Peculiarities of the raster images of hieroglyphic texts digitalization process organization

Kukunin Serhiy *

Received: 2022-03-22
Accepted: 2023-04-29
DOI: http://dx.doi.org/10.5281/zenodo.8109229

Abstract. An analysis of modern techniques for the use of software and neural network algorithms in solving the task of automating the digitization and subsequent restoration of text data, which include the Chinese characters "Hanzi", was carried out. It is noted that when restoring printed hieroglyphic texts, it is possible to effectively use procedures based on morphological operators, the adaptation of which to the given task makes it possible to reduce the load on the computing resource of the hardware and software platform of the machine analysis complex, as well as to reduce the delay when restoring text data presented in the form raster image. It is indicated that the efficiency of hieroglyphic text recovery significantly increases when the text block is divided into constituent elements and groups of elements according to the multi-level classification structure, such as the "sentence-phrase-word" structure. Similarly, effective restoration of a single hieroglyph is possible when it is divided into radicals, and radicals further into separate components, such as strokes and combinations of strokes of a fixed slope and thickness according to the font. A method of restoration of the hieroglyphic text presented in the form of raster images is proposed, based on the application of morphological procedures of erosion and dilation, which are used to remove defects of the image matrix and restore the structure of individual hieroglyphs. The specified approach consists in choosing the matrix of the graphic primitive of the morphological algorithm procedure and rotating the image matrix according to the thickness and angle of inclination of individual lines as elementary components of hieroglyphs. At the same time, within the framework of the study, the effectiveness of the recovery of the hieroglyphic text can also be carried out in an automated mode according to the target indicators of the accuracy of the reconstruction process. Thus, the task of optimizing the data conversion algorithm can be carried out by determining the maximum of the objective function, the arguments of which are the size and shape of the graphic primitive, the rotation angle of the image matrix, as well as the sequence of the erosion and dilation procedures.

Keywords: hieroglyphic text, text data reconstruction, image restoration, morphological algorithms, erosion procedure, dilation procedure.

* Lead Developer at Easy Health, Inc., ORCID: 0000-0001-5243-495X

INTRODUCTION

У рамках глобального тренду, пов'язаного з цифровізацією даних представлених у аналоговому вигляді, зазначається висока ***актуальність*** задачі переведення у цифрову форму текстових даних, зокрема формування символьних послідовностей на основі растрових зображень відсканованих документів. Якість представлення текстових даних визначається значною кількістю факторів, як то рівнем пошкодження аналогового носія, якістю обладнання для сканування та рівнем підготовки персоналу, а тому процедура відновлення растрових зображень низької якості є типовою у галузі переведення документів у цифрову форму. При цьому зазначається, що для текстів складених на основі латинської або кириличної абетки (зокрема для нестандартних шрифтів, курсиву, рукописного тексту, тощо) для відновлення символьних складових матриці зображення та розпізнавання тексту можуть бути ефективно використані як програмні, так і нейромережеві алгоритми [1-5]. Водночас, переведення у цифрову форму текстових даних, що включають у себе китайські ієрогліфи «ханьцзи» (китайська писемність), як то текстів китайською мовою, текстів японською мовою, що доповнюються символами складових абеток хіраґана і катакана, текстів в'єтнамською мовою (до 1945 року в'єтнамська писемність базувалась на китайських ієрогліфах), а також текстів корейською мовою, де китайські ієрогліфи використовуються на доданок до фонематичної абетки хангиль, залишається нетривіальною задачею [6-10].

Можна зазначити, що складність задачі автоматичного відновлення та розпізнавання ієрогліфічних текстів полягає у складній структурі ієрогліфів у порівнянні з символьними знаками латини або кирилиці, що збільшує вимоги до якості друку та сканування відповідних текстів, а також значній кількості ієрогліфів, що надзвичайно ускладнює процедуру їх розпізнавання відповідно до необхідності у розширення класифікації символів у 100-2000 разів у порівнянні зі стандартною абеткою (у залежності від типу писемності). При цьому посеред ієрогліфічних текстів є значна кількість документів, що становлять високу культурну, наукову та комерційну значимість, а тому на сьогоднішній день значна кількість досліджень присвячена вирішенню відповідної задачі.

Аналіз результатів наукових досліджень представлених у профільних виданнях присвячених автоматизації процесу розпізнавання ієрогліфічних текстів вказав на проблеми, що виникають при відновленні матриці зображення, класифікації ієрогліфів та цифровізації аналогових даних на рівні формування відповідного символьного ряду [6-10]. Найбільшу ефективність у відповідній галузі показує застосування нейромережевих алгоритмів, зокрема рекурентних нейронних мереж (RNN: Recurrent Neural Networks), а також архітектур побудованих на її основі, як то двонаправлених моделей (BRNN: Bidirectional Recurrent Neural Networks), моделей довгої короткочасної пам'яті (LSTM: Long Short-Term Memory) та RNN-перетворювачів [11-15]. Дослідники зауважують, що застосування нейромережевих алгоритмів збільшує навантаження на обчислювальний ресурс апаратно-програмної платформи, що відповідно збільшує складність організації і загальний кошторис системи машинного аналізу, а також час обробки вхідного набору даних. Тому для вирішення задачі переведення у цифрову форму друкованого тексту представленого на аналогових носіях актуально розробити ефективні програмні алгоритми відновлення растрових зображень з ієрогліфічними текстовими даними та класифікації їх складових елементів, що на сьогоднішній день розглядається як

невирішена частина загального дослідження.

Таким чином, ***метою роботи*** стала побудова комплексної методики відновлення структури друкованих ієрогліфічних текстів, представлених у вигляді растрових зображень, а також виділення і класифікації їх складових елементів (китайських ієрогліфів «ханьцзи»), на основі системи машинного аналізу, що базується на програмних алгоритмах і морфологічних операторах. При цьому задача машинного аналізу рукописних текстів або текстів, у яких використовуються шрифти, що імітують рукописний текст, виходить за рамки дослідження у зв'язку з тим, що у даному випадку більшу ефективність показують нейромережеві алгоритми.

РЕЗУЛЬТАТИ

1. Класифікація складових ієрогліфічного тексту представленого у вигляді растрового зображення

Постановка задачі виділення аналізу ієрогліфічного текстового блоку, включає у себе етап його поділення на складові елементи відповідно до багаторівневої класифікації:

- речення та предикативні центри речень, що поділяються за допомогою знаків пунктуації та союзів (у ієрогліфічних текстах зазвичай не використовуються пропуски між словами, що ускладнює задачу класифікації);
- блоки ієрогліфів слів предикативних центрів, що відокремлюються від знаків пунктуації, арабських цифр та символів абетки;
- окремі ієрогліфи, на основі яких складаються слова.

На рис. 1. наведено приклад застосування відповідної схеми до блоку тексту японською мовою присвяченого історії України. Тест складається з двох речень, кожне з яких включає у себе два предикативних центри. Символами, що використовуються у тексті поза ієрогліфів є елементи японської складової абетки хіраґана, що використовується для формування суфіксів, прикметників та союзів, символи складової абетки катакана, що використовується для слів іноземного походження, зокрема топонімів (у даному тексті це топоніми «Україна», «Київська Русь» та аналогічні), а також арабських цифр, що у японський писемності заміняють відповідні ієрогліфи у календарних датах та при розрахунках. Надалі блоки ієрогліфів згідно зі схемою поділяються на окремі символи. Характерно, що значна кількість японських слів є багатокореневими і складаються з кількох ієрогліфів, що можуть бути поділені відповідно їх лінійним розмірам.

Незалежно від складності структури ієрогліфу відповідний символ має уніфікований розмір і пропорції, причому це характерно як для друкованих, так і для рукописних ієрогліфічних текстів.

Аналогічним чином можна запропонувати багаторівневу схему машинного аналізу та класифікації ієрогліфу, що включає у себе поділ окремого ієрогліфу на радикали, а радикалів базові елементи, що представляють собою риски фіксованого нахилу та розміру, що залежить від шрифту, включення риски у окремий радикал та положення радикала у ієрогліфі (рис. 2).

На рисунку представлено класифікацію складових елементів ієрогліфу «OKU» (ставити, встановлювати), радикали якого цілком розкладаються на базові елементи горизонтальних та вертикальних рисок, що відрізняються довжиною та положенням у загальній структурі. Більшість ієрогліфів, тим не менш, включає більшу кількість базових елементів, що суттєво ускладнює задачу дослідження.

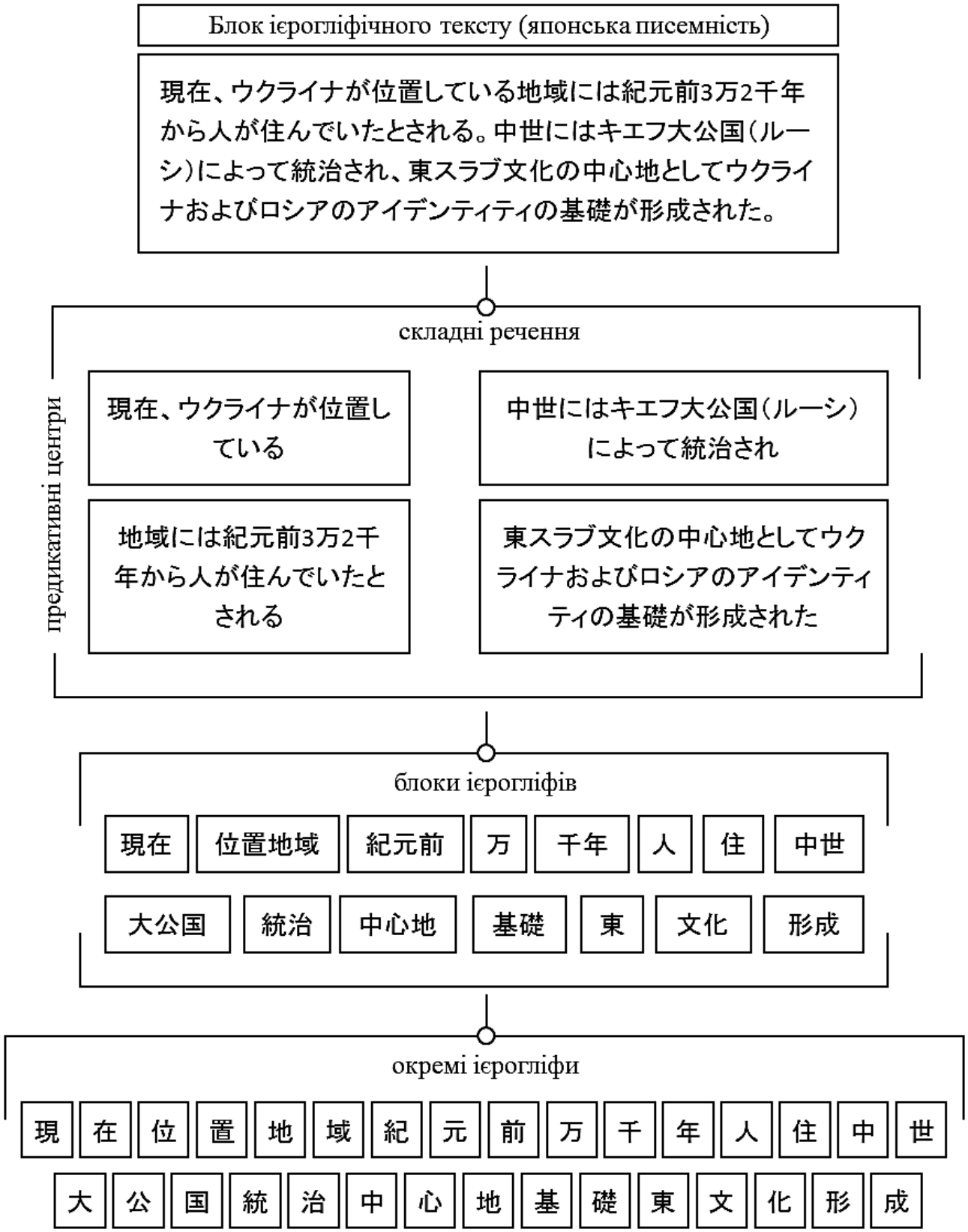

Рис. 1. Схема багаторівневої класифікації елементів блоку ієрогліфічного тексту

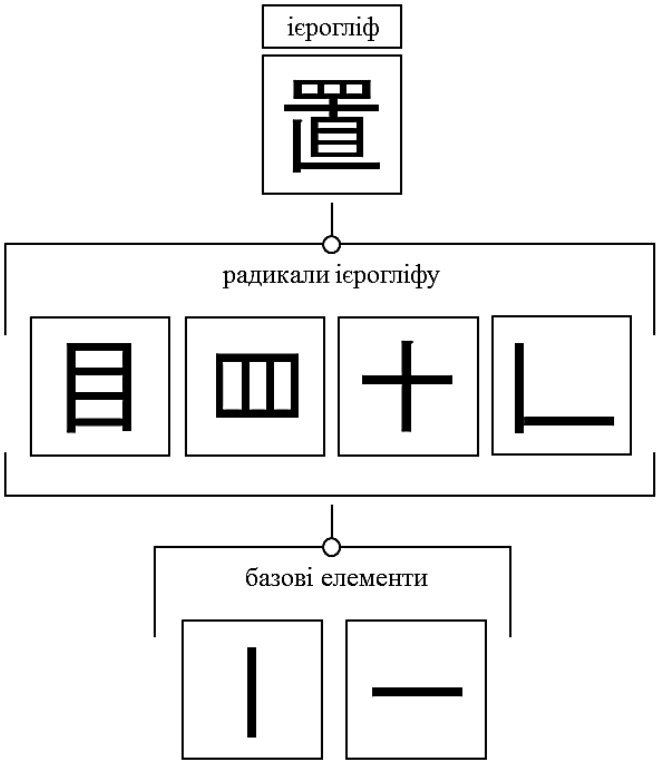

Рис. 2. Схема багаторівневої класифікації радикалів та базових елементів ієрогліфу

2. Морфологічні алгоритми відновлення ієрогліфічного тексту представленого у вигляді растрового зображення

При роботі по переведенню у цифрову форму текстових блоків представлених у вигляді растрового зображення вхідні дані попередньо переводяться у монохромних графічних даних через застосування порогових методів та налаштування параметрів зображення. Матриця монохромного зображення ієрогліфічного тексту, що характеризується роздільною здатністю $X_T \times Y_T$ при цьому має бути представлена як множина бінарних змінних $\mathbf{T}$: $\{t(x,y)\}$, де $\forall t(x,y) \in \{0;1\}$, $x \in [1;X_T]$, $y \in [1;Y_T]$, до якого надалі можуть бути ефективно застосовані морфологічні алгоритми [16-20] з метою виділення дефектів та відновлення структурних елементів, що складають ієрогліф. Зважаючи на те що, ієрогліф складається з окремих рисок, кут положення яких по відношенню до горизонталі складає фіксований набір значень $\{\theta_i\}$, де $i \in [1;I]$, графічний примітив може являти собою двовимірну матрицю бінарних елементів $\mathbf{P}$: $\{p(x,y)\}$, де $\forall p(x,y) \in \{0;1\}$, розмірність якої $X_T \times Y_T$ з одного боку відповідає максимально допустимому розриву у структурі зображення $x \in [1;X_P]$, що є меншим за мінімальну відстань між сусідніми базовими елементами ієрогліфу, а з іншого середній товщині риски ієрогліфа $y \in [1;Y_P]$ у відповідності до заданого шрифту.

На рис. 3 представлена схема застосування морфологічних алгоритмів при відновленні структурних елементів ієрогліфу «higashi» (схід). Відповідні структурні елементи представляють собою риски різною довжини розташовані під одним з чотирьох кутів фіксованого набору $\{\theta_1, \theta_2, \theta_3, \theta_4\}$. Через поворот матриці зображення на відповідні кути та застосування морфологічних алгоритмів може бути відновлена структура кожного з елементів ієрогліфу, що пропонується формалізувати у рамках дослідження.

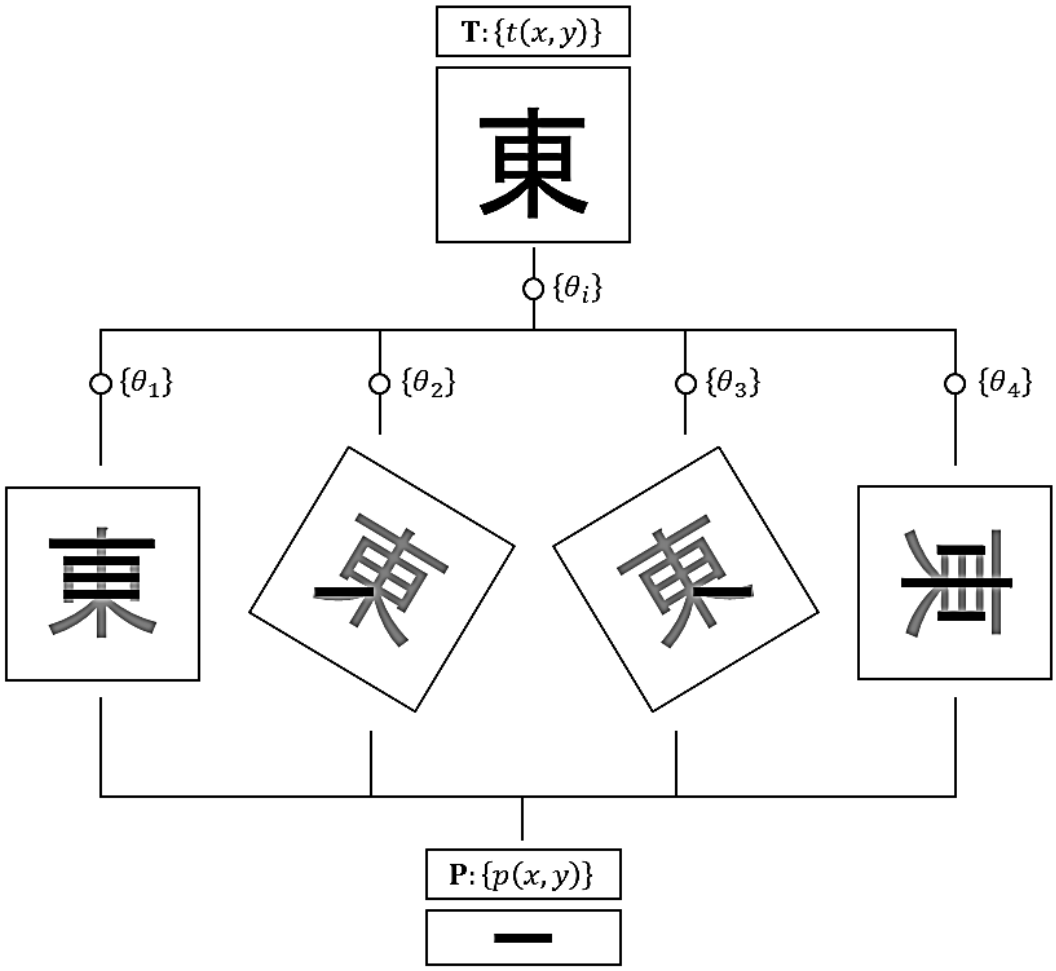

Рис. 3. Базова схема застосування морфологічних алгоритмів при відновленні структурних елементів ієрогліфу

Відповідно до запропонованого підходу алгоритм відновлення структури ієрогліфів текстового блоку представленого у вигляді монохромного растрового зображення $\mathbf{T}: \{t(x,y)\}$ через застосування графічного примітиву $\mathbf{P}: \{p(x,y)\}$ і морфологічних методів має включати у себе наступні етапи, що виконуються циклічно для повного набору кутів $i \in [1; I]$:

Вибір початкового значення кута з набору $\{\theta_1\}$.

Поворот матриці зображення, таким чином щоб відповідні риски ієрогліфи розташовувались горизонтально.

Розміщення графічного примітиву $\mathbf{P}: \{p(x,y)\}$ по відношенню до матриці зображення у положенні $x \in [1; X_P]$ і $y \in [1; Y_P]$.

Застосування оператору антиеквіваленції $t(x,y) \oplus p(x,y)$ для всіх x і y з метою нівелювання розривів у структурі зображення.

Перехід $x \to x+1$ до встановлення межі $x \in [X_T - X_P; X_T]$ і перехід до пункту «4».

Перехід $y \to y+1$ до встановлення межі $y \in [Y_T - Y_P; Y_T]$ і перехід до пункту «4».

Перехід до наступного кута $i \to i+1$ до проведення зазначеної процедури для останнього кута θ_I і перехід до пункту «2».

Наведена процедура може бути визначена як морфологічний алгоритм дилатації. Після цього до матриці зображення застосовується процедура дилатації, що включає у себе виконання наступних етапів:

Вибір початкового значення кута з набору $\{\theta_1\}$.

Поворот матриці зображення, таким чином щоб відповідні риски ієрогліфи розташовувались горизонтально.

Розміщення графічного примітиву $\mathbf{P}$: $\{p(x,y)\}$ по відношенню до матриці зображення у положенні $x \in [1; X_P]$ і $y \in [1; Y_P]$.

Застосування оператору кон'юнкції $t(x,y) \wedge p(x,y)$ для всіх x і y з метою нівелювання розривів у структурі зображення.

Перехід $x \to x+1$ до встановлення межі $x \in [X_T - X_P; X_T]$ і перехід до пункту «4».

Перехід $y \to y+1$ до встановлення межі $y \in [Y_T - Y_P; Y_T]$ і перехід до пункту «4».

Перехід до наступного кута $i \to i+1$ до проведення зазначеної процедури для останнього кута θ_I і перехід до пункту «2».

Відповідна процедура надає можливість зменшити товщину рисок ієрогліфу, що збільшується внаслідок дилатації, а також усунути дефекти зображення.

ВИСНОВКИ

У результаті проведеного дослідження було проаналізовано принципи застосування морфологічних алгоритмів при відновленні структури друкованих ієрогліфічних текстів, представлених у вигляді растрових зображень з метою збільшення ефективності процесу виділення і класифікації їх складових елементів.

При цьому у рамках дослідження було розроблено:

схему багаторівневої класифікації елементів блоку ієрогліфічного тексту;

схему багаторівневої класифікації радикалів та базових елементів ієрогліфу;

схему адаптації морфологічних алгоритмів при відновленні структурних елементів ієрогліфу.

References

1. Liu, Y., Cao, F., & Zhang, Y. (2022). Generative adversarial examples for sequential text recognition models with artistic text style. *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods.* https://doi.org/10.5220/ 0010866800003122.
2. Ahmed, S. B., Razzak, M. I., & Yusof, R. (2019). Progress in cursive wild text recognition. *Cursive Script Text Recognition in Natural Scene Images*, 85-92. https://doi.org/10.1007/ 978-981-15-1297-1_5.
3. Martinovska, C., Nedelkovski, I., Klekovska, M., & Kaevski, D. (2012). Recognition of old Cyrillic Slavic letters: Decision tree versus fuzzy classifier experiments. *2012 6th IEEE International Conference Intelligent Systems.* https://doi.org/10.1109/is.2012.6335113.
4. Kachalin, V. S., Panov, Y. N., & Popov, N.-L. E. (2022). Comparative analysis of various Optical Character Recognition Systems when working with text written using the Cyrillic alphabet. *Modern High Technologies, 8*(22), 65-70. https://doi.org/10.17513/snt.39268.
5. Martinovska, C., Klekovska, M., Nedelkovski, I., & Kaevski, D. (2013). Methodologies for recognition of old Slavic Cyrillic characters. *International Journal of Computational Intelligence Studies, 2*(3/4), 264. https://doi.org/10.1504/ijcistudies.2013.057639.
6. Lu, H., & Peng, Y. (2022). Named entity recognition of Chinese legal text based on BERT. *2022 4th International Conference on Applied Machine Learning (ICAML).* https://doi.org/10.1109/icaml57167.2022.00029.

7. Nguyen, K. C., & Masaki, N. (2016). Enhanced character segmentation for format-free Japanese text recognition. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR).* https://doi.org/10.1109/icfhr.2016.0037.
8. Gao, J., Zhu, B., & Nakagawa, M. (2014). Building compact recognizer with recognition rate maintained for on-line handwritten Japanese text recognition. *Pattern Recognition Letters, 35*, 169–177. https://doi.org/10.1016/j.patrec.2013.08.014.
9. Jeong, J.-ha. (2019). A study on the recognition of Heo Mok's archaic text. *The Korean Society of Calligraphy, (34)*, 69–85. https://doi.org/10.19077/tsoc.2019.34.3.
10. Xuan, T. L., Thi, H. P., & Do, H. N. (2022). Vietnamese text detection, recognition and classification in images. *2022 14th International Conference on Knowledge and Systems Engineering (KSE).* https://doi.org/10.1109/kse56063.2022.9953789.
11. Li, J., Zhang, D., & Wulamu, A. (2021). Chinese text classification based on ERNIE-RNN. *2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT).* https://doi.org/10.1109/cecit53797.2021.00072.
12. Zhang, X., & Yan, K. (2019). An algorithm of bidirectional RNN for offline handwritten Chinese text recognition. *Intelligent Computing Methodologies,* 423–431. https://doi.org/ 10.1007/978-3-030-26766-7_39.
13. Gao, L., Li, H., Liu, Z., Liu, Z., Wan, L., & Feng, W. (2021). RNN-transducer based Chinese Sign Language recognition. *Neurocomputing, 434*, 45–54. https://doi.org/10.1016/j.neucom.2020.12.006.
14. Messina, R., & Louradour, J. (2015). Segmentation-free handwritten Chinese text recognition with LSTM-RNN. *2015 13th International Conference on Document Analysis and Recognition (ICDAR).* https://doi.org/10.1109/icdar.2015.7333746.
15. Wang, Z.-R., & Du, J. (2021). Joint Architecture and knowledge distillation in CNN for Chinese text recognition. *Pattern Recognition, 111*, 107722. https://doi.org/10.1016/j.patcog.2020.107722.
16. Schmitt, M. (2018). On Two Inverse Problems in Mathematical Morphology. *Mathematical Morphology in Image Processing,* 151-169. doi:10.1201/9781482277234-5.
17. Roerdink, J. B. (2018). Mathematical Morphology with Noncommutative Symmetry Groups. *Mathematical Morphology in Image Processing,* 205-254. doi:10.1201/9781482277234-7.
18. Shih, F.Y. Grayscale morphology. (2017). Image Processing and Mathematical Morphology, 49–60. https://doi.org/10.1201/9781420089448-7.
19. Bao, L. I.-F., & Li, H.-M. (2020). Study on the morphology of mature ovarian follicles using Image Processing Toolbox. *Research Square.* https://doi.org/10.21203/rs.3.rs-42445/v1.
20. Vincent, L., & Heijmans, H. (2018). Graph Morphology in Image Analysis. *Mathematical Morphology in Image Processing,* 170-203. doi:10.1201/9781482277234-6.