Автоматизация информационных технологий создания цифровых графических документов со слабо формализованным описанием объектов

Ю.Г. Васин, Л.И. Лебедев
Центр информатики и интеллектуальных информационных технологий
Института информационных технологий, математики и механики
Нижегородского государственного университета им. Н.И. Лобачевского
603005, Россия, Нижний Новгород, ул. Ульянова, 10, ЦИИТ ИИТММ ННГУ.
E-mail: ya.vasinyuri@yandex.ru, lebedev@pmk.unn.ru

Аннотация

В работе описываются методы повышения эффективности технологии создания цифровых графических документов (ГД) на основе двухкритериальных алгоритмов распознавания с самообучением, способов их распараллеливания, а также автоматизации интерактивных процедур редактирования и присвоения кодов классификации автоматически построенным эталонам. Полученные результаты демонстрируется на примере оцифровки изображения реальных документов.

1. ВВЕДЕНИЕ

В настоящее время актуальной остается проблема создания электронных архивов большеформатных графических документов (топографических и морских навигационных карт, конструкторской документации, чертежей, схем и др.). При этом возникает необходимость перевода с графических изображений и бумажных носителей в цифровой вид, создание цифровых графических документов (ГД) в терминах соответствующей проблемной области. Данный процесс является крайне трудоемким. С целью автоматизации технологии ввода ГД в последние годы предлагается использование методов распознавания с учителем. Однако многие большеформатные графические документы выполнены ручным способом со слабо формализованным начертанием объектов. В связи с этим крайне сложно сформировать набор эталонов для систем распознавания с учителем, что приводит к низкому уровню распознавания объектов исходного изображения.

2. ПОСТАНОВКА ЗАДАЧИ

В работе с целью дальнейшего повышения эффективности технологий создания цифровых ГД предлагается использование разработанных методов распознавания с самообучением и в результате замены базовых наборов эталонов, сформированным в результате обучения с учителем по множеству исходных документов, совокупностью эталонов, порожденных объектами данного ГД.

3. МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ

Распознавание объектов на основе вычисления оценок сходства предлагается осуществлять корреляционно- экстремальным контурным методом (КЭКМ) [1]. В основу КЭКМ положено вычисление оценки близости ε_m по описаниям $w=(x,y)$ объекта O и $\overset{e}{w}=(\overset{e}{x},\overset{e}{y})$ эталона E, которые задаются посредством векторной (контурной) модели представления информации:

$$\varepsilon_m = \overset{e}{Dw} - R^2 \Big/ Dw,$$

где $\overset{e}{Dw}$, Dw - дисперсии эталона и объекта соответственно, а R - величина, вычисленная по значениям смешанных корреляционных моментов:

$$R^2 = Sn^2 + Cs^2,$$

$$Sn = \operatorname{cov}(\overset{e}{x},y) - \operatorname{cov}(\overset{e}{y},x),$$

$$Cs = \operatorname{cov}(\overset{e}{x},x) + \operatorname{cov}(\overset{e}{y},y).$$

Оценка близости ε_m корректна только для согласованных описаний контуров. В общем случае она зависит от положения начальной точки Q описания контура объекта $\varepsilon_m(Q)$. Задача нахождения согласованных описаний сводится к решению оптимизационной задачи вида

$$\varepsilon_{\min} = \min_{Q\in O} \varepsilon_m(Q)$$

Для решения этой задачи применяются эффективные по быстродействию метод относительных смещений и метод парабол [2-4]. В результате, КЭКМ инвариантен не только относительно ортогональных преобразований и масштабирования, но также и к циклическим описаниям контуров, что обеспечивает наиболее полный состав объектов в классах эквивалентностей. Методы распознавания на базе КЭКМ в режиме самообучения с автоматическим пополнением эталонов на основе описаний неопознанных объектов были реализованы в технологиях создания цифровых графических документов [4]. Таким образом, после работы КЭКМ в режиме самообучения на выходе будут сформированы два файла: файл эталонов и файл для параметрического описания распознанных объектов, представляющий совокупность записей постоянной длины. Каждая запись включает в себя данные, необходимые для восстановления объекта, в том числе ссылку на эталон с которым получено наибольшее значение оценки сходства.

Эффективность предлагаемой технологии во многом определяется быстродействием алгоритма распознавания КЭКМ с самообучением. Один из основных подходов увеличения быстродействия заключается в распараллеливании потоков данных. Известно, что сложность вычисления оценки сходства с использованием КЭКМ пропорциональна числу точек в описаниях эталона и объекта. В связи с этим для обеспечения равномерной загрузки ядер очередной эталон, получаемый при распознавании с самообучением, необходимо отнести к потоку, в котором суммарное количество точек в описаниях эталонов этой группы является наименьшим.

В целях повышения качества распознавания объектов на базе КЭКМ реализована возможность использования дополнительного критерия оценки сходства. Поэтому, при распознавании объектов на базе КЭКМ предусмотрена возможность использования дополнительной оценки δ

для расстояния Хаусдорфа d_H, вычисляемого по описаниям контуров исходного O и восстановленного объекта $\tilde{O}$,

$$d_H = \max\left\{ \sup_{w\in O}\inf_{\tilde{w}\in\tilde{O}}\|w-\tilde{w}\|, \sup_{\tilde{w}\in\tilde{O}}\inf_{w\in O}\|w-\tilde{w}\| \right\}$$

Метод нахождения величины d_H предусматривает вычисление этого расстояния по вспомогательным описаниям исходного и восстановленного объектов, для которых была получена минимальная оценка близости $\varepsilon_{\min}$ [2,3,6]. Расстояние Хаусдорфа d_H вычисляется только в том случае, если значение $\varepsilon_{\min} \le \lambda$, то есть когда оценка близости $\varepsilon_{\min}$ не будет превосходить заданный порог λ. Если в этом случае $d_H > \delta$, то рассматриваемый объект вносится в список эталонов. Так как в этот момент на всех ядрах процессора вычисления завершены, то алгоритм нахождения величины d_H в целях увеличения быстродействия также распараллеливается. Суть алгоритма распараллеливания здесь состоит в следующем. Вычисляются расстояния r_i между узловыми точками $\hat{w}_i \in O$ и $\hat{w}_i^e \in \tilde{O}$, $i = 1,2,...,n$, вспомогательных описаний контуров объекта и эталона $r_i = \left\| \hat{w}_i = \hat{w}_i^e \right\|$. Если $r_i > \delta$, то оба отрезка, прилежащих к i-тому узлу, заносятся в соответствующий набор данных. Далее, аналогичным образом, осуществляется распараллеливание на уровне потоков для полученного набора отрезков. В совокупности распараллеливание алгоритма распознавания позволяет при удвоении числа ядер увеличить быстродействие до 1.8 раз.

В результате решения задачи распознавания с самообучением все объекты разбиваются на классы эквивалентностей. Далее требуется классам эквивалентностей присвоить соответствующий идентификатор [7]. В соответствии с предлагаемой технологией присвоение идентификаторов осуществляется в интерактивном режиме при визуализации и просмотре эталонов. С целью оптимизации временных затрат на этапе интерактивной идентификации совокупности выделенных эталонов были разработаны алгоритмы логической фильтрации, позволяющие автоматически осуществлять исключение эталонов, явно не относящихся к классам распознаваемых объектов (шумовые точки, объекты, не удовлетворяющие заданным критериям по размерам и др.). Для множества оставшихся эталонов разработаны методы их кластеризации на подмножества в целях одновременного присвоения меток всем элементам, входящих в полученное подмножество. Для решения кластер- анализа эталоны представлялись в признаковом пространстве. В качестве компонент c_i^j признакового пространства R_n были использованы оценки близости эталонов E^j с базовым эталоном B при различных циклических заданиях начальной точки описания контура $E^j : c_i^j = \varepsilon_m(E^j, B, s_i = i\cdot\Delta S)$, $\Delta S = S^b / n$, S^b - длина базового контура. Получение этих оценок близости не требует дополнительных вычислительных затрат, так как они в обязательном порядке находятся при согласовании описаний в методе относительных смещений [5,6]. Для обеспечения инвариантности относительно положения на контуре начальной точки описания в качестве первого признака пространства R_n^* путем циклических сдвигов выбирается компонента с наименьшим значением оценки близости. Само решение задачи кластер–анализа в пространстве R_n^* осуществлялось с использование метрик Евклида (χ) и Чебышева (γ). Для построения очередного кластера выбирался непомеченный эталон с наибольшим числом распознанных по нему объектов. Описание c^{k*} этого эталона в признаковом пространстве R_n^* берется в качестве центра кластера. Элементами кластера первого уровня будут эталоны из числа непомеченных, описание которых в пространстве R_n^* с учетом визуальной идентификации удовлетворяет условиям: $\left\| c^{k*} - c^{j} \right\| < \chi^{k*}$, $\left| c_i^{k*} - c_i^{j} \right| < \gamma^{k*}$, $\forall i \in \overline{1,n}$. Далее по описаниям в пространстве R_n^* эталонов кластера первого уровня формируются новые центры множеств и определяются эталоны второго уровня, удовлетворяющие тем же условиям и которые не являются элементами множеств, как предыдущего уровня, так и уже ранее сформированных множеств рассматриваемого уровня. Этот процесс заканчивается, если на текущем уровне не будет выделено элементов кластера следующего уровня. Таким образом, кластер будет состоять из совокупности множеств всех уровней и представлять собой планарный граф с древовидной структурой без петель и самопересечений. Пороговые величины χ^{k*} и γ^{k*} приближенно можно оценить следующим образом. Рассмотрим согласованные с базовым эталоном E^b описания эталонов E^{k*} и E^j (следует отметить, что описания всех эталонов центрированы). Преобразуем описание контура E^j в соответствии с параметрами его оптимального наложения на контур E^{k*}. Учитывая инвариантность оценок близости к ОПМ, величины c_i^{k*} и c_i^j после этих преобразований не изменятся. Полученные описания эталонов будем рассматривать как точки в некотором пространстве Ω, в котором по осям откладываются значения компонент точек контуров. Тогда c_i^{k*} и c_i^j это длины сторон треугольника, образованного этими точками. Отсюда, их разность меньше третьей стороны, приближенно являющейся оценкой близости $\varepsilon_m(E^{k*}, E^j)$ эталонов E^{k*} и E^j. Для эталонов, принадлежащих одному классу, максимальное допустимое значение равное $\varepsilon_m(E^{k*}, E^j)$ и будет величиной γ^{k*}. Далее, если допустить, что величины $\left| c_i^{k*} - c_i^{j} \right|$ на отрезке $\left[0, \gamma^{k*}\right]$ подчиняются некоторому

закону распределения, то значение порога χ^{k*} также легко оценить.

В результате для интерактивной идентификации нужно предъявлять не отдельные эталоны, а кластеры, которых в разы меньше, что позволяет существенно снизить временные затраты на создание цифровых документов.

Для формирования сложных объектов (надписи, отметки глубин, структурированные линейные объекты и т. д.), состоящих из последовательности распознанных дискретных и линейных объектов, реализован режим автоматической сборки таких элементов содержания документа [7].

Еще одной возможностью снижения временных затрат является автоматический анализ большеформатных объектов. Часть этих объектов представляют собой составные объекты, образованные в результате прилипания дискретных объектов к линейным. В этом случае решение задачи разделения слипшихся линейных и дискретных объектов осуществляется на основе перехода к сегментно-узловой модели или применения методов распознавания составных объектов на базе КЭКМ [9-11].

4. ЗАКЛЮЧЕНИЕ

Выбор эталонов на основе методов распознавания КЭКМ с самообучением позволило существенно повысить достоверность распознавания, а методы автоматизации повысить эффективность интерактивных процедур. В итоге при оцифровке планшетов гидрографической съемки в зависимости от качества документа было получено снижение временных затрат в 10-20 раз.

Работа выполнена при поддержке Российского Научного Фонда, проект № 16-11-00068.

Список литературы

1. Васин Ю.Г., Лебедев Л.И., Пучкова О.В. Контурные корреляционно-экстремальные методы обнаружения и совмещения объектов видеоинформации./Автоматизация обработки сложной графической информации: Межвуз. темат. сб. науч. тр./Под ред. Ю.Г. Васина.- Горьков. гос. ун-т, Горький, 1987. С.97-112.
2. Васин Ю.Г., Лебедев Л.И. Задача нахождения согласованных описаний в корреляционно-экстремальных контурных методах распознавания. //Математические методы распознавания образов (ММРО-15): 15-ая Всеросс. конф.: Сборник докладов. / М.: Изд-во ООО «МАКС Пресс», 2011. С.342-345.
3. Vasin Yu.G.,Lebedev L.I. The problem of obtaining coherent contour descriptions in the calculation of similarity estimates. //8th Open German-Russian Workshop “Pattern recognition and image understanding” (OGRW-8-2011): Workshop Proceedings. 2011. P. 324-327.
4. Vasin Yu.G., Lebedev L.I. An effective format for representing graphic information // Pattern Recognition and Image Analysis, 2012, Vol. 22, No. 2, pp. 393–398. © Pleiades Publishing, Ltd., 2012.
5. Lebedev L.I., Vasin Yu.G. Optimization of the representation structure and complexity in the development of an intelligent format for graphic images information // Pattern Recognition and Image Analysis, 2014, Vol. 24, No. 4, pp. 530–534. © Pleiades Publishing, Ltd., 2014.
6. Лебедев Л.И. Корреляционно – экстремальные контурные методы распознавания. Теоретические основы: Учебное пособие./ Нижний Новгород, Изд-во Нижегородского государственного университета, 2013. 113 с. – ISBN 978-5-91326-308-7.
7. Vasin Yu.G, Lebedev L.I. Automation methods for technologies to produce digital graphic documents with weakly formalized description of objects // The 11-th International Conference "Pattern recognition and image analysis: new information technologies" (PRIA-11-2013). September 23-28, 2013. Conference proceedings, Samara: IPSI RAS, 2013. Vol. I. P.342-344.
8. Васин Ю.Г., Лебедев Л.И., Морозов В.А. Модификация двухуровневого алгоритма распознавания последовательностей графических изображений. //Математические методы распознавания образов (ММРО-10): 10-ая Всеросс. конф.: Тез.докл. / М.: Изд-во «АЛЕВ-В», 2001. С.176-179.
9. Васин Ю.Г., Лебедев Л.И. Распознавание составных объектов изображения на базе структурного и корреляционно-экстремальных методов. //Математические методы распознавания образов (ММРО-13): 13-ая Всеросс. конф.: Сборник докладов. / М.: Изд-во ООО «МАКС Пресс», 2007. С.285-288.
10. Vasin Yu.G, Gromov V.P. Structural analysis of raster images // The 11-th International Conference "Pattern recognition and image analysis: new information technologies" (PRIA-11-2013). September 23-28, 2013. Conference proceedings, Samara: IPSI RAS, 2013. Vol. I. P.338-341.