УДК 519.68

В. М. ГЛАДКИХ, канд. техн. наук,
Одеська національна академія зв'язку ім. О. С. Попова

МЕТОД КЛАСИФІКАЦІЇ ЦИФРОВИХ ЗОБРАЖЕНЬ ДОКУМЕНТІВ СУВОРОЇ ЗВІТНОСТІ ЗА КОЛІРНИМ КОНТЕНТОМ

Розроблено метод класифікації зображень документів суворої звітності для систем їх автоматизованого розпізнавання та оброблення. Метод ґрунтується на аналізі власних значень коваріаційної матриці, розрахованої для компонентів колірного простору RGB, де R,G,B ∈ [0,225]. Виділено три класи зображень і показано, що запропонований метод класифікації не залежить від інтенсивності спотворень, обраної роздільної здатності та типу сканера.

Ключові слова: класифікація; коваріаційна матриця; власні значення.

Інтенсивність досліджень систем оптичного розпізнавання (СОР), незважаючи на значні успіхи в їх розробці й практичному використанні та високу точність розпізнавання текстів [1], не зменшується, оскільки потреба в удосконаленні й розширенні функціональності цих систем постає дедалі гостріше, передусім через стрімке зростання сфери автоматизації документообігу та переведення інформації, що зберігається на паперових носіях, в електронну форму.

У загальному випадку СОР тих чи інших символів складається з компонентів, що виконують такі дії:

- попереднє оброблення зображення, поданого у градаціях сірого;
- сегментацію — виділення текстових регіонів;
- бінаризацію зображення;
- виділення рядків слів та символів;
- розпізнавання символів;
- виправлення помилок розпізнавання.

Що ж до конкретної галузі застосування, то структура СОР залежить насамперед від пристрою, використовуваного для отримання цифрового образу документа (відеокамера, мобільний телефон, сканер тощо); типу документа — структурований, слабоструктурований, не структурований, архівний тощо, друкований або рукописний.

Одне з актуальних завдань у сфері електронного документообігу полягає в автоматизації розпізнавання та оброблення документів суворої звітності. І хоча документ суворої звітності — це, як правило, структурований документ, усе ж виділення тексту для його подальшого розпізнавання вимагає розроблення специфічних методів. Використання традиційних методів, таких як гістограмний, перетворення Радона або Хафа, не дає належного ефекту, особливо стосовно виділення напису суми, яка подається словами, та інших захищених від підробки написів. Окрім того ці методи не сприяють скороченню часу оброблення документа.

Бланки документа можуть бути надруковані чорною або кольоровою друкарською фарбою. Заповнювати бланк також можна чорним або кольоровим чорнилом. З огляду на те, що заповнені бланки мають різний колірний контент, вважається за доцільне для кожного типу колірного контенту розробити спеціалізовані (нескладні в обчислювальному плані) методи щодо виділення тексту.

Наприклад, якщо бланк надруковано чорною друкарською фарбою, а напис виконано кольоровим чорнилом, то у просторі кольорів можна виділити два кластери — ахроматичний та хроматичний. Останній являє собою текст, який далі доведеться розпізнавати. Для того щоб реалізувати цей підхід, необхідно розробити метод автоматичного визначення типу колірного контенту переказу.

Мета цієї статті — дослідження колірного контенту зображень документів, отриманих зі сканера в повнокольоровому режимі, та розроблення методу автоматичної класифікації цих зображень за їхнім колірним контентом.

Для того щоб суттєво спростити обчислювальну складність подальшого оброблення документа суворої звітності, особливо стосовно виділення тексту, необхідно встановити, до якого з трьох зазначених далі типів належить заповнений документ:

1) чорно-біле тло, рукописний текст подано чорним кольором;

2) чорно-біле тло, рукописний текст — кольоровий;

3) бланк із кольоровим тлом, рукописний текст — чорний або кольоровий.

Розглянемо головні особливості сканування бланків документів суворої звітності в повнокольоровому режимі.

У процесі сканування в зображення документа вносяться спотворення кольорів. Інтенсивність цих спотворень залежить від спектральних характеристик лампи сканера, чутливості та шумових характеристик сканувальної головки, обраної роздільної здатності, а також від характеристик оригіналу. Бланк документа може бути надрукований на папері з різним ступенем білизни. Так само й друкарські фарби, якими надруковано бланк,

також можуть мати різні відбивальні властивості та вносити зміни у спектр відбитого світла. Усі ці чинники становлять непрогнозоване джерело спотворення кольорів при скануванні. Окрім того, спотворення кольорів можуть виникати й за рахунок розсіювання світла лампи на межі, наприклад, між областю з чорним і областю з білим кольором. У такому разі спостерігається розмивання межі та спотворення кольорів навіть за достатньо високої роздільної здатності.

Якщо проаналізувати колірний контент зображення поштового переказу першого типу, отриманого за роздільної здатності 200 dpi, то хоча таке зображення візуально сприймається як ахроматичне, насправді кількість кольорів у ньому значно перевищує 255 і становить 34 тис. для зображення, розташованого ліворуч, і 76 тис. для зображення, розташованого праворуч. Із підвищенням роздільної здатності до 300 або 600 dpi кількість кольорів, а отже інтенсивність спотворень лише збільшується.

Як показує практика, інтенсивність спотворень найбільша на межах тексту та на межах надрукованих атрибутів бланка поштового переказу.

З огляду на це постає потреба розробити такий метод, із використанням якого можна точно класифікувати зображення поштового переказу незалежно від інтенсивності спотворень кольору.

У просторі кольорів RGB зображення побудуємо коваріаційну матрицю C, елементи якої обчислюються за формулами

$$c_{pp} = \frac{1}{N}\sum_{i=1}^{N} p_i^2 - \overline{p}^2, \quad c_{pq} = \frac{1}{N}\sum_{i=1}^{N} p_i\, q_i - \overline{pq}.$$

Тут $p, q \in \{R,G,B\}$ — значення інтенсивності компонентів простору RGB ($R,G,B \in [0,225]$), $N = n \times m$ — кількість пікселів у зображенні (n, m — розміри зображення).

Власні значення квадратної матриці знаходять, розв'язуючи характеристичне рівняння

$$\det(C - \lambda I) = 0,$$

де I — одинична матриця; λ — власні значення.

Коваріаційна матриця простору кольорів RGB має розмір 3×3, а тому характеристичне рівняння в розгорнутому вигляді являє собою кубічне рівняння виду

$$\lambda^3 - \mathrm{tr}(C)\lambda^2 + (M_{RR} + M_{GG} + M_{BB})\lambda - \det(C) = 0, \quad (1)$$

де M_{pp} — мінори другого порядку коваріаційної матриці.

Для спрощення подальшого аналізу будемо використовувати коваріаційну матрицю з елементами, що обчислюються за формулою

$$\tilde{c}_{pq} = \frac{c_{pq}}{\mathrm{tr}(C)}, \quad (2)$$

де $\mathrm{tr}(C)$ — слід матриці, $\mathrm{tr}(C) = \sum_p c_{pp}$. За такої умови сума власних значень $\sum_i \lambda_i = 1$.

Відомо, що для зображення в градаціях сірого для кожного піксела значення інтенсивності компонентів простору RGB однакові, тобто $R = G = B$. Згідно з цією умовою дістаємо:

$$\tilde{c}_{pp} = \tilde{c}_{pq} = \frac{1}{3}, \ \forall p,q \in \{R,G,B\}.$$

Тоді рівняння (1) набирає вигляду

$$\lambda^2(1-\lambda) = 0. \quad (3)$$

Звідси $\lambda_1 = 1$ і $\lambda_{2,3} = 0$.

Отже, для зображення у градаціях сірого лише одне власне значення коваріаційної матриці відмінне від нуля.

Розглянемо випадок, коли бланк поштового переказу заповнено кольоровим, наприклад червоним, чорнилом. Тоді коваріаційна матриця подається так:

$$C = \begin{pmatrix} c_{RR} & c & c \\ c & \frac{1-c_{RR}}{2} & \frac{1-c_{RR}}{2} \\ c & \frac{1-c_{RR}}{2} & \frac{1-c_{RR}}{2} \end{pmatrix}. \quad (4)$$

Характеристичне рівняння цієї матриці

$$\lambda(\lambda^2 - \lambda + (M_{GG} + M_{BB})) = 0$$

має дійсні корені

$$\lambda_{1,2} = \frac{1}{2} \pm \sqrt{\frac{1}{4} - 2M_{GG}}, \ \lambda_3 = 0.$$

Тут

$$M_{GG} = M_{BB} = \frac{1}{2}c_{RR}(1-c_{RR}) - c^2 < 0.$$

Зображення поштового переказу, отримане зі сканера, містить візуально непомітні спотворення кольорів. Коваріаційну матрицю в цьому разі можна записати у вигляді

$$C_\varepsilon = C + \delta C,$$

де $\delta C = (\pm\varepsilon_{pq})$ — матриця збурень, зумовлених спотворенням кольорів при скануванні. Як відомо [2], власні значення неперевно залежать від значень елементів коваріаційної матриці. Через це збурення елементів коваріаційної матриці призведе до збурення власних значень. Оцінку збурення власного значення λ_p можна знайти зі співвідношення

$$\left|\delta\lambda_p\right| \le \frac{\left\|C_\varepsilon - C\right\|}{\left|s_p\right|},$$

де s_p — косинус кута між власними векторами збуреної C_ε та незбуреної коваріаційної матриці C, що відповідають власному значенню λ_p.

Оскільки коваріаційна матриця симетрична, то в цьому разі $s_p = 1$, а отже,

$$\left|\delta\lambda_p\right| \le \left\|C_\varepsilon - C\right\| = \left\|\delta C\right\|.$$

У цьому співвідношенні $\|\cdot\|$ — матрична норма.

Нехай тепер

$$\left\|\delta C\right\| = \max_p\left(\sum_k \varepsilon_{kp}\right) < 3\max_k(\varepsilon_{kp}).$$

Тоді

$$\left|\delta\lambda_p\right| \le 3\max_k(\varepsilon_{kr}),$$

де $r = \arg\max_p\left(\sum_k \varepsilon_{kp}\right)$.

Для аналізу збурень власних значень використовувалась база зі 100 зображень бланків поштових переказів у градаціях сірого та 100 зображень бланків у градаціях сірого з написом кольоровим чорнилом, а також додатково 50 зображень поштових переказів, надрукованих кольоровою друкарською фарбою. Половину зображень було отримано на сканері HP Scanjet 4500 c, а решту — на сканері Mustek BearPaw 1200 з роздільними здатностями 200, 300 і 600 dpi. За результатами аналізу встановлено, що збурення власних значень для бланків першого та другого типів за порядком такі: $\left|\delta\lambda_p\right| \approx O\left(10^{-4}\right)$, причому незалежно від типу сканера, властивостей оригіналу, інтенсивності спотворень, а також використаної роздільної здатності.

Окремі результати обчислення власних значень для різних зображень поштових переказів наведено в таблиці. Власні значення для зображень, що візуально сприймаються як ахроматичні, у таблиці не наведено, оскільки в усіх випадках найбільше з них дорівнює одиниці, незалежно від властивостей оригіналу та умов сканування.

Власні значення коваріаційної матриці бланків із різним колірним контентом

Кольоровий			У градаціях сірого, напис кольоровий		
λ_1	λ_2	λ_3	λ_1	λ_2	λ_3
0,94	0,057	0,003	0,996	0,004	0
0,957	0,04	0,003	0,993	0,007	0
0,913	0,086	0,001	0,994	0,006	0
0,917	0,078	0,005	0,995	0,005	0
0,92	0,076	0,004	0,997	0,003	0

Таким чином, метод класифікації зображень документів суворої звітності згідно з колірним контентом передбачає такі дії.

1. Обчислити елементи коваріаційної матриці та нормувати їх, скориставшись формулою (2).

2. Обчислити власні значення коваріаційної матриці, згідно з рівнянням (1) і округлити їх до третьої значущої цифри після коми, записавши в порядку спадання: $\lambda_1 > \lambda_2 > \lambda_3$.

3. Проаналізувати множину власних значень:

а) якщо $\lambda_1 = 1$, то колірний контент зображення заповненого бланка документа суворої звітності — переважно у градаціях сірого, а зображення належить до першого класу;

б) якщо $\lambda_1 + \lambda_2 = 1$, то колірний контент зображення документа — бланк у градаціях сірого, напис кольоровий і зображення належить до третього класу;

в) якщо всі власні значення відмінні від нуля, то колірний контент — кольоровий бланк, напис чорний або кольоровий і зображення належить третьому класу.

Висновки

1. Інтенсивність спотворення кольорів оригіналу документа суворої звітності при скануванні найбільша на межах областей з різною яскравістю. Із підвищенням роздільної здатності кількість спотворень лише зростає.

2. Власні значення коваріаційної матриці відбивають колірний контент зображення документа суворої звітності.

3. Запропонований метод класифікації зображень документів суворої звітності за колірним контентом дає змогу на підставі аналізу власних значень коваріаційної матриці забезпечити чітку класифікацію, незалежно від інтенсивності спотворень, типу сканера та роздільної здатності.

4. Класифікацію зображень документів суворої звітності за колірним контентом можна використати як процедуру попереднього оброблення системи автоматичного розпізнавання поштових переказів.

Подальші дослідження буде спрямовано на розробку методів виділення тексту для кожного класу зображень окремо.

Література

1. ***ABBYY*** *FineReader 10 / [Електронний ресурс] — Режим доступу: http://www.abbyy.ru/finereader/*
2. ***Улкинсон, Дж. Х.*** *Алгебраическая проблема собственных значений / Дж. Х. Улкинсон.— М.: Наука, 1970.— 564 с.*

В. Н. Гладких

МЕТОД КЛАССИФИКАЦИИ ЦИФРОВЫХ ИЗОБРАЖЕНИЙ ДОКУМЕНТОВ СТРОГОГО УЧЕТА ПО ЦВЕТОВОМУ КОНТЕНТУ

Разработан метод классификации изображений документов строгого учета для систем их автоматизированного распознавания и обработки. Метод основан на анализе собственных значений ковариационной матрицы, рассчитанной для компонент цветового пространства RGB, где R, G, B∈ [0,225]. Выделено три класса изображений и показано, что предложенный метод не зависит от интенсивности искажений, выбранной разрешающей способности и типа сканера.

Ключевые слова: классификация; ковариационная матрица; собственные значения.

V. M. Gladkyh

DIGITAL IMAGE CLASSIFICATION METHOD OF ACCOUNTING DOCUMENTS FOR COLOR CONTENT

The method of digital image classification of document for the system of postal orders automatic recognition and processing was developed. Proposed method is based on RGB, R, G, B∈ [0,225] and components of covariance matrix of eigenvalues analysis. There were three classes of images and it was shown that proposed method does not depend on the intensity of distortion, chosen resolution and scanner type.

Keywords: classification; covariance matrix; eigenvalues.