УДК 621.3 **doi: 10.26906/SUNZ.2022.1.043**

О. Ю. Барковська, В. О. Холєв, Д. А. Поліканов

Харківський національний університет радіоелектроніки, Харків, Україна

ЗНАЧУЩІСТЬ ОБЧИСЛЮВАЛЬНИХ СИСТЕМ ІЗ МАСОВИМ ПАРАЛЕЛІЗМОМ ПРИ ОБРОБЦІ СКАНОВАНИХ ДОКУМЕНТІВ

Анотація. У роботі запропоновано узагальнена модель системи класифікації сканованих документів, яка являє організаційно-функціональний, технологічний і програмно-технічний комплекс для класифікації або категоризації документу за ключовими словами, які визначаються частотним словником. Актуальність теми дослідження полягає у скороченні часу впорядкування нових інформаційних ресурсів, що надходять до сховища, завдяки збільшенню швидкості роботи **методів** покращення якості вихідного зображення безпосередньо перед обробкою та аналізом тексту. **Аналіз результатів** довів ефективність та доцільність використання обчислювачів із масовим паралелізмом для виконання таких задач, як шумопригнічення та зміна значення колірних каналів вихідного повнокольорового зображення, досягаючи прискорення до 53,51% у порівнянні із використанням обчислювального ресурсу центрального процесору.

Ключові слова: система, обробка, текст, обчислювач, графічний процесор, багатопоточність, зображення, передобробка, прискорення, інформаційний ресурс, сховище.

Вступ

Організація електронних бібліотек [1, 2], архівів [3, 4], сховищ електронних інформаційних ресурсів різних масштабів [5, 6] (від традиційних сховищ на підприємствах або персональних сховищ розміром до терабайту, до великих хмарних сховищ, які мають власну універсальну архітектуру) стикаються із процесом безперервного поповнення новими ресурсами (із різною тематичною приналежністю - наукові ресурси, соціальні ресурси, екологічні ресурси, законодавчі ресурси, нормативно-правові ресурси, статистичні ресурси, навчальні ресурси тощо), що, в свою чергу, потребує впорядкування (за автором, назвою, роком видання, видавництвом чи ключовими словами) для полегшення подальшого користування. Подібний процес «накопичення – аналіз – зберігання - доступ користувачів» є звичним для різного типу сховищ, які зустрічаються зараз та стосується усіх категорій інформаційних ресурсів (рис. 1). Задача впорядкування документів передбачає використання методів text processing та text mining, особливістю яких є відсутність структури у вихідних даних (текст) [7, 8]. Дослідження в цій сфері є актуальними на сьогоднішній день, про що свідчить широке використання отриманих результатів у багатьох проблемних галузях: системи категоризації текстових інформаційних ресурсів, пошукові системи тощо [9-11]. Однак, такі вимоги, як скорочення часу обробки та збільшення точності аналізу тексту, є затребуваними та висуваються безперервно незалежно від області практичного застосування.

Особливої уваги вимагає проблема обробки та аналізу тексту, який надходить до системи у вигляді не текстового вихідного файлу, наприклад, у вигляді фото або сканованих документів. Так, при додаванні до електронного архіву старих сканованих друкованих книг або інших друкованих видань, організація автоматичного впорядкування є більш складною процедурою, оскільки передбачає такий етап попередньої обробки документу, як переведення зображення у текстовий вигляд, а вже потім безпосередня робота з текстом [12]. **Аналіз існуючих рішень** показав, що системи розпізнавання тексту [13] застосовують у багатьох областях, наприклад: зчитування даних з бланків, анкет чи білетів; автоматичне розпізнавання автомобільного номера; розпізнавання паспортних даних; вилучення інформації з візитних карток до списку контактів; технологія для допомоги сліпим і слабозорим; оцифрування архівних документів чи старих книг. Критеріями оцінювання існуючих систем, серед яких були розглянуті найпотужніші рішення - ABBYY Finereader, ABBYY Flexicapture, Adobe Acrobat DC, було обрано точність розпізнавання, доступність, легкість використання. Результат аналізу наведено у табл. 1.

Таблиця 1 – **Аналіз існуючих аналогів у сфері текстового процесінгу**

	Швидкість	**Доступність**	**Легкість використання**
ABBYY Finereader	Задовільна	Висока ціна	Потрібен час для навчання
ABBYY Flexicapture	Задовільна	Висока ціна	Важко налаштувати
Adobe Acrobat DC	Задовільна	Висока ціна	Простий у використанні

Таким чином, виконання досліджень із метою досягнення скорочення часу обробки та аналізу ска-

Рис. 1. Категоризація інформаційних ресурсів

нованих документів для організації впорядкованого сховища текстових електронних інформаційних ресурсів є задачею **актуальною**.

Мета та задачі дослідження

Метою дослідження є оцінка значущості обчислювальних систем із масовим паралелізмом при обробці сканованих документів, а саме при вирішенні задачі шумопригнічення та фільтрації сканованих документів.

Для досягнення поставленої мети мають бути вирішені наступні задачі:

– дослідження загальних параметрів систем розпізнавання сканованих документів;

– дослідження існуючих методів шумопригнічення та сегментації зображень;

– розробка узагальненої моделі класифікації сканованих документів;

– адаптація методів попередньої обробки сканованих документів під обчислювачі із масовим паралелізмом;

– оцінка таймінгу роботи;

– аналіз отриманих результатів.

Дослідження, проведені в рамках даної роботи, зосереджені на скороченні часу попередньої обробки вихідних сканованих документів безпосередньо перед процесінгом тексту.

Рішення поставленої задачі

Актуальність теми дослідження полягає у скороченні часу впорядкування нових інформаційних ресурсів, що надходять до сховища, завдяки збільшенню швидкості роботи методів покращення якості вихідного зображення безпосередньо перед обробкою та аналізом тексту.

У роботі запропоновано узагальнену модель системи класифікації сканованих документів (рис. 2), яка являє організаційно-функціональний, технологічний і програмно-технічний комплекс для класифікації або категоризації документу за ключовими словами, які визначаються частотним словником. Запропонована узагальнена модель складається з блоків:

– попередньої обробки зображення;

– розпізнавання тексту;

– попередньої обробки тексту [14, 15];

– побудови частотного словника.

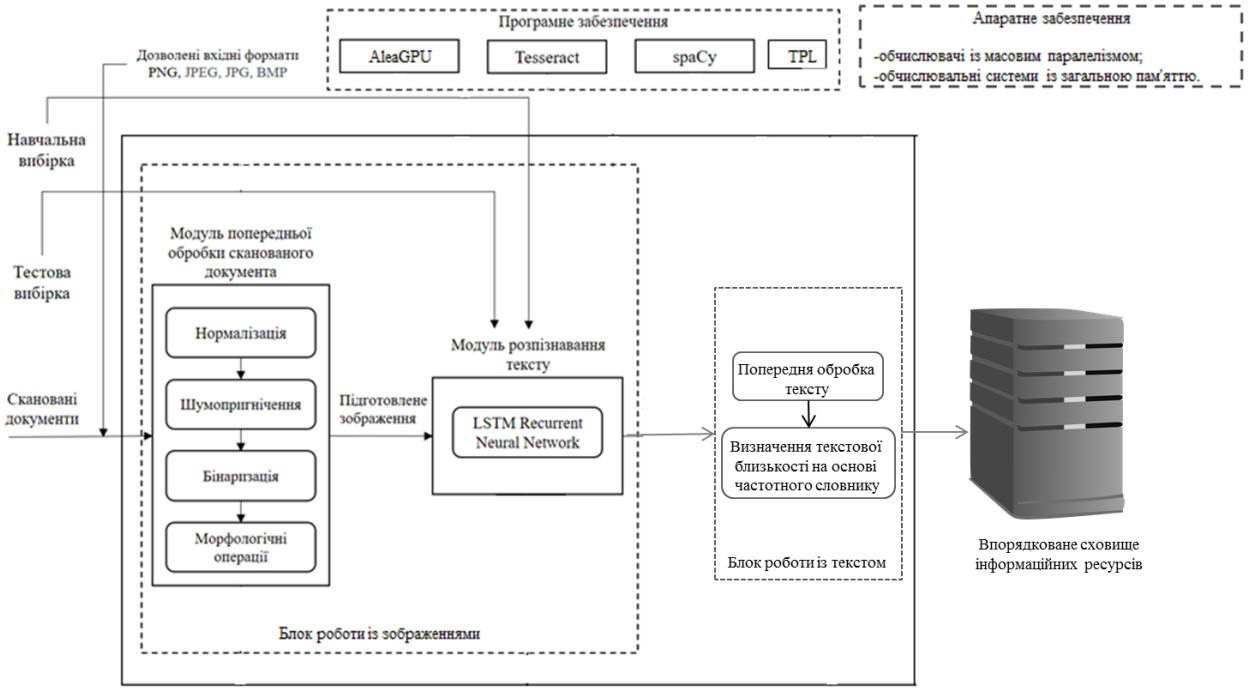


Рис. 2. Узагальнена модель системи класифікації сканованих документів

При цьому, скорочення часу роботи системи можливо як завдяки розпаралелюванню етапу роботи із зображенням, так і етапу роботи з текстом. Робота кожного з зазначених блоків складається з роботи окремих алгоритмів та методів, деякі з яких розглянуті в роботі та виконано їх вдосконалення.

На вхід системи подається сканований документ, який має бути нормалізований, очищений від шумів та бінарізований у модулі попередньої обробки зображення. Запропонована сукупність методів попередньої обробки сканованих документів дозволяє покращити вихідне зображення, що призводить до збільшення точності розпізнавання тексту.

Результатом роботи є побудований частотний словник із визначеною та впорядкованою TF-IDF мірою для зниження значущості слів які зустрічаються різних документах колекції, який подається на вхід класифікатора. Подальша робота полягає у аналізі слів із високим значенням параметру TF-IDF, оскільки слова, у яких значення TF-IDF близьке до нуля, відносяться до загальноуживаних слів, які не можуть дати об'єктивну характеристику документу.

Робота модулю попередньої обробки сканованого документу зосереджена на покращенні якості вихідного зображення перед подачею на модуль розпізнавання тексту. Обробка зображень має високі об-

числювальні вимоги. Низку операцій обробки зображень можна оптимізувати шляхом інтенсивного розпаралелювання обчислень. В даний час одним з найкращих варіантів паралельної обробки зображень є використання обчислень загального призначення на графічних процесорах (GPGPU).

Запропоновані методи досліджені за допомогою професійного стеку розробки CUDA для .NET - Alea GPU, яка дозволяє використовувати повний набір функцій пристрою CUDA, наданих NVIDIA LibDevice, а також паралельні внутрішні функції пристрою CUDA, такі як синхронізація потоків, атомарні операції тощо.

Alea GPU — це повноцінний компілятор, побудований на основі популярної інфраструктури компілятора LLVM і пакета SDK компілятора NVIDIA CUDA.

Код, зібраний за допомогою компілятора Alea GPU для графічного процесора, забезпечує таку ж продуктивність, як і еквівалент CUDA C/C++ або CUDA Fortran.

Використовуючи спеціальну бібліотеку SixLabors.ImageSharp.PixelFormats, виконується шумопригнічення та зміна значення колірних каналів, спочатку із багатопоточністю процесора, а потім використовуючи потоки відеоадаптера.

Для визначення продуктивності були використані три групи зображень із різною роздільною здатністю:

– перша група - зображення високої роздільної здатності (від 3336x2823 до 5893x5597);

– друга група - зображення середньої роздільної здатності (від 1538x864 до 1920x1080);

– третя група - зображення низької роздільної здатності (від 720x539 до 1280x720).

Експеримент проводився на різній кількості зображень (1, 10, 40) при пакетній обробці для того щоб показати ефективність використання потоків графічного обчислювача у залежності від обчислювальної інтенсивності задачі.

Апаратною базою для виконання експерименту є відеоадаптер GEFORCE GTX 1050 TI із графічним процесором GP107 на базі архітектури Pascal:

(кількість кластерів потокових текстур – 6,
кількість потокових мультипроцесорів – 6,
кількість потокових процесорів – 768).

Приклади роботи усіх запропонованих та досліджуваних методів зображено на рис. 3.

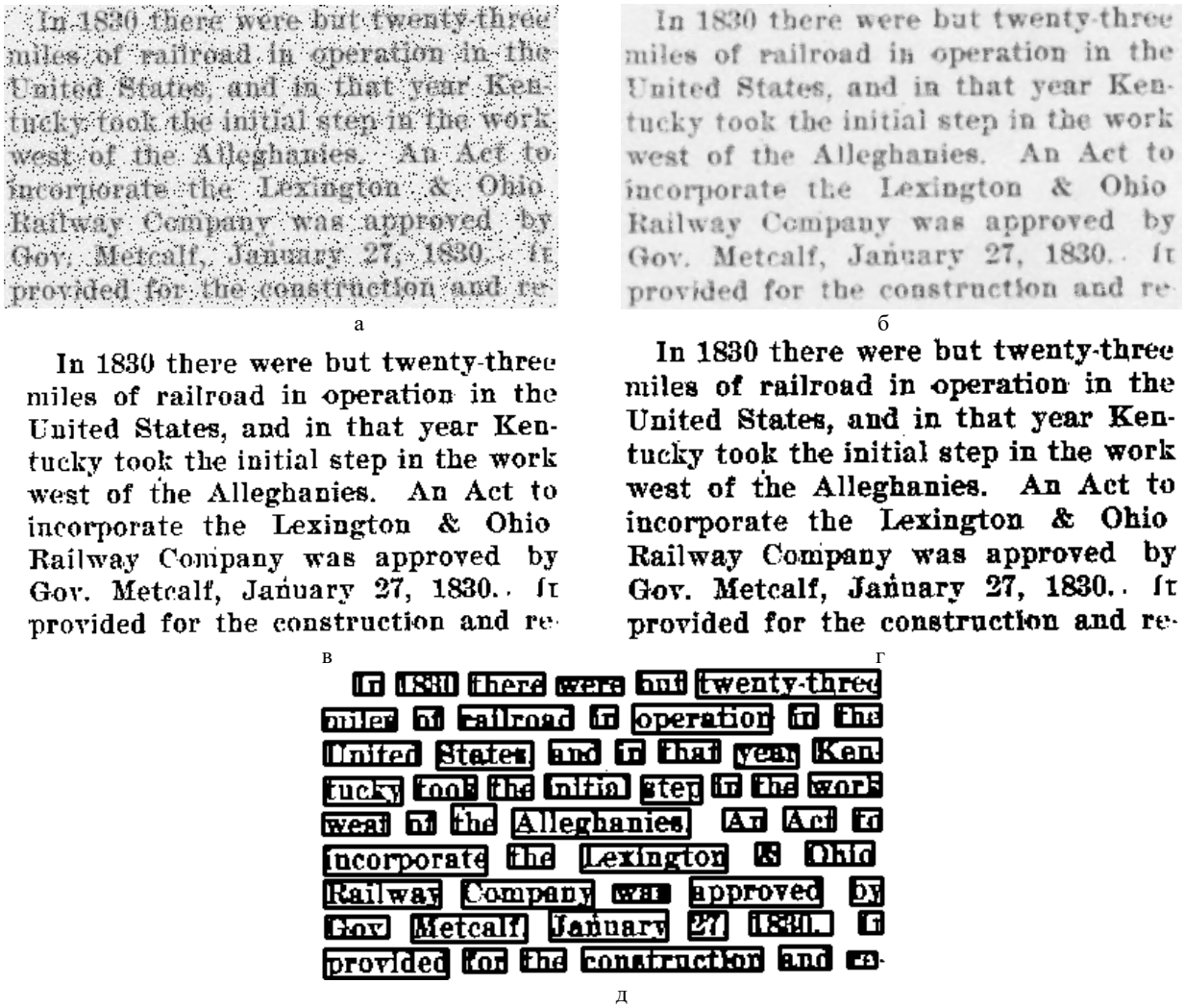


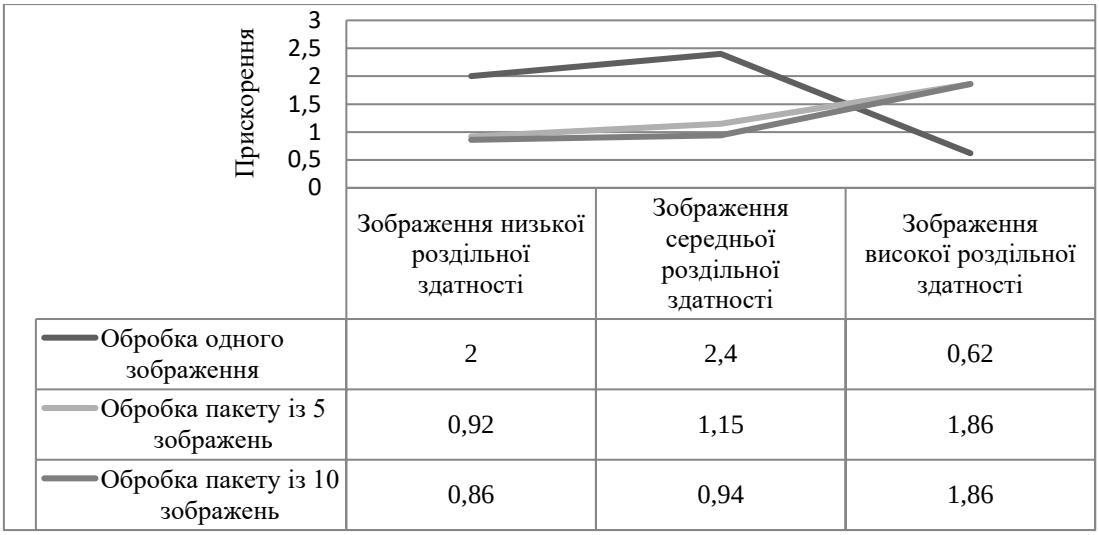
Рис. 3. Результати роботи модулю попередньої обробки вихідного зображення:
а – вихідне зображення, б – результат застосування медіанного фільтру,
в – результат бінарізації, г – результат морфологічних операцій, д – результат сегментації

При бінарізації методом Оцу, пікселі розділяються на два класи, за розрахованим порогом, при якому досягається мінімальне значення дисперсії. Для зменшення кількості можливих артефактів та збільшенню чіткості тексту до зображення застосовуються морфологічні операції.

Аналіз отриманих результатів (табл. 2, рис. 4) показав, що для зображень високої роздільної здатності потоки процесору із масовим паралелізмом працюють на 53,51% швидше, ніж потоки центрального процесору при найвищому досліджуваному навантаженні при обробці пакету із 10 зображень.

Таблиця 2 – **Визначення часу процесінгу зображень**

	Час обробки одного зображення, *мс*			Час обробки п'яти зображень, *мс*			Час обробки десяти зображень, *мс*		
	Перша група	Друга група	Третя група	Перша група	Друга група	Третя група	Перша група	Друга група	Третя група
Багатопоточність, забезпечена центральним процесором	0.006	0.012	0.109	0.024	0.053	0.626	0.043	0.101	0.981
Багатопоточність, забезпечена графічним процесором	0.003	0.005	0.175	0.026	0.046	0.336	0.05	0.107	0.525



	Зображення низької роздільної здатності	Зображення середньої роздільної здатності	Зображення високої роздільної здатності
Обробка одного зображення	2	2,4	0,62
Обробка пакету із 5 зображень	0,92	1,15	1,86
Обробка пакету із 10 зображень	0,86	0,94	1,86

Рис. 4. Прискорення, отримане при паралельній обробці зображень із різною роздільною здатністю на обчислювальних системах із загальною пам'яттю та системах із масовим паралелізмом

Окрім того, результати показують, що збільшення кількості зображень із одного до десяти, при обробці вихідних зображень із високою роздільною здатністю на системах із масовим паралелізмом призводить до збільшення часу обробки лише у 3 рази, що свідчить про особливості вирішення задач загального призначення на графічних процесорах – час на копіювання даних із хосту на пристрій та у зворотньому напрямку не може буди вищим за час, витрачений на виконання обчислень. Так само, як і обробка одного зображення із третьої групи на графічному процесорі не є доцільною.

Вищенаведені результати підтверджують доцільність використання масивно-паралельних обчислювачів для рішення таких задач, як:

фільтрація,

бінарізація,

виконання морфологічних операцій,

сегментація вихідних повно кольорових зображень різної роздільної здатності.

Висновки

При дослідженні загальних параметрів існуючих систем розпізнавання сканованих документів (ABBYY Finereader, ABBYY Flexicapture, Adobe Acrobat DC) було виявлено такі недоліки, як недостатня швидкість отримання розпізнаного тексту, а також велика вартість застосунків, що доводить необхідність досліджень у даній галузі із метою скорочення часу впорядкування нових інформаційних ресурсів, що надходять до сховища, завдяки збільшенню швидкості роботи методів покращення якості вихідного зображення безпосередньо перед обробкою та аналізом тексту.

Запропонована у роботі узагальнена модель системи класифікації сканованих документів, являє собою організаційно-функціональний, технологічний і програмно-технічний комплекс для класифікації або категоризації документу за ключовими словами, визначеними у частотному словнику.

Аналіз отриманих результатів (рис. 4) показав, що для зображень високої роздільної здатності потоки процесору із масовим паралелізмом працюють на 53,51% швидше, ніж потоки центрального процесору при найвищому досліджуваному навантаженні при обробці пакету із 10 зображень. Окрім того, результати показують, що збільшення кількості зображень із одного до десяти, при обробці вихідних зображень із високою роздільною здатністю на системах із масовим паралелізмом призводить до збільшення часу обробки лише у 3 рази, що говорить про доцільність використання обчислювачів із масовим паралелізмом для пакетної обробки.

СПИСОК ЛІТЕРАТУРИ

1. Rusyn B., Lytvyn V., Vysotska V., Emmerich M., Pohreliuk L. (2019) The Virtual Library System Design and Development. In: Shakhovska N., Medykovskyy M. (eds) Advances in Intelligent Systems and Computing III. CSIT 2018. Advances in Intelligent Systems and Computing, vol 871. Springer, Cham. https://doi.org/10.1007/978-3-030-01069-0_24.
2. Cao, G., Liang, M., & Li, X. (2018). How to make the library smart? The conceptualization of the smart library. Electron. Libr., 36, 811-825.
3. T. Hermawan and R. W. Wardhani, "Implementation AES with digital signature for secure web-based electronic archive," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), 2016, pp. 1-6, doi: 10.1109/ICITEED.2016.7863268.
4. Y. Wang, "Design and Implementation of Electronic Archives Information Management Under Cloud Computing Platform," 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2019, pp. 154-158, doi: 10.1109/ICMTMA.2019.00041.
5. Y. Yang and J. Shieh, "Data Warehouse Applications in Libraries -- The Development of Library Management Reports," 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2016, pp. 88-91, doi: 10.1109/IIAI-AAI.2016.129.
6. S. Savanur and K. S. Shreedhara, "Automated data validation for data warehouse testing," 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2016, pp. 223-226, doi: 10.1109/ICEECCOT.2016.7955219.
7. Barkovska, O., Pyvovarova, D., Kholiev, V., Ivashchenko, H., & Rosinskyi, D. (2021). Information object storage model with accelerated text processing methods. In CEUR Workshop Proceedings (pp. 286-299).
8. Barkovska, O., Kholiev, V., Pyvovarova, D., Ivaschenko, G., & Rosinskiy, D. (2021). INTERNATIONAL SYSTEM OF KNOWLEDGE EXCHANGE FOR YOUNG SCIENTISTS. Advanced Information Systems-Sučasnì ìnformacìjnì sistemi, 5(1), 69-74.
9. M. Lan, C. L. Tan, J. Su and Y. Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 721-735, April 2009, doi: 10.1109/TPAMI.2008.110.
10. Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," in IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466, Nov. 2006, doi: 10.1109/TKDE.2006.180.
11. Yefeng Zheng, Huiping Li and D. Doermann, "Machine printed text and handwriting identification in noisy document images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 3, pp. 337-353, March 2004, doi: 10.1109/TPAMI.2004.1262324.
12. R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.
13. Y. -M. Su, H. -W. Peng, K. -W. Huang and C. -S. Yang, "Image processing technology for text recognition," 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), 2019, pp. 1-5, doi: 10.1109/TAAI48200.2019.8959877.
14. Olesia Barkovska, Oleg Mikhal , Daria Pyvovarova , Oleksii Liashenko , Vladyslav Diachenko and Maxim Volk, Local Concurrency in Text Block Search Tasks, International Journal of Emerging Trends in Engineering Research. - Volume 8. No. 3, March 2020. – P.690-694.
15. Barkovska O., Pyvovarova D. and Serdechnyi V., Pryskorenyj alghorytm poshuku sliv-obraziv u teksti z adaptyvnoju dekompozycijeju vykhidnykh danykh. [Accelerated word-image search algorithm in text with adaptive decomposition of input data]. Systemy upravlinnja, navighaciji ta zv'jazku 4 (56), 28-34. (in Ukrainian)

Received (Надійшла) 30.11.2021
Accepted for publication (Прийнята до друку) 19.01.2022

The importance of massively parallel computing systems in scanned documents processing

O. Barkovska, V. Kholiev, D. Polikanov

Abstract. The paper proposes a generalized classification model for scanned documents, which represents an organizational-functional, technological and software-hardware complex for document classification or categorization by keywords defined in a frequency dictionary. The relevance of the research topic lies in time reducing for new scanned information resources streamline, due to the increase in the speed of the **methods** for improving the quality of the original image immediately before processing and analyzing the text on images. **The analysis of the results** showed the effectiveness and expediency of using massively parallel computers to perform tasks such as noise reduction and changing the value of color channels of the original full-color image, achieving an acceleration of up to 53,51% compared to using the computing resources of the central processor.

Keywords: system, processing, text, processor, GPU, multithreading, image, preprocessing, speedup, information resource, storage.