РАЗРАБОТКА СПЕЦИАЛИЗИРОВАННЫХ АЛГОРИТМОВ ПАКЕТНОГО УСВОЕНИЯ ДАННЫХ САМОПИШУЩИХ ИЗМЕРИТЕЛЬНЫХ ПРИБОРОВ

В.В. Долотов, С.И. Казаков*, А.С. Кузнецов*

Морской гидрофизический институт
НАН Украины
г. Севастополь, ул. Капитанская, 2
E-mail: vdolotov@mail.ru
* ЭО МГИ НАН Украины
г. Ялта, пгт. Кацивели,
ул. Академика Шулейкина, 9

В работе описывается программный инструмент, основанный на алгоритмах оцифровки бумажных лент самопишущих приборов. Инструмент используется для автоматизированной оцифровки записей самопишущего измерителя уровня моря – мареографа.

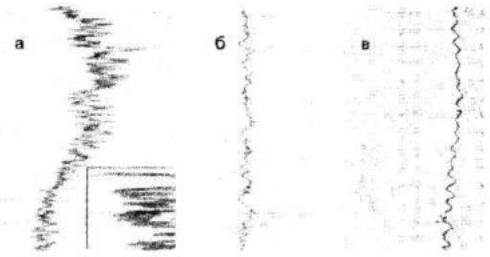
В настоящее время продолжает сохраняться актуальность совершенствования методов оцифровки аналоговых графических данных. Наиболее распространенным и постоянно развивающимся направлением такого рода является оцифровка картографических материалов [1 – 3 и др.], однако в некоторых случаях заслуживают внимания и более простые задачи. Так, например, в ЭО МГИ НАН Украины многие годы наблюдений за уровнем моря и некоторыми другими гидрометеорологическими параметрами окружающей среды привели к накоплению массивов в виде тщательно сохраняемых лент регистрирующих приборов, содержащих уникальную и неповторимую информацию. В качестве примера можно привести записи уровня моря, выполняемые в период с 1899 г. по настоящее время. Если современные приборы выводят непосредственно цифровую информацию, то результаты более ранних измерений, необходимые для анализа различных трендов, существуют лишь на бумажных лентах самопишущих измерителей уровня (СУМ) – мареографов. Обработка этих данных чрезвычайно трудоемка, вследствие чего, большинство записей лежат «мертвым грузом», несмотря на их практическую значимость.

Настоящая работа описывает результаты использования разработанной авторами программы, позволяющей значительно ускорить выполнение необходимых преобразований с получением соответствующих цифровых массивов.

В качестве тестовых использовались записи на лентах, полученныее в Экспериментальном отделении Морского гидрофизического института (ЭО МГИ) НАН Украины.

Эта же программа впоследствии использовалась авторами и для оцифровки аналогичных данных из других источников, представленных в виде сканированных изображений.

Характеристика тестируемых материалов. Для тестирования и отладки методики были выбраны три варианта записей (рис. 1).



Р и с. 1. Фрагменты тестовых образцов записей: а – низкое качество; б – высокое качество; в – современный формат

Первые два представляли собой записи 1960–70-х годов, достаточно пожелтевшие от времени и различающиеся своим качеством. Так запись на рис. 1а выполнена синими жидкими чернилами, что вследствие растекания и впитывания чернил в бумагу привело к существенно неоднородной плотности линии (см. врезку на рис. 1а). Вторая запись (рис. 1б), выполненная чернилами красного цвета, представляет собой вариант относительно высокого качества и имеет достаточно равномерную плотность и четкость линии. Третий вариант (рис. 1в) – современная «стандартная» запись,

достаточно четкая, выполнена синими чернилами.

Все записи содержат пометки, выполненные карандашом, либо чернилами другого цвета, как за пределами линии, так и непосредственно «поверх» ее (хорошо видно на рис. 1б). Дополнительно, все записи включают линии стандартной специальной разметки лент, выполненной в серых цветах.

Основные алгоритмы. При разработке методики рассматривался лишь один естественный вариант: сканирование с последующей программной оцифровкой. Другой способ, основанный на использовании цифровых планшетов с привязкой кривых и ручным их обводом был отброшен, как требующий значительно больших затрат времени, трудоемкости и аккуратности с получением результатов меньшей точности.

Параметры сканирования. Параметры сканирования подразделяются на «глубину цвета» (количество распознаваемых цветов) и «разрешение» (количество точек на дюйм изображения). Известно, что первый параметр измеряется в «bpp» (bit per pixel), второй в «dpi» (dot per inch).

Оценка первого параметра производилась по изображениям, отсканированным в режимах:

1. черно-белое изображение (1 bpp);
2. градации серого (8 bpp);
3. 16 цветов (4 bpp);
4. 256 цветов (8 bpp);
5. "true color" (24 bpp).

Анализ указанных типов изображений, полученных с тестовых записей, показал, что ни один из первых четырех параметров не может использоваться для распознавания, поскольку линия, частично или полностью «сливается» не только с пометками, но даже с линиями сетки ленты самописца. Таким образом, в дальнейшем использовался только пятый вариант глубины цвета, который, помимо прочего, является вариантом «по умолчанию» для большинства современных сканеров и позволяет в дальнейшем детектировать цвета, как составляющие RGB (red, green, blue), что упрощает в дальнейшем проблему цветового декодирования.

Значения второго параметра не столь критичны и определяются, в основном, необходимой точностью оцифровки. Напомним некоторые, принятые как некий «стандарт» значения разрешения: 72 dpi – для экранного представления, 300 dpi – для печати, 150 – 200 dpi – настойчиво предлагается некоторыми сканерами. Анализ наиболее качественного тестового изображения (рис. 1б), отсканированного с разрешением 300 dpi показал, что в ширине линии укладывается около 10 точек, таким образом разрешение в 100 – 200 dpi в данном случае является вполне достаточным. Рассматриваемый параметр определяет и погрешность оцифровки, которую можно определить, разделив значение всей шкалы на количество точек в ней, определяемое как *ширина ленты* (см) / 2,54 * *разрешение*. Так при ширине ленты в пределах шкалы 28 см и значении шкалы 100 ед. при разрешении 300 dpi погрешность оцифровки составит

$$100 / (28 / 2{,}54 * 300) = 0{,}03 \text{ ед. шкалы},$$

а при разрешении 150 dpi погрешность, соответственно составит

$$100 / (28 / 2{,}54 * 150) = 0{,}06 \text{ ед. шкалы}.$$

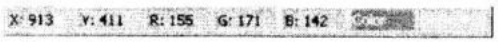
При этом, по-видимому, данные величины можно в обоих случаях считать незначительными.

Параметры распознавания. Алгоритмы распознавания цвета хорошо известны и основная трудность заключается в идентификации цвета кривой, особенно при низком ее качестве.

Распознавание цвета кривой можно осуществить двумя способами: указав на кривую и программно определив ее цвет, либо обнаружив кривую по ее положению в совокупности с контрастностью по отношению к другим присутствующим цветам. Забегая вперед, отметим, что в программе распознавания реализованы оба метода, однако применение первого из них требует некоторой постоянной (плавающей) корректировки, по-

скольку кривая состоит не только из указанного цвета, но и близких к нему цветов, образующихся при впитывании чернил в бумагу. Разработанный же алгоритм корректировки позволяет в большинстве случаев детектировать линию автоматически.

С целью анализа цветовой гаммы изображения был разработан специальный программный модуль, который впоследствии был использован в качестве одного из основных в программе оцифровки. На рис. 2 представлена строка статуса этого модуля, которая позволяет получить информацию о цвете пикселя изображения под курсором.



Р и с. 2. Строка статуса программы

В строке отображаются координаты курсора, отдельные составляющие цвета (RGB), сам цвет пикселя (Color), а также дополнительное поле (Detected), которое показывает о срабатывании алгоритма детектирования цвета линии. Наличие последнего значительно ускорило подбор алгоритма распознавания цвета и, в дальнейшем, в случае некачественной оцифровки позволяет подобрать параметры, описанные ниже (привязать алгоритм к конкретному изображению).

Уверенность в возможности реализации заданного алгоритма заключалась в том, что на всех тестовых изображениях линии сетки и большинство служебных отметок (особенно выполненных карандашом) состоят примерно из одинакового уровня величин R, G, B, т.е. представляют собой оттенки, близкие к серому цвету, в то время как линии, требующие оцифровки и, к сожалению, некоторые служебные отметки, выполненные чернилами, характеризуются преобладанием какого-либо одного (R, G, B) или пары (RG, RB, GB) цветов.

С учетом этого, основной алгоритм оцифровки включает следующие процедуры:

– последовательное сканирование каждой сканированной строки изображения от левого края или от указанной точки начала шкалы до точки, цвет которой соответствует цвету линии (в случае отсутствия такой точки строка исключается из анализа, о чем производится запись в журнал);

– последовательное обратное (справа налево) сканирование каждой строки изображения от правого края или от указанной точки окончания шкалы до точки, цвет которой соответствует цвету линии;

– расчет значений, соответствующих левому и правому границам линии на основании заданной шкалы;

– расчет среднего значения на основании минимального и максимального;

– вывод значений в итоговую таблицу.

Во всех процедурах, работающих с цветом, используются известные алгоритмы повышения контрастности изображения [4 – 6 и др.] с целью достижения более высокого качества распознавания.

Алгоритмы коррекции. Следует отметить, что первая реализация указанного алгоритма показала достаточно четкое распознавание линии, однако при этом наблюдалось немало ложных определений, выражавшихся в следующем:

– ложное определение цвета линии (реакция на линии другого цвета, соответствующие различного рода служебным отметкам);

– ложное определение левого или правого границ линии в виду наличия в сканируемой строке пятен, цвет которых соответствует или близок цвету линии (часто это просто пятна чернил, случайно попавшие в процессе заправки самопишущего прибора или мелкие дефекты производства ленты).

В дальнейшем, с целью минимизации ложных срабатываний алгоритм был усовершенствован дополнением его следующими функциями:

– функцией начального детектирования цвета линии и сравнения этого цвета с цветом анализируемых точек;

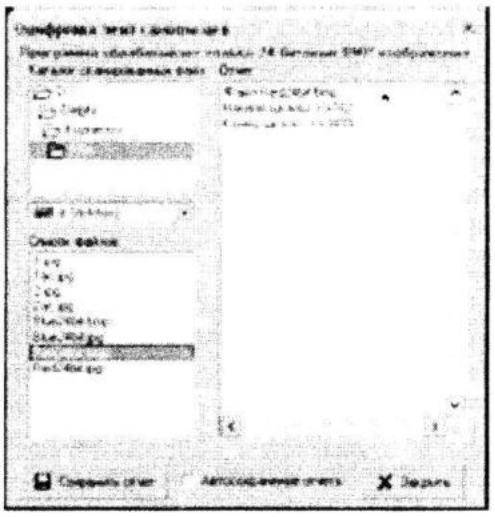
– функцией игнорирования мелких пятен.

Первая из перечисленных функций определяет весь массив контрастных (удовлетворяющих условиям детектирования) точек в пределах 10 первых строк изображения. Далее определяется цвет точек, представленный максимальным количеством с учетом разброса цвета (допуска), который можно регулировать. После этого, в любом случае обнаружения контрастного цвета, последний сравнивается с детектированным на первом этапе «цветом линии» и, в случае значительного несовпадения, исключается из анализа. Реализация данной функции в общем алгоритме оцифровки сразу показала практически полное прекращение ложных срабатываний на посторонние контрастные цвета.

Алгоритм игнорирования мелких пятен основан на анализе цвета окружающих точек и, в первую очередь, следующих далее за анализируемой. В случае, если в пределах нескольких точек (количество задается программно) контрастный цвет пропадает, анализируемая точка игнорируется.

В результате реализации всей совокупности алгоритмов программа показала вполне работоспособный вариант, который даже при «средних» значениях регулируемых параметров одинаково четко оцифровывает все три тестовых изображения.

Описание программы. Как принято при разработке программного обеспечения, взаимодействующего с пользователем, программа оцифровки загружается в виде основного окна (рис. 3), позволяющего выбрать одно из заранее отсканированных изображений. В правой части окна реализован журнал работы, обновляемый для каждого изображения и сохраняемый автоматически или по требованию пользователя в текстовом файле. В дальнейшем в журнал заносятся все параметры оцифровки и сообщения о программно обнаруженных проблемах.

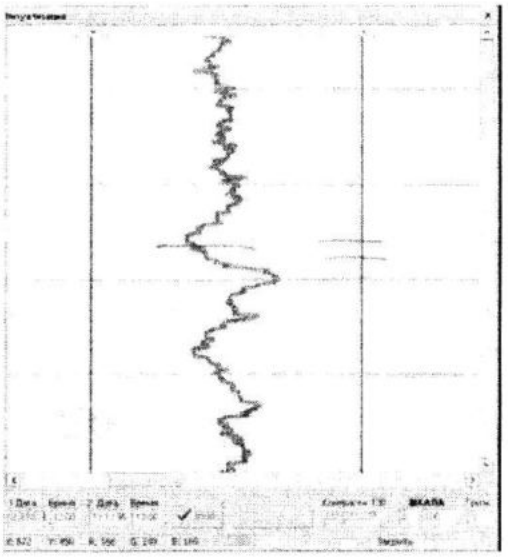


Р и с. 3. Основное окно программы

После выбора изображения последнее загружается в отдельное окно (рис. 4) с увеличением, пропорциональным параметру разрешения при сканировании. Помимо изображения окно представлено следующими элементами (строка статуса описывалась ранее на рис. 2):

– верхняя узкая полоса с метками начала (Н) и окончания (К) шкалы;

– панель параметров оцифровки, несколько видоизменяющаяся в зависимости от режима работы.

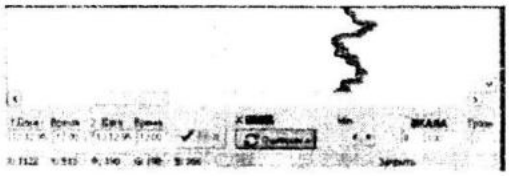


Р и с. 4. Окно параметров оцифровки

Метки начала и окончания шкалы могут перемещаться по щелчку в соответствующей полосе, что позволяет их корректировать в случае неверного автоматического распознавания. Их положение отмечается и на изображении соответствующими вертикальными линиями.

Панель параметров позволяет указать время начала и окончания временного интервала, соответствующего записи на ленте, при этом допускается указать временной период, соответствующий любым двум точкам изображения с последующим автоматическим пересчетом на начало и окончание ленты, о чем производится соответствующая запись в журнал. Кнопка «Ввод» фиксирует временные интервалы в программе и активирует кнопку «Оцифровка». Указанная кнопка активирует процесс оцифровки с учетом указанных значений начала и окончания шкалы и количества пикселей, допустимых для «грязных пятен». Бегунок «Контраст» позволяет установить «четкость» линии, которая в соответствии с пределами изменения интенсивности каждой составляющей цвета может изменяться в пределах 0 – 255, однако, поскольку контраст не может быть «нулевым» и максимальным, значения крайних положений движка ограничены значениями «80» и «180».

При нажатии на кнопку «Оцифровка» выполняется собственно оцифровка изображения с отображением прогресс – индикатора в виде значения процентов на кнопке. После окончания процесса оцифровки панель параметров выглядит следующим образом (рис. 5).

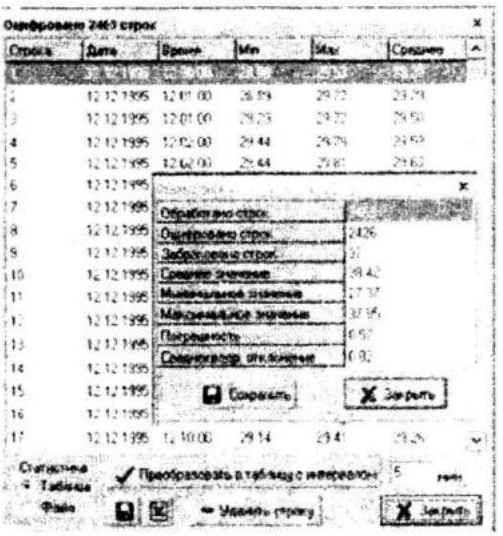


Р и с. 5. Панель параметров после оцифровки

Бегунок «Контраст» заменяется переключателем «Min – Среднее – Max», позволяющем накладывать на изображение оцифрованную кривую, соответствующую левой (Min) или правой (Max) границе линии либо линию, рассчитанную как их среднее арифметическое.

В процессе оцифровки в журнал могут выводиться сообщения о забракованных линиях. Этот брак практически не сказывается на результатах, т.к. возникает на «отдельных» линиях, в то время как частота линий чаще всего избыточна. Кроме того, при количестве линий изображения около 2500 (как в тестовых вариантах), пропуск даже нескольких десятков одиночных линий мало влияет на результаты. Причины брака заключаются либо в слабой насыщенности линии оцифровки, либо в «замазывании» ее другим цветом (служебные отметки).

Процедура оцифровки завершается отображением результирующего окна с таблицей (рис. 6) с соответствующим окном результирующей статистики, которая, впрочем, может записываться во внешний файл автоматически. Погрешность при расчете статистики вычисляется методом «скользящего» среднего, т.е. для каждой оцифрованной строки изображения. Это же значение используется и в расчетах среднеквадратичного отклонения.



Р и с. 6. Результирующая таблица

С целью приведения измерений к единому временному интервалу предусмотрен режим пересчета с генерированием новой таблицы, включающей лишь усредненные данные. Алгоритм пересчета таков, что при наличии измерений в заданный интервал их среднее значение записывается в таблицу, при их отсутствии выполняется интерполяция между ближайшими значениями. Сам временной интервал может задаваться произвольным.

Возможности редактирования оцифрованных данных. В любом случае в процессе оцифровки вероятно появление ошибок. С целью их идентификации и удаления предусмотрен ряд мер. Так, строки результирующей таблицы интерактивно связаны с оцифрованными линиями изображения и наоборот, что позволяет щелчком мыши на «дефектной» строке изображения мгновенно переместиться в соответствующую строку таблицы. Для удаления дефектных записей предусмотрена соответствующая кнопка.

Заключение. Разработанная и описанная программа оцифровки в тестовых вариантах на стандартных компьютерах средней мощности позволяла выполнять оцифровку записей самописцев физического размера около 28 см длиной (предел используемого сканера) в течение нескольких секунд с обработкой около трех тысяч сканированных строк изображения и получением соответствующего количества табличных записей. Учитывая более значительное время, требуемое для сканирования, в расчетах трудоемкости следует использовать именно его, но поскольку оно измеряется несколькими (около одной) минутами после прогрева сканера, то общий процесс оцифровки можно квалифицировать как чрезвычайно эффективный.

На момент подготовки статьи были обработаны данные самописца с 2005 по 2009 гг. Это составило порядка 6000 «кадров» сканера обрабатываемых лент самописца. Дискретность снятия информации – 1 мин. Полученные данные загружаются в специализированную базу данных «Мареограф» [7], которая в настоящее время включает 2223253 записи.

Следует отметить, что до разработки описанной программы обработка лент выполнялась вручную с оцифровкой данных мареографа с дискретностью 1 час. Проводить съем информации вручную с дискретностью 1 мин практически невозможно. Разработанная программа предоставляет такую возможность при сохранении необходимой точности снятия данных и многократно ускоряет этот процесс.

СПИСОК ЛИТЕРАТУРЫ

1. *Программа* EasyTrace. www.dataplus.ru /Soft/VECTORIZ/Easy7.5/easytrace.htm
2. *ArcGIS* Desktop. www.dataplus.ru/Soft /ESRI/ArcGIS/ArcGIS.htm
3. *Программа* Golden Software Diger. www.goldensoftware.com/products/ didger/didger.shtml
4. *Увеличение* резкости фотографий. – fotopit.ucoz.ua/publ/3-1-0-11
5. *Сойфер В.А.* Компьютерная обработка изображений. Методы и алгоритмы // Самарский государственный аэрокосмический университет, 1996. www.pereplet.ru/obrazovanie/stsoros/68.html
6. *Сканирование* и коррекция изображений. arttower.ru/tutorial/Svetilkin/ spravochnicki/Photoshop_for_WEB/ Charter4/1.htm
7. *Иванов В.А., Долотов В.В., Казаков С.И., Кузнецов А.С.* Развитие субрегиональной информационно-аналитической системы научного центра междисциплинарных исследований НАН УКРАИНЫ на базе черноморского экспериментального полигона «КАЦИВЕЛИ» // Сб. "Экологическая безопасность прибрежной и шельфовой зон и комплексное использование ресурсов шельфа". – Севастополь: МГИ НАН Украины, 2010. – Вып. 21. – С. 10 – 24.