Проблемы формализации разметки графического образа документа

М. В. Будаковский, А. А. Михайлов

Аннотация. Работа посвящена вопросу создания разметки графического образа документа. Перечислены виды графической информации, дана классификация документов и элементов разметки. Вводится формальное описание разметки, описывается ее структура. Систематизированы проблемы при решении задач создания разметки и оценки качества построенной разметки. Описываются основные подходы построения разметки.

Ключевые слова: *разметка изображения документа, формализация разметки, уровни разметки, оценка качества, layout analysis.*

Введение

В современном мире люди заполняют огромное количество различных документов, которые затем используются в документообороте. Все большее распространение получают системы, автоматизирующие процессы ввода и обработки документов [1, 2]. Подобные технологии позволяют ускорить обработку документов в десятки раз. После оцифровки документа с помощью сканера или другого устройства, изображение документа подается на вход системе обработки и распознавания. Одной из самых сложных задач, существующих в системах такого рода, является задача построения разметки цифрового изображения.

В компьютерных системах построение разметки является одним из первых этапов анализа изображения. Конечный успех компьютерных процедур анализа изображений во многом зависит от корректной разметки документа. По этой причине значительное внимание должно быть уделено повышению ее надежности.

В процессе глобальной компьютеризации появляется всё больше прикладных областей, в которых актуально применение разметки изображения — это системы распознавания документов, системы поиска и сжатия изображений.

1. Виды графической информации

При анализе оцифрованной графической информации часто требуется выделить на изображении атомарные объекты для последующей обработки и распознавания, то есть создать разметку изображения.

Существует достаточно обширный класс изображений, содержащих текстовую информацию. Авторами предлагается разделить данный класс на два подкласса в зависимости от доли объектов, содержащих текстовую информацию:

- Изображения, в которых доля объектов содержащих текстовую информацию, мала;
- Графические образы, содержащие существенную долю информацию в виде текстов.

К последнему относятся изображения различных видов документов, оцифрованные специализированной техникой. Системы потокового массового ввода и распознавания работают на изображениях документов, содержащих как текстовую информацию в печатном виде, так и текст в письменной форме. Подобные документы содержат:

- печати, штампы;
- схемы, таблицы, графы, чертежи, элементы разграфки;
- тексты;
- формулы; (1)
- штрих-коды, двумерные графические коды;
- логотипы, фото;
- подписи.

Документы, содержащие текстовую информацию, авторы предлагают сгруппировать следующим образом:

- финансовые документы: счета-фактуры, платежные поручения;
- административные: договора, заявления, свидетельства, удостоверения личности;

- формы: анкеты, опросные листы, бланки;
- публицистика: журнальные и газетные страницы, книги.

Документы, относящиеся к первым трем категориям, в дальнейшем будем называть структурированными документами. Структура таких документов может быть достаточно сложной. Далее будем рассматривать разметку образов структурированных документов, так как их анализ в современном мире осуществляется в промышленных масштабах и имеет большое прикладное значение.

2. Объекты интереса разметки

Разметка изображения документа позволяет локализовать области изображения, содержащие однородные объекты интереса. Это требуется в задаче распознавания для семантического анализа документа [3]. В качестве объектов интереса в зависимости от цели построения разметки выступают графические и текстовые блоки, таблицы и элементы, составляющие данные объекты. От назначения разметки зависит и ее степень детализации, еще до этапа построения решается вопрос какие именно объекты интереса считать элементарными для данной разметки. Таким образом, разметка — это описание геометрической структуры информационного объекта с заданной степенью детализации.

3. Результирующая разметка

Разметка изображения документа описывает структуру всего документа как отдельного информационного объекта. В общем случае разметка документа является иерархической структурой, в корневом узле которой находится изображение документа как целостного информационного объекта. От корня к листьям каждая вершина в иерархии представляет собой некоторое описание (дескриптор) объекта, который входит в состав объекта более низкой детализации, соответствующего предкам текущей вершины. Вершина иерархического графа задается как:

$$V = (\{\alpha^1, \dots, \alpha^n\}, \{\delta^1, \dots, \delta^m\}), \qquad (2)$$

где α^i— предок текущей вершины, $i = 1, \dots, n$, n — количество предков предыдущего уровня детализации для вершины V и соответствующего объекту разметки;

δ^i — потомок текущей вершины, $i = 1, \dots, m$, m — количество объектов, из которых состоит текущий объект, соответствующий вершине V;

В общем случае текущая вершина может иметь несколько предков. Такой спецификой обладают следующие элементы:

- таблицы,
- графы,
- схемы,
- формулы.

В случае таблиц ячейка является составляющим элементом как строки таблицы, так и столбца, на пересечении которых она находится. Вершины графа и элементы схем допускают в силу своей природы несколько предков в иерархии. В формулах множественность предков возникает при наличии нескольких знаков группового суммирования $\sum$ или произведения $\prod$, знаков групповых операций над множествами, пределов, интегралов. В формулах элементы матрицы, так же как и табличные элементы, имеют несколько предков.

Совокупность потомков для данной вершины есть разметка объекта изображения, которому соответствует данная вершина иерархического графа.

4. Уровни разметки

На первом уровне документ — самый крупный элемент разметки и корневой узел в иерархическом графе. Если документ представляет собой составной элемент разметки, то он может быть разбит на текстовые и графические блоки (логотипы, фото), а также блоки, содержащие таблицы и прочие вышеперечисленные элементы разметки (1). Первому уровню разметки согласно (2) соответствует вершина

$$\sigma_{root} = (\emptyset, \{\delta^1, \dots, \delta^m\}). \qquad (3)$$

В дальнейшем каждый блок разбивается на составляющие его части. Так, для текстового блока составляющими частями служат текстовые строки, отдельные слова, символы и слоги при переносе.

5. Минимальная детализация разметки

Подобное разбиение может оканчиваться примитивными объектами, атомами документа, дальнейшее деление которых лишено смысла в рамках конкретной задачи. Атомарному объекту разметки согласно (2) соответствует вершина

$$\sigma_{atom} = (\{\alpha^1, \dots, \alpha^n\}, \emptyset). \qquad (4)$$

Если рассматривать проход по иерархическому графу от потомков к предкам, то атомарные объекты могут объединяться в более сложные, например, текст, разделители, выделители и указатели объеди-

няются в таблицу. Комплексный объект O_V образуется объектами следующего уровня детализации следующим образом:

$$O_V = (\bigcup_{i=1}^{m} O_{\delta^i}) / \bigcup_{i,j,i\neq j} O_{\delta^i} \cap O_{\delta^j}, \quad (5)$$

где O_{δ^i} — объект разметки, соответствующий в иерархическом графе δ^i вершине, являющейся потомком текущей вершины V.

6. Свойства выделенного объекта

Каждая вершина иерархического графа помимо информации о потомках и предках имеет набор атрибутов, описывающих конкретный объект на изображении документа. Набор атрибутов сильно зависит от поставленной задачи, для решения которой строится разметка документа. Авторами выделены следующие группы атрибутов A:

- описывающие тип объекта A_{type};
- представляющие геометрические характеристики $A_{geometric}$. К этой группе относятся геометрические границы A_{border}, угол наклона объекта относительно одной из сторон документа A_{angle}, области пересечения и соприкосновения с другими объектами $A_{interaction}$;
- описывающие статистические особенности объекта $A_{statistic}$, то есть дескрипторы области изображения, являющиеся мерой свойств области: гладкости, шероховатости и регулярности. В качестве дескрипторов используют такие статистические характеристики, как среднее, дисперсия, моменты более высокого порядка, энтропию, которые определяют по гистограмме яркости области изображения [4];
- описывающие спектральные особенности объекта $A_{spectral}$ [4], представляющие направленности присутствующих в изображении периодических или квазипериодические двумерных структур;
- описывающие логические связи между объектами $A_{logical}$.

С другой стороны, атрибуты можно подразделить на группу атрибутов, задающих отношения между объектами разметки $A_{relations}$, и индивидуальные атрибуты $A_{individual}$, которые присущи конкретному объекту.

$$A_{relations} = < A_{interaction}, A_{logical} >, \quad (6)$$

$$A_{individual} = < A_{type}, A_{geometric} \setminus \setminus A_{interaction}, A_{statistic}, A_{spectral} >. \quad (7)$$

Зачастую индивидуальные атрибуты объекта наследуют значения от объектов-предков, или меняются в зависимости от шага разбиения при построении разметки.

Набор атрибутов объекта σ задаются следующим образом:

$$A_\sigma = < A_{individual}, A_{relations} >. \quad (8)$$

Таким образом, каждый объект разметки σ описывается парой V_σ (2) и A_σ(8):

$$\sigma = < V_\sigma, A_\sigma >. \quad (9)$$

Разметка Ω представляет собой:

$$\Omega = \bigcup_d \bigcup_i \sigma_i^d, \quad (10)$$

где σ_i^d — i-объект разметки уровня детализации d.

7. Оценка качества разметки

Для оценки качества при создании разметки зачастую используют супервизорные критерии [5], основанные на вычислении меры отличия текущей разметки от идеальной. При этом идеальная разметка создается экспертами. Классификация каждого пикселя — задача трудоемкая, поэтому имеет смысл заканчивать декомпозицию на объектах-примитивах. В результате сравнения анализируемой разметки с идеальной необходимо получить количественную оценку качества анализируемой разметки, характеризующую меру отклонения от идеальной разметки. Задача сравнения двух разметок состоит из задачи о назначении элементов анализируемой разметки σ_i^d элементам идеальной разметки $ideal\sigma_i^l$ и подсчете некоторого функционала качества, который определяет меру отличия элемента разметки от соответствующего ему элемента идеальной разметки. Задача о назначении объектов формализуется следующим образом:

$$\max_{Gr} \sum_i \sum_m \mu(ideal\sigma_m^l, \sigma_i^d), \quad (11)$$

где Gr — множество сопоставленных между собой пар объектов.

Решением данной задачи будет являться множество

$$\widehat{Gr} = \{< \sigma, ideal\sigma >\}.$$

Для оценки качества для каждой пары из $\widehat{Gr}$ необходимо посчитать функционал $\mathcal{F}: \widehat{Gr} \to \mathcal{R}$. Итоговое качество разметки можно задать:

$$K = \frac{1}{N_B^{ideals}} \cdot \sum_{i=1}^{|\widehat{Gr}|} \mathcal{F}(g_i), \quad (12)$$

где N_B^{ideals} — количество элементов сопоставления, размеченных экспертом;

$$g_i \in \widehat{Gr}, \forall\, i \in \left[1, \left|\widehat{Gr}\right|\right].$$

Легко видеть, что идеальная разметка не может быть задана однозначно. Существует конечное множество различных вариантов идеальной разметки, мощность которого даже в случаях простой разметки достаточно велика, чтобы идеальная разметка была представлена перечислением всех допустимых вариантов. Корректная идеальная разметка должна максимизировать качество результатов дальнейшего анализа изображения, к примеру, качество результатов распознавания. Решением для данной проблемы может стать создание несколько идеальных вариантов разметки, максимально различных между собой. Анализируемая разметка сравнивается с идеальными разметками, и в результате максимальное значение функционала качества будет количественной характеристикой качества анализируемой разметки. Однако, создание группы идеальных разметок, если процесс не автоматизирован, также является достаточно трудоемким, поэтому можно учесть некоторые отклонения от идеальной разметки как вариант нормы при подсчете функционала качества в виде параметров. Для получения точных оценок также необходимо анализировать вариации идеальной разметки и определения допустимых параметров функционала качества для каждого элемента разметки. Однако, в первом приближении, можно считать данные параметры константами для конкретной идеальной разметки, и получить оценку качества как результат сравнения идеальной и анализируемой разметок.

8. Методы построения разметки

Для построения иерархического графа объектов часто возникает задача декомпозиции объекта-предка на составляющие элементы. Данная задача является частным случаем задачи сегментации и не имеет однозначного решения. Большинство алгоритмов, строящих разметку, основываются на однородности объектов по некоторому набору признаков. Иерархический граф для разметки может быть построен как от корня к листьям (нисходящий подход), так и наоборот — от листьев к корню (восходящий подход); оба подхода имеют свои преимущества и недостатки.

Для методов, относящихся к восходящему подходу, характерно на первоначальном этапе рассмотрение низкоуровневых объектов — примитивов, которые в дальнейшем объединяются и классифицируются как области документа. Таким образом, восходящий подход способствует выделению областей, имеющих сложную форму, но не учитывает высокоуровневых объектов. С другой стороны, создание изображения документа посредством нисходящего подхода начинается с анализа изображения в целом, постепенно разбивая его на сегменты. Разбиение основано на критерии однородности: процедура разбиения прекращается, когда элементы изображения становятся однородными по некоторым признакам. Алгоритмы, основанные на нисходящем подходе, обладают высокой скоростью работы, так как не содержат трудоемких операций, но они не позволяют обрабатывать документы со сложной структурой.

Заключение

Разметка структурированного документа является важным промежуточным этапом в процессе анализа документа. Создание разметки структурированного документа в общем случае сводится к построению иерархического графа. Детализация разметки сильно зависит от ее дальнейшего применения.

Основные трудности при формализации разметки:

- большое количество (десятки) различных объектов интереса,
- большое количество (десятки) различных типов связей между объектами интереса,
- множественное подчинение объектов интереса (структура графа не в виде дерева),
- большая логическая сложность модели разметки (ближайший единственный предок объекта интереса может отстоять далеко по графу),
- отсутствие корреляции между разметкой по содержанию (документ — логический блок — … утверждение — слово) и разметкой по представлению (том — страница … абзац — строка),
- многосвязная область локализации объекта интереса,
- сложная форма области локализации объекта интереса,
- размытость области локализации объекта интереса (множественность правильного решения),
- сложность определения меры ошибки для неточной локализации объектов интереса (трудность предсказания границы между регулярным и сингулярным возмущением для распознавания символа),
- концептуальная сложность оценки качества разметки, содержащей логические ошибки ввиду сильной вариативности ущерба для конкретных задач,
- пересекаемость областей интереса,
- неестественное форматирование,
- ошибки представления информации,
- противоречивость требований к локализации объектов интереса (возможное отсутствие точного решения при конфликте требований тривиальности и адекватности).

Литература

1. *Арлазаров В. В.* Управление информационными потоками в системе автоматического ввода документов // Сборник трудов Института системного анализа РАН. М.: URSS, 2002. С. 5–11
2. *Емельянов Н. Е.*, *Арлазаров В. Л.* Прикладные аспекты построения систем на основе документооборота // Труды ИСА РАН «Документооборот. Прикладные аспекты». М.: URSS, 2004. С. 5–11.
3. *Stašák Jozef.* A Contribution to Image Semantic Analysis // INFORUM 2004: 10th Conference on Professional Information Resources, 2004.
4. *Гонсалес Р.*, *Вудс Р.*, *Эддинс С.* Цифровая обработка изображений // М.: Техносфера, 2005. С. 947–956.
5. *Zhang Y. J.*, Advances in image and video segmentation // IBM Press, USA, 2006.

Будаковский Максим Викторович. Аспирант ИСА РАН. Окончил в 2013 г. МФТИ. Область научных интересов: обработка изображений, распознавание образов, искусственный интеллект. E-mail: budakovsky@gmail.com

Михайлов Александр Александрович. Н. с. ИСА РАН. Окончил в 1986 г. МФТИ. Количество печатных работ: 20. Область научных интересов: обработка изображений, распознавание образов. E-mail: almi@cs.isa.ru