УДК 551.501

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

Использование современных компьютерных технологий для автоматизации усвоения данных самопишущих измерительных приборов

В работе описывается разработанный авторами вариант прикладной программы автоматической оцифровки записей ленточных самопишущих приборов, основанной на использовании сканированных изображений и специальных алгоритмов распознавания линий. Результаты тестирования программы, полученные с использованием записей различного качества, показали высокую избирательность к цвету и толщине записанных линий. В сочетании с развитыми средствами оперативной коррекции ошибок это позволяет в автоматическом режиме и с высокой скоростью обрабатывать информацию и помещать ее в соответствующие базы данных.

Ключевые слова: самопишущие приборы, оцифровка, распознавание, технология, базы данных, уровень моря.

Введение

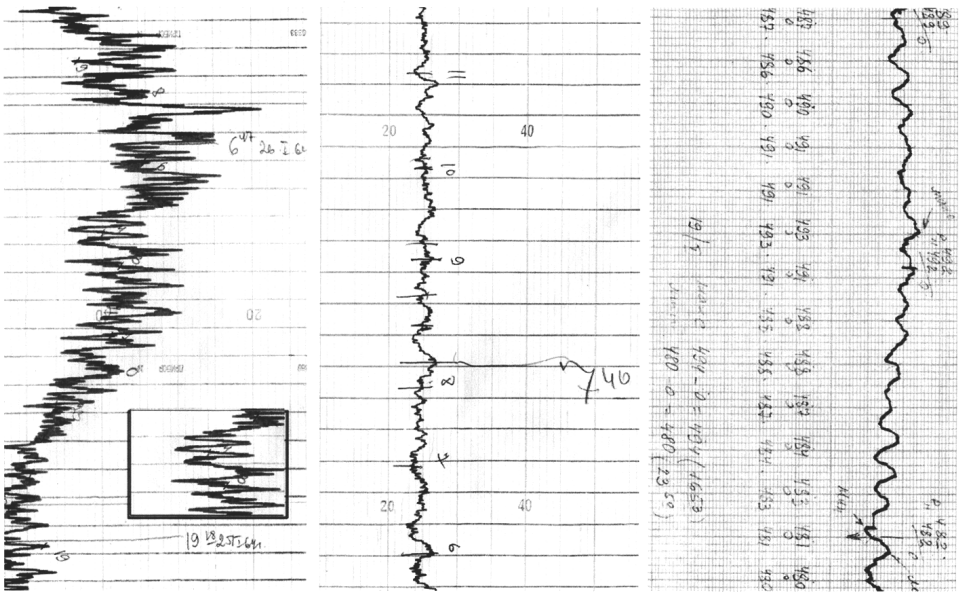
В океанологии, как и в других науках о природе, очень часто используются алгоритмы вычисления различных трендов. При этом очевидна ценность архивных данных о параметрах окружающей среды для прогнозирования изменений климата. До появления цифровых автоматических приборов такие данные накапливались на бумажных лентах самопишущих измерителей. Размещение этой информации в современных базах цифровых данных представляет очевидную проблему.

Так, например, в Экспериментальном отделении Морского гидрофизического института НАН Украины за многие годы наблюдений за уровнем моря и некоторыми другими гидрометеорологическими параметрами окружающей среды накопились огромные архивы в виде лент регистрирующих приборов, содержащих уникальную информацию. Наиболее типичный пример – записи самопишущего измерителя уровня моря (мареографа). Обработка этих данных вручную чрезвычайно трудоемка, вследствие чего большинство записей обрабатывается медленно и с низким разрешением (1 ч).

Настоящая работа основана на результатах использования разработанной авторами программы, позволяющей значительно ускорить выполнение необходимых преобразований с получением соответствующих цифровых массивов более высокого разрешения.

Характеристика тестируемых материалов

Для тестирования и совершенствования программных алгоритмов были выбраны три варианта записей (рис. 1). Первые два из них представляют собой записи, сделанные в 1960 – 1970-х годах, достаточно пожелтевшие от времени и различающиеся своим качеством. Так, запись на рис. 1, *а* сделана синими чернилами низкого качества (жидкими) и вследствие растекания и впитывания чернил в бумагу имеет существенно различающуюся плотность линии (см. врезку на рис. 1, *а*). Вторая запись (рис. 1, *б*), выполненная чернилами красного цвета, представляет собой вариант относительно высокого качества с равномерной плотностью и четкостью линии. Третий вариант (рис. 1, *в*) – современная запись, достаточно четкая, выполнена синими чернилами.



Р и с. 1. Фрагменты тестовых образцов записей самопишущего измерителя уровня моря: *а* – низкое качество, *б* – высокое качество, *в* – современный формат

Все записи содержат пометки, сделанные карандашом либо чернилами другого цвета как за пределами линии, так и непосредственно поверх нее, и, кроме того, включают серые линии стандартной специальной разметки.

Основные алгоритмы

В качестве основного рассматривался лишь один вариант обработки данных: сканирование с последующей программной оцифровкой. Другой способ, основанный на использовании цифровых планшетов с координатной привязкой кривых и обводом их вручную, был отброшен как трудоемкий и требующий значительно больших затрат времени. Кроме того, полученные результаты имели бы меньшую точность. Рассмотрим первый вариант более подробно.

Параметры сканирования. Параметры сканирования подразделяются на «глубину цвета» (количество распознаваемых цветов) и «разрешение» (количество точек на дюйм изображения). Принято первый из них измерять в «*bpp*» (*bit per pixel*), второй – в «*dpi*» (*dot per inch*). В тестовых испытаниях были определены оптимальные для эффективного распознавания значения параметров: 24 *bpp* и 300 *dpi*. При этом анализ наиболее качественного тестового изображения (рис. 1, *б*) показал, что в ширине линии укладывается около 10 точек, таким образом, и разрешение в 200 *dpi* в данном случае является вполне достаточным. Рассматриваемый параметр определяет и погрешность оцифровки, которую можно вычислить следующим образом:

диапазон шкалы / (ширина ленты, см / 2,54 · разрешение).

Так, например, при ширине тестовой ленты в пределах шкалы 28 см и диапазоне шкалы 100 ед. при разрешении 300 *dpi* погрешность оцифровки составит

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

а при разрешении 150 *dpi* погрешность, соответственно, в два раза выше. При этом, по-видимому, эти 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 курсором на кривую и программным способом определив ее цвет либо обнаружив кривую по ее положению и контрастности по отношению к другим присутствующим цветам. В описываемой программе распознавания реализованы оба метода, однако разработанный алгоритм коррекции помех позволяет в большинстве случаев детектировать линию автоматически.

Уверенность в возможности реализации заданного алгоритма заключалась в том, что на всех тестовых изображениях линии сетки и большинство служебных отметок (особенно сделанных карандашом) представляют собой оттенки серого цвета (близкие по значению величины цветовых составляющих *RGB*). 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 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 левой и правой границам линии на основании заданной шкалы;

– расчет среднего значения положения линии на основании определенных ранее минимального и максимального;

– вывод значений в итоговую таблицу.

Дополнительно во всех процедурах, связанных с идентификацией цвета, используются известные алгоритмы повышения контрастности изображения (см., например, [1 – 3]) 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 левой или правой границ линии ввиду 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 определений алгоритм был усовершенствован и дополнен 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 идентифицируется цвет точек по их максимальному непрерывному массиву с учетом разброса цвета (допуска), который можно регулировать. После этого при обнаружении 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.

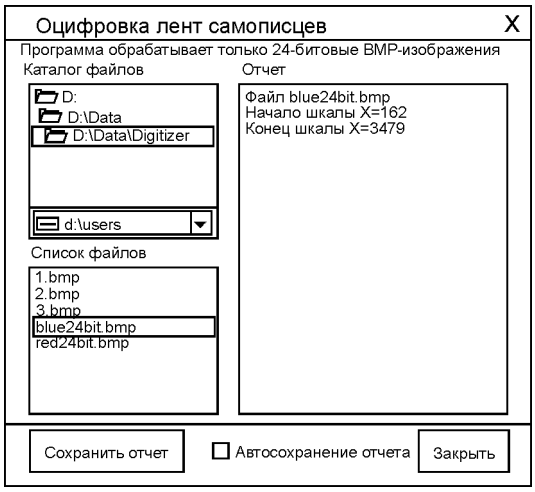
Алгоритм игнорирования мелких пятен основан на анализе цвета окружающих точек и, в первую очередь, следующих далее за анализируемой. В случае если в пределах некоторого заданного оператором интервала контрастный цвет пропадает, анализируемая точка игнорируется.

В результате реализации всей совокупности алгоритмов был получен вполне работоспособный вариант программы, в котором даже при средних значениях регулируемых параметров одинаково четко оцифровываются все три тестовых изображения.

Описание программы

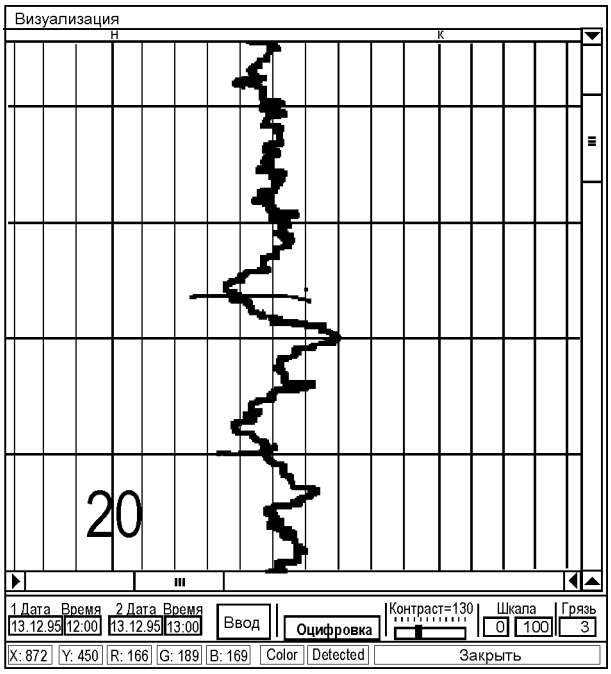
Как принято при разработке программного обеспечения, взаимодействующего с пользователем, программа имеет оконный интерфейс (рис. 2), позволяющий выбрать одно из предварительно отсканированных изображений. В правой части окна расположен журнал работы, обновляемый для каждого изображения и сохраняемый автоматически или по требованию пользователя

в текстовом файле. В дальнейшем в журнал заносятся все параметры оцифровки и сообщения об обнаруженных проблемах.



Р и с. 2. Основное окно программы

После выбора изображения последнее загружается в отдельное окно (рис. 3) с увеличением, пропорциональным параметру разрешения при сканировании.



Р и с. 3. Окно параметров оцифровки

Помимо изображения в окне представлены следующие элементы:

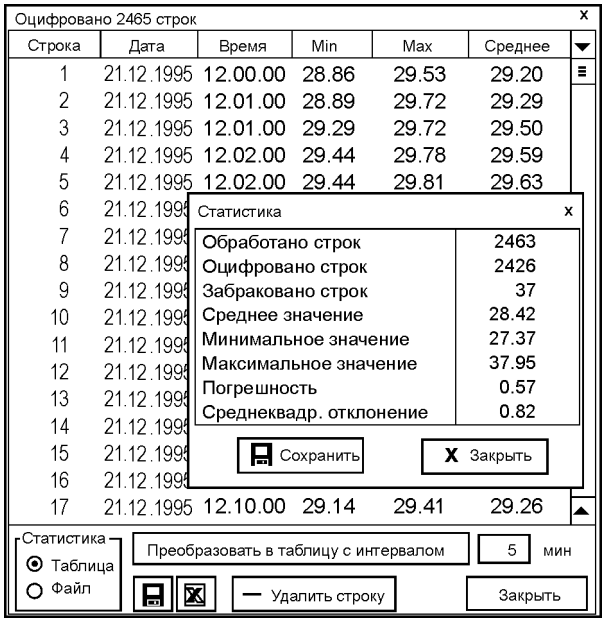
– верхняя узкая полоса с метками начала (Н) и окончания (К) шкалы;

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

Метки начала и окончания шкалы могут перемещаться пользователем, что позволяет корректировать их положение в случае неверного автоматического распознавания. Положение меток обозначается и на изображении соответствующими вертикальными линиями.

Панель параметров оцифровки 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. Кнопка «Ввод» фиксирует временные интервалы в программе и активирует кнопку «Оцифровка». Регулятор «Контраст» позволяет в некоторых пределах повысить четкость линии.

В процессе оцифровки в журнал выводятся сообщения о забракованных линиях. Брак практически не сказывается на результатах, так как это касается лишь единичных линий, в то время как их частота и количество чаще всего избыточны. Например, 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 цвета детектируемой линии, либо в смешивании его с другим цветом (например, с цветом служебных отметок).



Р и с. 4. Итоговая таблица данных

Процедура оцифровки завершается отображением итоговой таблицы данных (рис. 4) и предоставлением статистических результатов оцифровки, которые могут записываться во внешний файл автоматически. Погрешность при расчете статистических характеристик вычисляется методом скользящего среднего. С целью приведения измерений к единому временному интервалу предусмотрен режим пересчета с генерированием новой таблицы, включающей лишь усредненные по заданному временному интервалу данные. Алгоритм пересчета таков, что при наличии измерений в заданном интервале их среднее значение записывается в таблицу, а 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. Это позволяет щелчком мыши на дефектной линии изображения мгновенно переместиться в соответствующую строку таблицы результатов и удалить дефектную запись.

Заключение

Разработанная и описанная программа оцифровки в тестовых вариантах на стандартных компьютерах средней мощности позволяла выполнять оцифровку записей длиной около 30 см (предел используемого сканера) в течение нескольких секунд с обработкой около 3000 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 секундами после прогрева сканера, то общий процесс оцифровки можно квалифицировать как чрезвычайно эффективный.

На момент подготовки статьи были обработаны данные самописцев с 2005 по 2009 гг. Это составило порядка 6000 сканированных страниц. Дискретность оцифровки – 1 мин. Полученная информация загружена в специализированную базу данных «Мареограф» [4], которая в настоящее время включает 2 223 253 записи.

Следует отметить, что до создания описанной программы обработка лент выполнялась вручную с оцифровкой данных мареографа с дискретностью 1 ч. Проводить оцифровку вручную с дискретностью 1 мин практически невозможно. Представленная авторами программа позволяет производить обработку данных в автоматическом режиме с высокой дискретностью при сохранении необходимой точности.

СПИСОК ЛИТЕРАТУРЫ

1. *Увеличение* резкости фотографий. – http://fotopit.ucoz.ua/publ/3-1-0-11.
2. *Сойфер В.А.* Компьютерная обработка изображений. Методы и алгоритмы. – Самарский государственный аэрокосмический университет, 1996. – http://www.pereplet.ru/obrazovanie/stsoros/68.html.
3. *Сканирование* и коррекция изображений. – http://arttower.ru.
4. *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» // Экологическая безопасность прибрежной и шельфовой зон и комплексное использование ресурсов шельфа. – Севастополь: МГИ НАН Украины, 2010. – Вып. 21. – С. 10 – 24.

Морской гидрофизический институт НАН Украины,
Севастополь
E-mail: vdolotov@mail.ru
Экспериментальное отделение
Морского гидрофизического института НАН Украины,
пос. Кацивели
E-mail: edmhi@ukr.net

Материал поступил
в редакцию 08.12.11
После доработки 27.02.12

АНОТАЦІЯ У роботі описується розроблений авторами варіант прикладної програми автоматичного оцифрування записів стрічкових самописних приладів, яка базується на використанні сканованих зображень і спеціальних алгоритмів розпізнавання ліній. Результати тестування програми, отримані з використанням записів різної якості, показали високу вибірковість до кольору та товщини записаних ліній. У поєднанні з розвиненими засобами оперативної корекції помилок це дозволяє в автоматичному режимі та з високою швидкістю обробляти інформацію та поміщати її у відповідні бази даних.

Ключові слова: самописні прилади, оцифровка, розпізнавання, технологія, бази даних, рівень моря.

ABSTRACT A variant of the applied software developed by the author for automatic digitizing of records of tape recording devices based on application of scanned images and special algorithms for lines recognition is described. The results of program testing obtained due to application of records of various quality show high color and thickness selectivity among the recorded lines. Being coupled with the developed tools for error correction, it permits to process and download the information to the corresponding databases in the automatic mode and at high rate.

Keywords: recording devices, digitizing, recognition, technology, databases, sea level.