Компресія зображень тексту на основі класифікуючої метрики з подавленням шумів друку та сканування.

В.Г. Іванов, М.Г. Любарський, Ю.В. Ломоносов, С.В. Котляр

Національна юридична академія України імені Ярослава Мудрого, Харків
nuau@bestnet.kharkov.ua

Анотація

Представлена нова метрика, яка визначає ступінь близькості зображень двох символів при їх класифікації. Ця метрика мало чутлива до шумів друку і сканування, оскільки заснована на стабільних характеристиках, які не враховують (подавляють) контурні шуми порівнюваних символів при їх накладенні, що в значній мірі впливає на якість класифікації зображень символів.

1. Вступ.

Методи класифікації є вельми перспективним і багатообіцяючим напрямком в теорії і практиці стиску зображень [1-5]. Особливе значення дані методи можуть мати при стиску зображень тексту, які повсюдно використовуються для перекладу друкарської продукції в електронний вигляд.

У даній роботі запропонований і досліджений метод стиску зображень тексту, заснований на виділенні нероздільних символів (букв і знаків пунктуації) і такий їх класифікації, що в кожен клас потрапляють тільки зображення одного і того ж символу. Складність цього завдання викликана шумами, що виникають при друці тексту на папері і подальшому його скануванні.

2. Метод виделення символів та їх класифікація.

Основне завдання, що вирішується приведеним нижче алгоритмом, полягає в наступному:

- розділити всі символи, що входять в зображенні тексту, на класи так, щоб в кожному класі містилися тільки зображення одного символу
- кількість класів має бути мінімальне можливим.

В ідеалі кількість класів дорівнює числу різних символів в зображенні тексту, але це, як правило, недосяжно із-за шумів друку і сканування, тобто спотворень форми символу при друці тексту на папері і подальшому скануванні.

Наявність вказаних шумів приводить до того, що майже всі зображення якого-небудь символу відрізняються один від одного. Отже, знайти однакові зображення символу набагато складніше, ніж різні. Існує ще одна важлива особливість шумів друку і сканування – ці шуми не є структурними, а носять виключно контурний характер, тобто виникають на межах чорних і білих областей зображення тексту. Ефективність пропонованої нижче класифікації великою мірою заснована на врахуванні цієї обставини

Пропонований метод можна умовно розділити на декілька окремо вирішуваних завдань:

1. Виділення із зображення тексту нероздільних символів у вигляді мінімальних прямокутних областей, що містять цей символ;
2. Попередня класифікація отриманих зображень символів за простими ознаками (висота, ширина, повний периметр);
3. Основна процедура – розбиття сукупності всіх зображень нероздільних символів на класи, кожен з яких містить зображення тільки одного символу. Знаходження усередненого «представника» для кожного класу;
4. Створення «графічного словника», що містить сукупність усереднених

«представників» і побудова карти регіонів, яка показує розміщення кожного символу з графічного словника на площині зображення тексту.

3. Класифікуюча метрика з подавленням шумів друку та сканування.

Основна класифікація, проводиться за алгоритмом «просіювання» [6, 7].

При порівнянні двох зображень символів S_1 і S_2 з допустимими відхиленнями по висоті, ширині і периметру (ΔH, ΔW і ΔP) ці зображення накладаються один на одного за допомогою плоскопаралельного перенесення так, щоб їх центри тяжіння збігалися. Вони виражають індивідуальні особливості символів, тобто гарантовано розрізняють, наприклад, такі букви, як «n» і «r». Тому для порівняння використовують безрозмірну, тобто не залежну ні від роздільної здатності сканування, ні від розміру шрифту величину

$$\Delta P = \frac{|P_1 - P_2|}{\sqrt{P_1 P_2}} 100\%, \qquad (1)$$

де P_1 і P_2 – периметри порівнюваних символів. Задовільна попередня класифікація виходить, якщо периметри символів рахувати досить близькими при виконанні умови $\Delta P \le 10\%$.

Далі підраховуються дві величини: $R(S_1, S_2)$ – кількість точок «істотних відмінностей», і $D(S_1, S_2)$ – кількість загальних точок збігу, рис.1.

Перша величина – це кількість неспівпадаючих по яскравості (біле – чорне) крапок, які не є суміжними для сукупності загальних чорних крапок. Таким чином, кількість істотних відмінностей $R(S_1, S_2)$ ігнорує неспівпадання в тих крапках, які лежать на периметрах зображень і, як правило, є шумами друку і сканування. Друга величина - потрібна для втрати розміру першою, щоб діапазон можливих значень величини

$$\varepsilon(S_1, S_2) = \frac{R(S_1, S_2)}{D(S_1, S_2)} 100\% \qquad (2)$$

для всіх пар символів не мінявся при зміні розміру шрифту і роздільній здатності сканування.

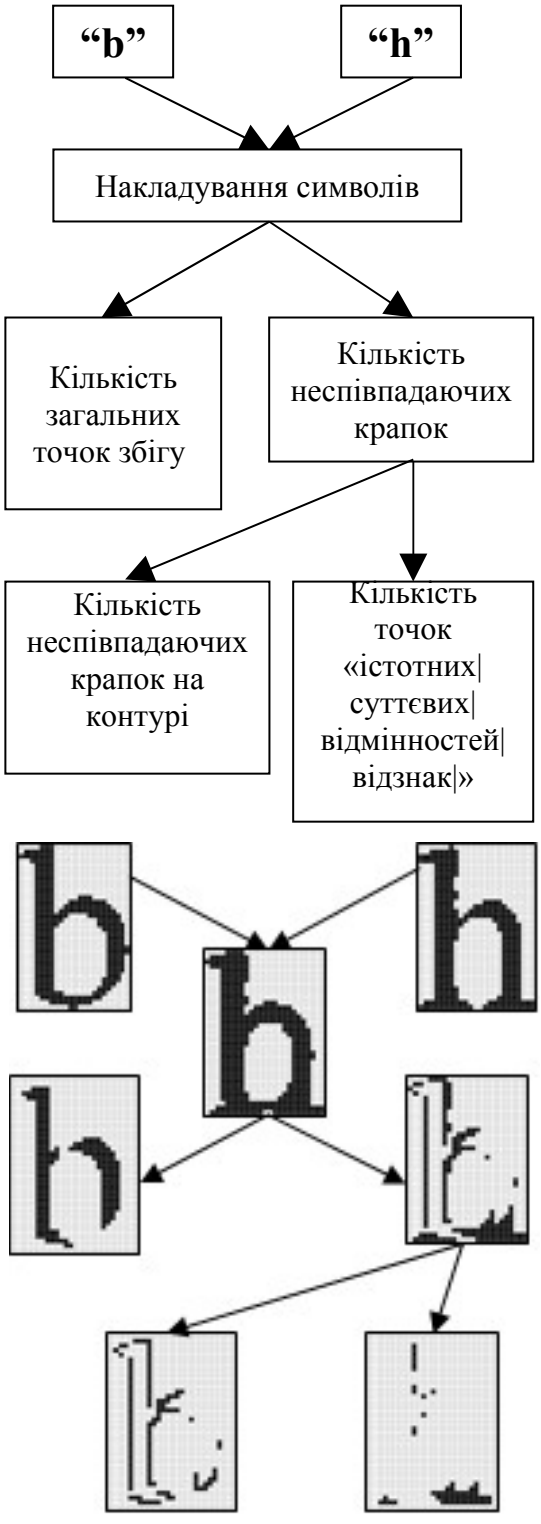

Рис.1. Схема порівняння зображень символів “b” і “h”.

Функція $R(S_1,S_2)$ визначається з урахуванням ваги. Ваговий коефіцієнт кожної крапки в $R(S_1,S_2)$ тим більше, чим більше у даної точки таких же суміжних крапок [8]. Таким чином, пропонована метрика ε (2) що визначає ступінь близькості зображень двох символів при класифікації алгоритмом «просіювання», мало чутлива до шумів друку і сканування. Вона заснована на стабільних характеристиках $R(S_1S_2)$ і $D(S_1,S_2)$ які пригнічують (не враховують) контурні шуми порівнюваних символів при їх накладенні з суміщеними центрами тяжіння.

Після вибору метрики потрібно підібрати таке її порогове значення $\varepsilon_{\max}$ що при виконанні умови $\varepsilon(S_1,S_2)<\varepsilon_{\max}$ можна вважати, що два порівнювані зображення S_1 і S_2 належать одному і тому ж символу. Дуже великі значення параметра $\varepsilon_{\max}$ приведуть до неприпустимих помилок – ототожнюватимуться (потрапляти в один клас) зображення різних символів. Дуже мале значення приведе до того, що деякі зображення одного і того ж символу кваліфікуватимуться як зображення різних символів (і потрапляти в різні класи). В цьому випадку при класифікації виникне велике число класів, що знизить ефективність стиску зображення тексту. Цей діапазон при різних значеннях роздільної здатності зображення тексту дорівнює універсального порогу ε_{opt}=6%. Даний показник визначен експериментальним шляхом.

Порівняння з кращим в даний час спеціальним алгоритмом для стиску зображень тексту – JB2, включеним у формат DjVu, показало, що якість класифікації у пропонованого методу значно вища, ніж у алгоритму JB2. Кількість класів, що виходить в результаті запропонованої класифікації, більш ніж в два рази менше при всіх значеннях роздільної здатності сканування (таблиця 1).

Це є основною якісною характеристикою методу і дає широкі можливості підвищення інформативності цього алгоритму в інженерних реалізаціях.

Таблиця 1.

Роздільна здатність зображення (dpi)	Кількість класів в зображенні оригіналу	Кількість класів пропонованого алгоритму	Кількість класів JB2
600	3558	72	314
500	3557	72	259
400	3557	71	199
300	3545	95	235
200	3890	148	451

В таблиці 2 приведені кількісні характеристики приведеного метода, що до вихідного розміру файла при застосуванні різних форматів представлення символьних зображень.

Таблиця 2.

4. Висновки.

Використовуючи контурний характер шумів, що вносяться до зображень символів (букв і розділових знаків) при друці і подальшому скануванні, в роботі запропонована нова метрика в просторі чорно-білих зображень символів, яка дозволяє добре розрізняти ці символи навіть за наявності сильних шумів друку і сканування. Відповідна класифікація зображень символів, що використовує цю метрику, дає число класів, близьке до мінімально можливого. Це дозволило створити ефективний алгоритм компресії зображень тексту, заснований на виділенні зображення символів і подальшої їх класифікації.

Приведений метод в порівнянні з спеціальним алгоритмом стиску зображень тексту – JB2, що включений у формат DjVu, показав більш якісну класифікації. Це дало можливість підвищити ступінь компресії символьних зображень ніж у алгоритму JB2 (табл. 2). Кількість класів, що виходить в результаті запропонованої класифікації, в двічі меньше при всіх значеннях роздільної здатності зображення тексту (див. табл. 1).

Реалізований алгоритм дозволяє зменшити розміри вихідних даних, в порівнянні з алгоритмом JB2 при всіх значеннях роздільної здатності зображень тексту (від 8% до 28,6% при відповідних значеннях dpi, табл. 2), що в середньому складає близько 20%.

5. Література

1. Земсков В.Н. Сжатие изображений на основе автоматической классификации [Текст] / В.Н. Земсков, И.С. Ким // Известия вузов. Электроника. – 2003. – № 2. – С. 50-56.
2. Gupta Maya R., Stroilov A. Segmenting for wavelet compression [Электронный ресурс]: Data Compression Conference, 2005. Proceedings. DCC 2005, 29-31 March 2005, USA, Utah, Snowbird. – 462 p. - Режим доступа: http://www.computer.org/portal/web/csdl/proceedings/ - 10.04.2010 г.
3. Иванов В.Г. Сокращение содержательной избыточности изображений на основе классификации объектов и фона [Текст] / В.Г. Иванов, М.Г. Любарский, Ю.В. Ломоносов // Проблемы управления и информатики. – 2007. – № 3. – С. 93-102.
4. Иванов В.Г. Сжатие изображений на основе автоматической и нечеткой классификации фрагментов [Текст] / В.Г. Иванов, Ю.В. Ломоносов, М.Г. Любарский // Проблемы управления и информатики. – 2009. – №1 – С. 52-63.
5. Іванов В.Г., Любарський М.Г., Ломоносов Ю.В. Стиск зображень на основі виділення і кодування об'єктів з різною візуальною якістю // Праці Восьмої Всеукраїнської міжнародної конференції “Оброблення сигналів і зображень та розпізнавання образів” (УкрОБРАЗ‘2006). – Київ: Міжнародний науково-навчальний

Роздільна здатність сканування (dpi)	200	300	400	500	600
Початковий розмір файлу (kb)	505,3	1080,2	2003,9	3111,2	4498,0
Методи	**Розмір файлу після стиску (kb)**				
Формат JPEG 2000	132,8	288,6	532,4	830,0	1200,3
Формат PDF	61,4	96,1	119,6	148,9	178,9
JB2 у форматі DjVu	9,6	8,7	9,9	11,4	13,6
пропонований алгоритм	8,1	8,0	8,0	8,5	9,7

центр інформаційних технологій та систем, 2006. – С. 159-163.
6. Прикладная статистика: Классификация и снижение размерности [Текст]: справочник / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков и др.; под общ. ред. С.А. Айвазяна.– М.: Финансы и статистика, 1989. – 607 с.
7. Іванов В.Г., Ломоносов Ю.В., Любарський М.Г., Стиснення зображень на основі класифікації і декорелюючих перетворень // Праці Дев′ятої Всеукраїнської міжнародної конференції “Оброблення сигналів і зображень та розпізнавання образів” (УкрОБРАЗ‘2008). – Київ: Міжнародний науково-навчальний центр інформаційних технологій та систем, 2008. – С. 159-163.
8. Прэт У.К. Комбинированная система сжатия факсимильных данных с подбором символов [Текст] / У. К.. Прэт, П. Дж. Капитан, Чжань Вэнсюнь, Э. Р. Хамилтон, Р.Х. Уоллис // Цифровое кодирование графики. Тематический выпуск. ТИИЭР. – 1980. – Т. 68, № 7. – С. 40-49.