных параметров, можно анализировать более сложные топологические особенности, например, оценивать степень симметричности изображения.

Получаемые количественные характеристики могут использоваться при перемещении по сети решений. На их основе может осуществляться отображение дискретного объекта в метрическое пространство, после чего возможно применение классических методов распознавания. Кроме того появляется возможность оценки степени неопределенности результата распознавания.

Рассмотрим подробнее некоторые характеристики фрагментов оптических образов изображений, анализируемых по методу проекций. Как отмечено выше фрагментами являются прямоугольные области с коор-

динатами $(X_л, Y_в)$ левого верхнего угла и $(X_п, Y_н)$ правого нижнего угла, образованные $m = X_п + X_л + 1$ столбцами и $n = Y_н + Y_в + 1$ строками.

Фрагменты характеризуются

а) направлением $D(\leftarrow, \rightarrow, \uparrow$ или $\downarrow)$, по которому образуется проекция;

б)граничными фиксированными координатами (Z_0, Z_1, Z_2) фрагмента:

$$(Z_0, Z_1, Z_2) = (X_л, X_п, Y_в), (Z_0, Z_1, Z_2) = (X_л, X_п, Y_н),$$

$$(Z_0, Z_1, Z_2) = (Y_в, Y_н, X_л), (Z_0, Z_1, Z_2) = (Y_в, Y_н, X_п),$$

при направлениях $\leftarrow, \rightarrow, \uparrow, \downarrow$ образования проекции соответственно;

в) кодовым словом $V \in \{B, BW, WB, WBW, Z\}$, где W обозначает любую последовательность точечных элементов фона, B — любую последовательность элементов изображения, Z —любую последовательность точечных элементов, отличающуюся от последовательностей, соответствующих кодовым словам B, BW, WB, WBW.

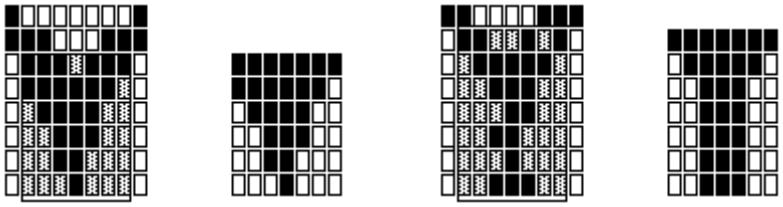
Опишем действие оператора $Pr(D, (Z_0, Z_1, Z_2), V)$ формирования и анализа фрагментов по методу проекций.

Сначала образуется фрагмент с координатами (Z_0, Z_1, Z_2, Z_2) при направлениях $\downarrow, \uparrow$ проектирования или фрагмент (Z_2, Z_2, Z_0, Z_1) при направлениях $\leftarrow, \rightarrow$. Одновременно образуются и их проекции.

Затем добавляется столбец снизу или сверху (строка слева или справа), и образуется проекция расширенного фрагмента. Это действие повторяется до тех пор, пока не будет достигнута граница изображения символа или не будет построен фрагмент, проекция которого соответствует кодовому слову V.

В процессе построения последовательности фрагментов и их проекций формируются числовые параметры, отражающие моменты изменения кодовых слов проекций. Кроме того фиксируется число M изменений проекции, число N, равное наибольшему количеству шагов, при выполнении которых проекция фрагмента не изменялась.

Пример. На рис. *а,б* показаны нижняя часть оптического образа символа "V"и последовательность проекций, получающихся при работе оператора $Pr(\uparrow, 2, 8, 12, B)$. При этом выделяется фрагмент в прямоугольнике на рис. *а*, и формируются его числовые характеристики, в частности, $M = 5$ и $N = 1$.



На рис. *в,г* показаны нижняя часть оптического образа символа "Y"и последовательность проекций, получающихся при работе того же оператора. При этом выделяется фрагмент в прямоугольнике на рис. *в*, и формируются, в частности, параметры $M = 2, N = 5$. Сравнивая подобные параметры для различных пар оптических образов символов "V"и "Y", можно заметить устойчивую тенденцию различия по указанным двум числовым характеристикам оптических образов символа "V"и символа "Y". Таким образом, принятое на предварительном этапе распознавания решение, что оптический образ соответствует одному из этих символов, может быть уточнено по этим характеристикам.

5 Постобработка. Распознавание полиязыковых текстов

Создание OCR-программы, способной различать тексты, в которых имеются фрагменты на разных языках, например, использующих алфавит кириллицы и латинский алфавит, (полиязыковыеые тексты) связано с проблемой различения символов разных алфавитов, имеющих одинаковые изображения. Необходимость этого возникает потому, что символы одного алфавита в тексте на языке, использующем другой алфавит, хотя и не нарушают его читаемости, исключают возможность использования лексических корректоров и других программ обработки текстов. Кроме того необходимо обеспечить возможность шрифтового выделения в распознанном тексте слов разных языков.

Результат распознавания оптических образа текста на уровне шейпов представляется как некоторый список, элементами которого являются коды шейпов отдельных символов, сопровождаемые координатами их оптических образов. В этом списке порядок расположения элементов может не соответствовать взаимному расположению оптических образов символов в образе текста документа. Кроме того коды шейпов могут

не соответствовать требуемым алфавитным кодам символов. Устранение такого рода неопределенностей и является задачей постобработки.

Для замены кодов шейпов правильными алфавитными кодами символов используется ряд унарных и бинарных предикатов на множестве символов. Ими описываются множества пар прописных и строчных символов и подмножество пар таких символов с одинаковыми шейпами, множества символов, являющихся цифрами, знаками, символами алфавита того или иного языка и др. Кроме того используется ряд функций на множестве всех символов, значения которых качественно определяют уровень верхней или нижней границы (относительно воображаемой или вычисляемой линии уровня строки текста) изображения символа или его ширины.

Такие отношения и функции с учетом информации о координатах оптических образов символов позволяют решить такие задачи постобработки как упорядочение результата первичного распознавания в соответствии с расположением образов символов в образе текста, определение пробелов между словами, различение прописных и строчных символов и соответствующая замена кодов символов, определение языка отдельных слов по уникальным символам одного языка и соответственно замена кодов шейпов алфавитными кодами символов.

Таким образом устраняется неоднозначность трактовки результатов первичного распознавания шейпов и достигается представление результата в виде текстового файла, коды символов которого соответствуют кодам символов распознаваемого оптического образа.

Заметим, что рассмотренные методы преобразования результатов первичного распознавания относятся к теории структурного распознавания образов [4] и основаны (при их точном описании) на использовании специального вида формальных грамматик.

Описанный подход к проблеме оптического распознавания полиязыковых текстов позволяет достаточно просто расширять возможности OCR-программы путем дополнений в описания используемых отношений и функций, не изменяя собственно механизмов распознавания или контекстного анализа.

Разработанной экспериментальной OCR-программой, в подавляющем большинстве случаев осуществляется правильное кодирование подобных символов алфавита кириллицы или алфавитов на основе латинских символов. Решение обеспечивается методами контекстного анализа, базиру-

ющимися на рассмотренных выше предикатах и функциях на множестве символов.

Использование описанных в настоящей работе методов первичного распознавания обеспечило устойчивость программы к перекосам и нарушениям линейности строк и другим искажениям, возникающим вследствие дефектов сканирования или полиграфии.

6 Заключение

В настоящей работе отражены результаты исследований проблемы распознавания при наличии искажений оптических образов текстов содержащих фрагменты на разных языках. Предлагаемые подходы основаны на использовании комбинаторно-логических моделей, методов и алгоритмов. В рамках общей схемы распознавания частично упорядоченных объектов реализуются описанные в работе метод назначения и проекций как базовый и уточняющий методы принятия решений. Процесс распознавания организуется на сети принятия решений, что позволяет управлять выбором методов и параметров алгоритмов на различных стадиях распознавания.

В исследованиях также участвовали Фролов Д.А., Хомяков П.Б., Савельев В.Е., Долотова О.А., Гудалевич Л. и Фарамазян Н.В. Условие полиязыковости было сформулировано как требование к OCR программе Клаусом Вашиком. Исследование осуществлялось при поддержке фирмы LINK&LINK (ФРГ, Дортмунд). При содействии Строгалова А.С и Зайцева И.В. экспериментальная OCR-программа была использована в составе программного обеспечения рабочего места озвучивания оптических образов текстов для слепых.

Список литературы

[1] Ту Дж., Гонсалес Р. Принципы распознавания образов. — М.: Мир, 1978.

[2] Винер Н. Кибернетика или управление и связь в животном и машине. (М.: "Советское радио", 1968.

[3] Кудрявцев В.Б.. Алешин С.В. Комбинаторно-логический подход к распознаванию образов. Интеллектуальные системы. Том 1, выпуск 1 — 4, 1966.

[4] Фу К.. Структурные методы в распознавании образов. — М.: Мир. 1977.

[5] Биркгоф Г. Теория решеток. — М.: Наука, 1984.

[6] Фролов А.Б., Яко Э. Алгоритмы распознавания частично упорядоченных объектов и их применение. Изв. АН СССР. Серия Техническая кибернетика, 1990, N5.

[7] Фролов А.Б., Фролов Д.А. Алгоритмы распознавания упорядоченных объектов в системах принятия решений функционального типа. Вестник МЭИ, 1996, N 6.

[8] Фролов А.Б., Фролов Д.А., Яко Э. Функциональные схемы для распознавания упорядоченных объектов. Известия РАН. Серия Теория и системы управления N 5, 1997.

[9] Фролов А.Б., Фролов Д.А., Четрафилов И.Д. Распознавание в интеллектуальных системах функционального типа. Интеллектуальные системы. Том 2. вып.1—4, 1997.

[10] Холл М. Комбинаторика. — М.: Мир. 1970.