АКЦИОНЕРНОЕ ОБЩЕСТВО
«ВЕДУЩИЙ ПРОЕКТНО-ИЗЫСКАТЕЛЬСКИЙ И НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ ПРОМЫШЛЕННОЙ ТЕХНОЛОГИИ»
(АО «ВНИПИпромтехнологии»)

Центр развития цифрового инжиниринга

Интеллектуальная обработка исходных данных для гидрогеологического моделирования

Работу выполнил:

_____ Ильина С.А.
(подпись)

Согласовано:
Руководитель центра

_____ Перепелкин А.В.
(подпись)
«05» 02 2025 г.

2025 г.

Введение

В настоящий момент, специалисты, занимающиеся созданием гидрогеологических моделей, часто получают исходные данные в виде растровых изображений (отсканированных файлов, фотографий и т.п.), что затрудняет процесс переноса данных в программные средства для моделирования. Специалисты вынуждены осуществлять ручной перенос необходимой информации из изображений в редактируемые файлы текстового формата. Данный рутинный процесс требует больших трудозатрат, поэтому было предложено разработать ряд программных средств, позволяющих реализовать автоматическое извлечение данных из растровых изображений определённых форматов.

Целью данной работы является сокращение времени, затрачиваемого на подготовку исходных данных в процессе построения гидрогеологических моделей, и повышение качества данных, извлекаемых из растровых изображений. Для её достижения необходимо выполнить следующие задачи:

- Разработать алгоритм извлечения данных из растровых изображений колонок технологических и разведочных скважин с точностью распознавания символов чисел – 100% и точностью распознавания буквенных символов не менее 85%;
- Разработать алгоритм извлечения данных из растровых изображений геологических разрезов с точностью извлечения символов чисел – 100% и точностью определения литологического состава не менее 95%;
- Создать комплекс программных средств с графическим пользовательским интерфейсом для возможности применения разработанных алгоритмов в производственном процессе.

Результаты данной работы направлены на совершенствование производственных процессов НИЛ-5, но могут быть использованы и на других предприятиях дивизиона, для автоматического извлечения данных из аналогичных изображений, для которых не сохранилась версия в редактируемом формате.

Решение проблемы извлечения данных из геологических колонок

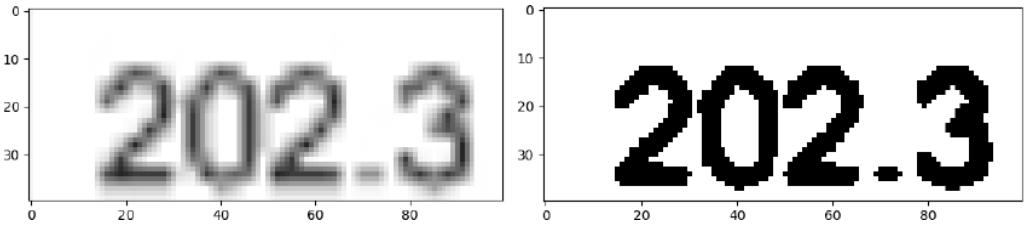
Геологические колонки скважин представляют собой табличную структуру, где каждая строка является отдельным слоем верхней части земной коры, а столбцы могут содержать различную информацию в зависимости от типа колонки. В данной работе рассматривается два вида колонок: разведочные и технологические; различающиеся типом скважин, по которым они были построены.

Извлекаемая из геологических колонок информация представлена на растровых изображениях двумя способами: текст и шкалы глубин. Для автоматического извлечения текстовой информации можно воспользоваться методами оптического распознавания символов (Optical Character Recognition – OCR) направленными на решение задачи перевода поступившего на вход изображения текста в формате графического файла, в формат текстовых данных, использующийся для представления символов в компьютере. Однако, существующие реализации методов OCR имеют ряд ограничений, в том числе не все программные средства осуществляют извлечение данных представленных в виде таблиц таким образом, чтобы сохранить их структуру. Также качество распознавания снижается при наличии на изображении посторонних элементов

(символы условных обозначений, линии границ ячеек и другие графические элементы) вблизи текстовых блоков, в случае слишком малого или слишком большого пустого пространства изображения (однотонный фон) вокруг текстового блока и т. д. [1]. Поэтому стандартные реализации метода OCR не позволяют добиться качественного извлечения данных из изображений геологических колонок. Для автоматического извлечения данных шкалы глубин не существует готовых решений, однако автоматизация позволила бы ускорить процесс и ограничить возможные ошибки пределами погрешностей, определяемых качеством изображения.

Учитывая обозначенные проблемы существующих решений, для сохранения структуры таблицы, в данной работе предлагается применять метод OCR только к фрагментам исходного изображения, которые являются отдельными ячейками таблицы. Поиск ячеек таблицы реализуется обнаружением границ раздела чёрного и белого цветов. Границы таблицы являются строго вертикальными и горизонтальными чёрными линиями, а фон изображения имеет белый цвет. Таким образом передвигаясь вниз и вправо от верхнего левого угла изображения можно найти все необходимые границы таблицы, причём обнаружение ячеек будет происходить в заданной известной последовательности.

Когда определены границы ячейки, попадающая в них область изображения выделяется отдельным фрагментом, к которому применяется алгоритм OCR. На этапе предобработки алгоритма OCR необходимо избавиться от шумов и уточнить контуры символов. В данной работе, для этих целей применяется бинаризации. Эта процедура разделяет все пиксели исходного изображения (в цветовой схеме оттенков серого) на два класса – чёрные и белые. Таким образом, можно получить изображение с максимальной контрастностью и, при качественном выборе метода бинаризации, точными контурами символов. В данной реализации, по результатам вычислительных экспериментов, был выбран метод бинаризации Оцу, пример работы метода приведён на рисунке 1. Этот метод ищет такое пороговое значение, которое позволило бы минимизировать дисперсию внутри выделяемых классов и максимизировать между.



а) б)

Рисунок 1 — а) исходное изображение текста, б) результат бинаризации методом Оцу

Ячейки обрабатываемых таблиц имели различные размеры, т.к. высота каждой ячейки, в некотором масштабе определяет толщину соответствующего ей слоя. По этой причине, при извлечении ячеек таблицы как фрагментов изображения, возникали ситуации, в которых текст занимал всё пространство получаемого изображения. В таких случаях, на этапах сегментации и классификации алгоритма

OCR могут возникнуть ошибки [1], поэтому, когда высота изображений была меньше или равна высоте ожидаемых символов, к фрагментам добавлялись белые рамки. При распознавании глубин, выраженных числом, были возможны и обратные случаи, высота ячейки могла оказаться во много раз больше ожидаемой высоты символов, в этом случае весь текст находился в нижней части ячейки и мог быть частично или полностью упущен на этапе сегментации [1]. Чтобы избавиться от ошибок, в подобных случаях извлекался фрагмент изображения, верхняя граница которого находилась ниже соотвтетствующей границы ячейки. Данные процедуры производились до этапа бинаризации, что также способствовало лучшему качеству предобработки.

Этап сегментации алгоритма OCR производился методами OSD (Orientation and script detection) [2], целью данного этапа является выделение фрагментов изображения содержащих отдельные символы. На следующем этапе, классификации, определяется какой именно символ изображён на выделенном фрагменте. В данной работе этап реализован посредством применения комбинированной модели, совмещающей в себе LSTM (Long short-term memory) нейронную сеть и алгоритм классификации на основе векторов признаков, обученной на наборе символов русского и английского алфавитов [3]. Для извлечения значений глубин, использовалась более ограниченная версия модели, классифицирующая символы цифр от 0 до 9, запятые и точки.

Часть данных в некоторых таблицах представлена при помощи шкалы глубин. Для работы с данным элементом применялись методы обнаружения границ чёрного и белого цветов для определения координат делений шкал. Указанные значения делений также извлекались методами OCR, аналогично глубинам из ячеек таблиц. Далее, по двум известным значениям шкалы в метрах, вычислялся масштаб данной шкалы глубин в метрах в пиксели по формуле:

$$mpp = \frac{y_2 - y_1}{d_2 - d_1}, y_2 > y_1, d_2 > d_1 \quad (1)$$

где $y_{1,2}$ – координаты двух известных делений шкалы по оси ординат изображения, $d_{1,2}$ – указанные значения делений измеренные в метрах от поверхности. Полученное значение масштаба использовалось для перевода ординат пересечений границ строк таблицы со шкалой глубин в метры, таким образом получая глубины разделов слоёв.

Извлекаемые таким образом данные сохраняются в виде переменных строк. Каждая строка соответствует конкретной ячейки таблицы, поэтому полученные результаты можно записывать в файл формата .XLSX в правильном порядке. Результаты извлечения глубин отдельно переводились в формат дробных чисел, для возможности использования их в вычислениях.

Для разработанного алгоритма было определено два параметра качества: точность распознавания чисел и точность распознавания текста. Оценка точности производилась по следующей формуле:

$$accuracy(y, \hat{y}) = \frac{1}{n_{characters}} \sum_{i=0}^{n_{characters}-1} 1(\hat{y}_i = y_i) \quad (2)$$

где $n_{characters}$ – общее число распознаваемых символов, y_i – действительное значение *i*-го символа, $\hat{y}_i$ – значение *i*-го символа, которое вернул алгоритм.

В качестве тестирования алгоритма, была произведена обработка 10 изображений, всего содержащих 280560 символов текста и 1454 символа чисел. Согласно оценке, вычисленной по формуле (2), точность распознавания чисел составила 100%, точность распознавания текста составила 93% для изображений колонок разведочных скважин и 86% для изображений колонок технологических скважин, которые составляли 30% от общей выборки. Отметим, что часть извлечённых чисел была получена не методом оптического распознавания символов, а при помощи обработки шкал глубин. Извлечённые таким методом значения не входили в оценку точности, параметром качества для них является величина погрешности, зависящая от разрешения изображения – чем больше разрешение, тем меньше вычислительная погрешность. На тестовом наборе данных, средняя погрешность при работе со шкалой глубин составила 0,18 м. Полученные оценки качества показывают удовлетворительные результаты, поэтому данное решение уже внедрено в производственный процесс НИЛ-5.

Решение проблемы извлечения данных из геологических разрезов

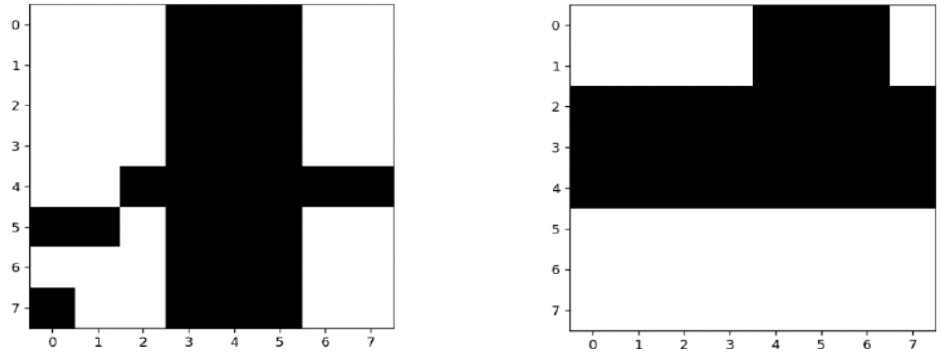
Геологические разрезы содержат в себе информацию о литологическом составе слоёв, о положении скважин относительно азимута, их глубине и уровне их устья, также на разрезах могут быть указаны уровень подземных вод на участке или области оруднения с определённым процентом содержания урана. Несмотря на современные возможности векторной графики, большое количество более давних разрезов хранится в виде растровых изображений, а информация, представленная на них, остаётся актуальной, например, для построения прогнозных гидрогеологических моделей местности. В процессе извлечения информации из исходных данных такого формата возникают проблемы, из-за отсутствия масштабной линейки на изображениях разрезов процесс требует большого количества времени, при этом полученная информация может содержать большие погрешности и ошибки. Большая часть данных на геологических разрезах представлена при помощи различных условных обозначений, только некоторые численные данные представлены в виде текста. По этой причине, автоматическое извлечения данных из подобных изображений существующими средствами невозможно. Однако, потенциально, автоматическое чтение геологических разрезов может позволить десятикратно сократить продолжительность процесса.

Учитывая обозначенную проблему, в данной работе предлагается решение в виде алгоритма, применяющего методы OCR для распознавания всей возможной информации, представленной в текстовом виде. Для работы с условными обозначениями разрезов предлагается применить методы попиксельной обработки изображений, бинаризацию, а также классификацию фрагментов изображений по предварительно выделенным признакам.

Так как на вход поступает полное изображение разреза, первым этапом алгоритма является обнаружение вертикальных проекций скважин на разрезе. Поиск скважин осуществляется путём поиска их номеров методами OCR вблизи линии азимута. Азимут всегда располагается ниже разреза, а отметки скважин

вдоль азимута определяют положение забоя и устья проекции скважины на разрезе по оси абсцисс. Зная положение устья скважины по оси абсцисс, её положение по оси ординат определяется обнаружением верхней границы разреза, вблизи устья скважины методами OCR распознаются её номер и абсолютная отметка устья.

Далее алгоритм анализирует пиксели вдоль линии скважины, для обнаружения пересечений слоёв. Для анализа выделяются квадраты размером 8 х 8 пикселей, бинаризируются методом Оцу [4], и проверяются на наличие на них вертикальной линии соответствующей линии скважины и линии, ориентированной горизонтально (рисунок 2 а), соответствующей разделу слоёв, уровню подземных вод или границе оруднения, в зависимости от цвета линии (чёрный, синий, красный или зелёный соответственно) на исходном изображении.



а) б)

Рисунок 2 — Бинаризированные фрагменты вертикальной проекции скважины содержащие а) горизонтально ориентированную линию, б) отметку забоя

Ниже отметки забоя (рисунок 2 б) расположено значение его глубины, которое распознаётся методами OCR. Зная глубину скважины, абсолютный уровень её устья и количество пикселей между двумя этими отметками, можно вычислить абсолютные и относительные глубины разделов слоёв, границ областей оруднения. Для определения масштаба использована формула:

$$mpp = \frac{deep_p - wellhead_p}{deep_m} \quad (3)$$

где $deep_p$, $wellhead_p$ – положение забоя и устья скважины по оси ординат в пикселях, $deep_m$ – указанная глубина скважины в метрах.

В процессе анализа скважины, при обнаружении линий раздела слоёв, вырезается фрагмент изображения, не менее 30 пикселей в высоту и 100 пикселей в ширину, содержащий условные обозначения слоя, расположенного между двумя линиями раздела и содержащий условные обозначения литологического состава. Эти фрагменты поступают на вход классифицирующей модели, которая определяет литологический состав, соответствующий данным условным обозначениям. В случае обнаружения уровня подземных вод (синяя, горизонтально ориентированная линия), вырезается фрагмент изображения правее и выше обнаруженной точки, содержащий глубину уровня вод в метрах, которая извлекается из данного фрагмента методами OCR.

В разработанном алгоритме, для определения литологического состава предлагается использование классифицирующей модели, работающей по методу опорных векторов (Support Vector Classification – SVC). Данный метод принимает на вход вектор обучающих примеров *X* и вектор соответствующих классов *Y*. Цель метода найти такие ω, b чтобы при определении класса по формуле $\omega^T x + b$, как можно больше обучающих примеров были определены верно. Для нахождения коэффициентов решается следующая задача:

$$\min_{\omega,b} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}\max\left(0,1-y_i\left(\omega^T x_i + b\right)\right) \tag{4}$$

где *C* – константа, определяющая уровень штрафа, при неправильной классификации образца на этапе обучения [5].

Используемый метод классификации работает с признаками изображений, а не со всей матрицей. Для извлечения признаков из классифицируемых изображений использовался оператор локального двоичного шаблона (Local Binary Patterns – LBP). Данный оператор, принимает на вход значение пикселя в оттенках серого, далее, на основании значений *P* соседей радиуса *R* данного пикселя, вычисляется новое значение по следующим формулам:

$$\begin{aligned} LBP_{P,R} &= \sum_{i=0}^{P-1} S\left(g_i - g_c\right)2^i \\ S\left(x\right) &= \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \end{aligned} \tag{5}$$

где g_i, g_c – значения *i*-го пикселя соседа и входного (центрального) пикселя, соответственно, в оттенках серого [6]. После применения LBP оператора ко всем пикселям изображения, по выходному массиву строится гистограмма, для использования в качестве вектора признаков алгоритма классификации.

Извлекаемые таким образом данные сохраняются в виде переменных строк и записываются в файл формата .XLSX. Результаты извлечения глубин отдельно переводились в формат дробных чисел, для возможности использования их в вычислениях.

Для разработанного алгоритма было определено три параметра качества: точность распознавания чисел, точность определения литологического состава, погрешность определения глубин. Оценка точности распознавания чисел, как текстовой информации извлекаемой методами OCR производилась по формуле (2). Для оценки погрешностей глубин, использовались данные о глубинах границ урановых оруднений, полученные в результате геологических изысканий, которые сравнивались с извлечёнными значениями. Оценка точности классификации при определении литологического состава производилась по следующей формуле:

$$accuracy = \frac{P}{N} \tag{6}$$

где *P* – количество верно классифицированных объектов, *N* – общее количество объектов в выборке.

Для тестирования алгоритма, произведена обработка 5 разрезов, суммарно содержащих 18 скважин, 342 символа чисел, 12 областей оруднений, а также из изображений было выделено 162 фрагмента с условными обозначениями. Согласно оценке, вычисленной по формуле (1.3), точность распознавания чисел составила 100%. Средняя погрешность определения глубин, полученная в результате сравнения извлечённых значений границ оруднений с известными, составила 0,167 м. Точность определения литологического состава посредством использования разработанной классифицирующей модели, была оценена по формуле (2.5) и составила 87,5%. При этом модель производила классификацию 8 классов, а обучающая выборка содержала 46 векторов, состоящих из 18 признаков, выделенных LBP оператором по 16 соседям радиуса 10.

Несмотря на высокую точность распознавания чисел, точность определения литологического состава является недостаточной для успешного внедрения решения в производственные процессы, поэтому планируется дальнейшая работа по увеличению и балансировке обучающей выборки, с целью добиться точности классификации не менее 95%.

Заключение

На основе программных реализаций алгоритмов ведётся разработка комплекса десктопных приложений, обрабатывающих файлы растровых цветных изображений формата .JPG и .PNG различного разрешения. Для автоматического извлечения данных из геологических колонок, разработано приложение «Чтение геологических разрезов». Данное приложение поддерживает обработку четырёх разновидностей входных данных: колонки разведочных скважин, полученные с трёх разных площадок («Хиагда», «Далур», «ППГХО»), которые различаются представлением информации о глубине (часть колонок вместо столбца глубин содержат шкалу глубин) и разрешением исходных изображений, и колонки технологических скважин. Пользователю доступен выбор одного из четырёх источников информации, загрузка одного или нескольких изображений соответствующего формата и выбор директории для сохранения результатов извлечения данных. По результатам работы приложения создаются отдельные файлы формата .XLSX для каждого входного изображения. Для удобства, создаваемым файлам присваивается то же имя, что и у входного изображения. Данное приложение уже внедрено в производственный процесс НИЛ-5 и показывает положительный эффект в вопросе сокращения времени, затрачиваемого на обработку данных геологических колонок.

Приложение для автоматического извлечения данных из растровых изображений геологических разрезов, в данный момент находится в разработке и будет представлено пользователям после повышения точности алгоритма классификации литологических составов.

Применение разрабатываемого программного комплекса, по оценкам сотрудников НИЛ-5, позволит сократить трудозатраты по обработке исходных данных не менее чем в десять раз. Кроме того, ожидается сокращение ошибок, связанных с человеческим фактором, которые возникают при стандартном подходе по переводу исходных данных из графического формата в редактируемый.

Список используемых источников

1. Tesseract documentation [Электронный ресурс]. – Режим доступа: https://tesseract-ocr.github.io/, свободный (дата обращения 06.06.2024)
2. Unnikrishnan R. Combined Script and Page Orientation Estimation using the Tesseract OCR engine / Ranjith Unnikrishnan, Ray Smith. – Текст : электронный // ACM International Conference Proceeding. – 2009. – Series. 6. – 10.1145/1577802.1577809.
3. Smith R. An Overview of the Tesseract OCR Engine / Ray Smith. – Text : electronic // Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). – 2007. – Vol. 2. – P. 629 - 633. – DOI: 10.1109/ICDAR.2007.4376991.
4. Otsu N. A Threshold Selection Method from Gray-Level Histograms // IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, Jan. 1979. — P. 62–66.
5. Alex J. Smola, Bernhard Schölkopf, A Tutorial on Support Vector Regression // Statistics and Computing archive vol. 14, Issue 3, Aug. 2004. – P. 199-222.
6. Roohum Jegan, R. Jayagowri, Windowed modified discrete cosine transform based textural descriptor approach for voice disorder detection // Inteligent Data-Centric System, vol. Implementation of Smart Healthcare Systems using AI, IoT, and Blockchain – Academic Press, 2023. – P. 147-167.