УДК 004.932:519.237

DOI 10.31651/2076-5886-2025-1-46-57

PACS 07.05.Mh

ЗАБОЛОТНІЙ Юрій Леонідович
викладач спеціальних дисциплін Тальнівського будівельно-економічного фахового коледжу Уманського національного університету
e-mail: Henry23@i.ua
ORCID 0009-0003-1377-7804

ЛЕВЧЕНКО Алла Василівна
учитель математики Тальнівської загальноосвітньої школи І-ІІІ ступенів №2 Тальнівської міської ради Черкаської області
e-mail: allalevchenko70@gmail.com
ORCID 0009-0009-4211-5950

ТИСЯЧНА Катерина Русланівна
учениця Тальнівської загальноосвітньої школи І-ІІІ ступенів №2 Тальнівської міської ради Черкаської області
e-mail: katerinatisacna@gmail.com

СТАТИСТИЧНИЙ АНАЛІЗ КОЛІРНИХ ПРОФІЛІВ ЦИФРОВИХ ЗОБРАЖЕНЬ ТА ЇХ КЛАСТЕРИЗАЦІЯ МЕТОДОМ K-СЕРЕДНІХ

У статті розглянуто методи статистичного аналізу цифрових зображень з метою виявлення характерних колірних ознак та їх автоматизованого групування. Предметом дослідження є колірні профілі цифрових зображень у просторах RGB та HSV, а також їх кластеризація за допомогою алгоритму k-середніх. Цифрове зображення формалізовано як статистичну вибірку у вигляді множини піксельних векторів, над якою виконуються операції обчислення основних числових характеристик: середнього значення, дисперсії, стандартного відхилення та моди для кожного з колірних каналів. Для побудови колірних профілів зображень використано гістограми яскравості та кольору, а для порівняння розподілів застосовано евклідову метрику подібності.

Алгоритм k-середніх реалізовано як метод некерованої кластеризації, що ітеративно мінімізує суму квадратів відстаней між колірними профілями зображень і центроїдами кластерів. Оптимальна кількість кластерів визначається за допомогою методу Elbow (аналіз SSE) та коефіцієнта Silhouette Score. Для оцінки статистичної значущості відмінностей між кластерами застосовано дисперсійний аналіз (ANOVA).

Програмна реалізація виконана мовою Python з використанням бібліотек OpenCV, NumPy, scikit-learn, Matplotlib, Pandas та Tkinter. Проведено експериментальне дослідження на наборі з десяти тестових зображень різних категорій. Порівняльний аналіз результатів у просторах RGB та HSV показав, що колірний простір HSV забезпечує вищу якість кластеризації та більш змістовну інтерпретацію статистичних характеристик, оскільки краще відповідає особливостям людського зорового сприйняття. Практичне значення результатів полягає у можливості застосування розробленої методики для автоматизованого сортування та класифікації фотографій, систем комп'ютерного зору, медичної діагностики та аналізу супутникових знімків.

Ключові слова: *статистичний аналіз зображень, колірний профіль, простір RGB, простір HSV, кластеризація, метод k-середніх.*

Вступ

У сучасному світі цифрові зображення є невід'ємною частиною багатьох сфер діяльності – від фотографії та соціальних мереж до медицини, промисловості та штучного інтелекту. Зростання обсягів візуальної інформації зумовлює нагальну потребу у розробці ефективних методів автоматизованої обробки та класифікації зображень [1, 2].

Одним із ефективних підходів до автоматизованого групування зображень є статистичний аналіз їх кольорових профілів, який дозволяє формалізувати інформацію про відтінки, яскравість та насиченість піксельних даних і зробити їх придатними для математичної обробки [3, 13]. Методи некерованої кластеризації, зокрема алгоритм k-середніх (k-means), належать до найбільш поширених і відносно простих підходів до автоматичного групування даних без попереднього маркування [4, 7, 9].

Вибір колірного простору для представлення пікселів зображень суттєво впливає на якість статистичного аналізу та кластеризації. Незважаючи на широке використання RGB, цей простір недостатньо ефективно розділяє яскравісну та кольорову складові зображення. Колірний простір HSV, орієнтований на особливості людського зорового сприйняття, розділяє ці характеристики, що може підвищити інформативність аналізу [13-15].

Актуальність теми визначається практичною потребою у вдосконаленні підходів до класифікації зображень в умовах зростання обсягів візуальної інформації та розвитку технологій машинного навчання.

Метою роботи є розробка та апробація методики статистичного аналізу кольорових профілів цифрових зображень із застосуванням алгоритму k-середніх для кластеризації, а також порівняльний аналіз впливу вибору колірної моделі (RGB та HSV) на інформативність статистичних характеристик і якість кластеризації зображень.

Виклад основного матеріалу:

1. Огляд методів розв'язання задачі

Задача автоматичної класифікації зображень за колірними ознаками активно досліджується в галузях комп'ютерного зору та машинного навчання. Узагальнений огляд методів кластерного аналізу подано у роботах [1, 4, 9, 10]. Зокрема, у [8] представлено комплексний аналіз алгоритму k-means та підкреслено його ефективність для задач обробки зображень. У роботах [13-15] розглянуто особливості різних колірних просторів та їх придатність для аналізу кольору. Підходи до порівняльного аналізу колірних просторів RGB та HSV у задачах сегментації та розпізнавання об'єктів описано у [13, 14]. Метрики оцінки якості кластеризації (SSE, Silhouette Score, ANOVA) систематизовано у роботах [5, 6, 11, 12].

Проте питання порівняльної ефективності кластеризації колірних профілів зображень у різних просторах та вплив вибору простору на результати автоматичного групування досліджено в наявній літературі недостатньо, що зумовлює актуальність даної роботи.

2. Постановка задачі

Нехай задано набір з N цифрових зображень. Кожне зображення подається як функція:

$$f : \{1, 2, \ldots, M\} \times \{1, 2, \ldots, N\} \to \mathbb{R}^d$$

де M, N – розміри у пікселях (рядків і стовпців), d – кількість каналів (для кольорового зображення $d = 3$). Кожному пікселю (i, j) відповідає вектор з трьох значень інтенсивностей у відповідному колірному просторі.

Для кожного цифрового об’єкта формується кольоровий профіль у вигляді вектора статистичних ознак каналів. Задача кластеризації полягає у розбитті множини профілів на k груп таким чином, щоб об'єкти всередині кожної групи були максимально схожими між собою, а між групами – максимально відмінними [4, 5].

3. Методи розв'язання

3.1. Статистичні характеристики кольорових каналів

Для кожного зображення обчислюються чотири основні характеристики кожного кольорового каналу [3, 17].

Середнє значення інтенсивності каналу визначає загальну «світлість» цього каналу і обчислюється як:

$$\overline{I} = \frac{1}{N}\sum_{i=1}^{N} I_i ,$$

де I_i – значення інтенсивності пікселя; N – загальна кількість пікселів.

Дисперсія показує, наскільки сильно розкидані значення яскравості навколо середнього:

$$D = \frac{1}{N}\sum_{i=1}^{N} (I_i - \overline{I})^2 .$$

Зростання дисперсії відповідає підвищенню контрастності та неоднорідності кольорового розподілу.

Стандартне відхилення – це квадратний корінь із дисперсії:

$$\sigma = \sqrt{D} .$$

Стандартне відхилення інтерпретується як середня «амплітуда» відхилень піксельних значень від середнього. Воно є показником контрастності або насиченості кольору у межах каналу.

Мода використовується як оцінка домінантної кольорової компоненти зображення. Іншими словами, мода характеризує найтиповіше значення досліджуваної величини. У задачах аналізу кольорових зображень мода може інтерпретуватися як найпоширеніше значення яскравості або кольорової компоненти, що дозволяє виявити домінантний тон або відтінок зображення.

3.2. Колірні простори

У дослідженні використано два колірні простори [13, 14, 15, 16]:

Колірний простір RGB (Red, Green, Blue) базується на адитивній моделі кольорів. Кожен колір описується трьома значеннями у діапазоні [0, 255], що визначають інтенсивність червоного, зеленого та синього каналів.

Колірний простір HSV (Hue, Saturation, Value) є нелінійним перетворенням RGB, орієнтованим на особливості людського зорового сприйняття. Компонента Hue (H) задає відтінок кутом від 0° до 360°; Saturation (S) характеризує насиченість кольору (0 – ахроматичний, 1 – максимально насичений); Value (V) визначає яскравість (0 – чорний, 1 – максимальна яскравість).

3.3. Алгоритм k-середніх

Алгоритм k-середніх є методом некерованої кластеризації, що ітеративно мінімізує суму квадратів відстаней від кожного об'єкта до центроїда свого кластера SSE (Sum of Squared Errors) [4, 7, 8, 9]:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - \mu_i \right\|^2 ,$$

де C_i – множина об'єктів i-го кластера, μ_i – центр i-го кластера.

Алгоритм складається з таких кроків: (1) вибір k випадкових початкових центроїдів; (2) призначення кожного об'єкта до найближчого центроїда за евклідовою відстанню; (3) перерахунок центроїдів як середніх значень об'єктів кластера; (4) повторення кроків 2-3 до збіжності (стабілізації центроїдів).

3.4. Програмна реалізація

Програмний комплекс розроблено мовою Python з використанням таких бібліотек: OpenCV (cv2) – зчитування та конвертація колірних просторів; NumPy – статистичні обчислення; scikit-learn – реалізація алгоритму KMeans та обчислення метрик; Matplotlib – побудова гістограм та графіків; Pandas – відображення результатів у табличній формі; Tkinter – графічний інтерфейс користувача.

Розроблений програмний модуль реалізує повний конвеєр аналізу зображень: завантаження та уніфікація розмірів зображень; перетворення у вибраний колірний простір; розрахунок статистичних характеристик; формування матриці кольорових профілів; кластеризація методом k-means; оцінка якості за показниками SSE, Silhouette Score та ANOVA; візуалізація результатів.

Розроблене програмне забезпечення є універсальним і дозволяє автоматично аналізувати зображення, отримувати числові характеристики кольорів, візуалізувати гістограми та здійснювати кластеризацію об’єктів за кольоровою подібністю. Інтерфейс програми наведено на рис. 1.

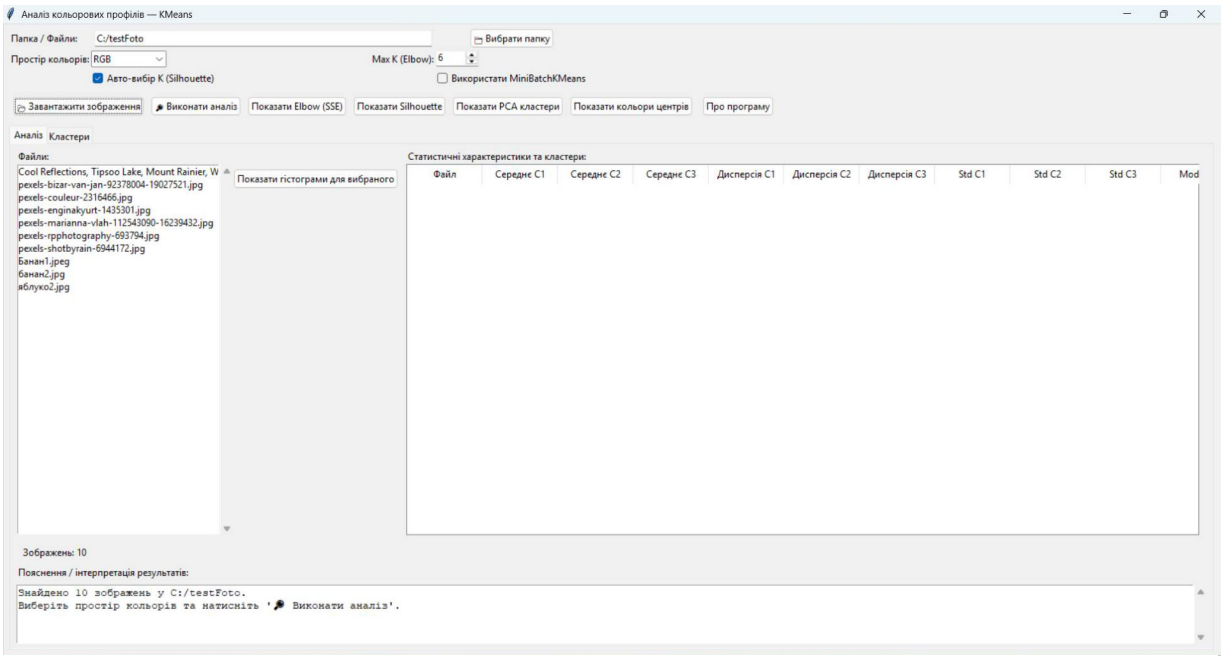


Рис. 1. Інтерфейс розробленого програмного комплексу

3.4.1 Попередня обробка зображень

Усі зображення проходять етап попередньої обробки. Цей етап необхідний для того, щоб привести дані до єдиного формату та забезпечити коректність подальших обчислень.

Для забезпечення однакових умов обробки всі зображення масштабуються до однакового розміру (наприклад, 200×200 пікселів). Залежно від положення перемикача користувача, зображення конвертуються у простір RGB або HSV.

Отримані значення кожного каналу зберігаються у вигляді числових масивів, які далі використовуються для обчислення статистичних характеристик і побудови кольорових профілів у вигляді числових векторів з трьома координатами.

3.4.2 Статистичний аналіз зображень

Після завантаження вибірки проводиться статистичний аналіз кольорових характеристик кожного зображення. Для цього реалізовано інтерфейс із кількома функціональними елементами управління:

- «Завантажити зображення» – відкриває папку, підраховує кількість знайдених зображень і показує їх у списку.
- «Показати гістограму» – будує гістограми розподілу кольорів для трьох каналів вибраної моделі (R, G, B або H, S, V).
- «Розрахувати статистику» – виконує розрахунки основних характеристик: середнє значення, дисперсія, стандартне відхилення і мода.
- «Перемикач RGB / HSV» – дозволяє аналізувати дані в різних кольорових просторах, що дає змогу порівняти отримані результати.

3.4.3 Кластеризація зображень

Для виявлення закономірностей і групування зображень за кольоровими профілями використано метод k-середніх, який реалізує наступні кроки:

1. Формується матриця колірних профілів усіх зображень $[x_R,x_G,x_B]$ або $[x_H,x_S,x_V]$.
2. Обирається кількість кластерів k (попередньо задається число N). Для k=1…N:
   - у Elbow Method обчислюється SSE і будується графік для визначення «ліктя», що вказує на оптимальне k;
   - у Silhouette Score обчислюється відповідний коефіцієнт для вибору оптимального k, що відповідає найбільшому значенню цього параметра.

   За цими двома критеріями обираємо оптимальне значення k.
3. Алгоритм ітеративно розподіляє зображення між кластерами, мінімізуючи відстань до центрів.
4. Після завершення кластеризації відображаються:

- кольори центрів кластерів (середні профілі груп);
- зображення групуються за кластерами;
- метрики якості кластеризації:

SSE – показує компактність кластерів: менше значення відповідає кращій групі;

Silhouette Score – оцінює наскільки об’єкт близький до свого кластера та наскільки далекий від інших. Значення близьке до 1 вказує на хорошу кластеризацію, а близьке до 0 відповідає перекриттю кластерів;

- *ANOVA (дисперсійний аналіз)* – це статистичний метод, призначений для перевірки гіпотези про рівність середніх значень у двох або більше незалежних групах.

У задачах аналізу даних ANOVA дозволяє визначити, чи є статистично значущі відмінності між групами, або ж спостережувані відмінності зумовлені випадковими коливаннями. Метод базується на порівнянні міжгрупової дисперсії (варіація між середніми значеннями груп) та внутрішньогрупової дисперсії (варіація всередині кожної групи). Якщо міжгрупова дисперсія істотно перевищує внутрішньогрупову, роблять висновок про статистично значущі відмінності між групами.

Результати кластеризації автоматично супроводжуються текстовою інтерпретацією отриманих метрик. Таким чином, кластеризація дозволяє автоматично визначати подібність об’єктів за їх кольоровими характеристиками.

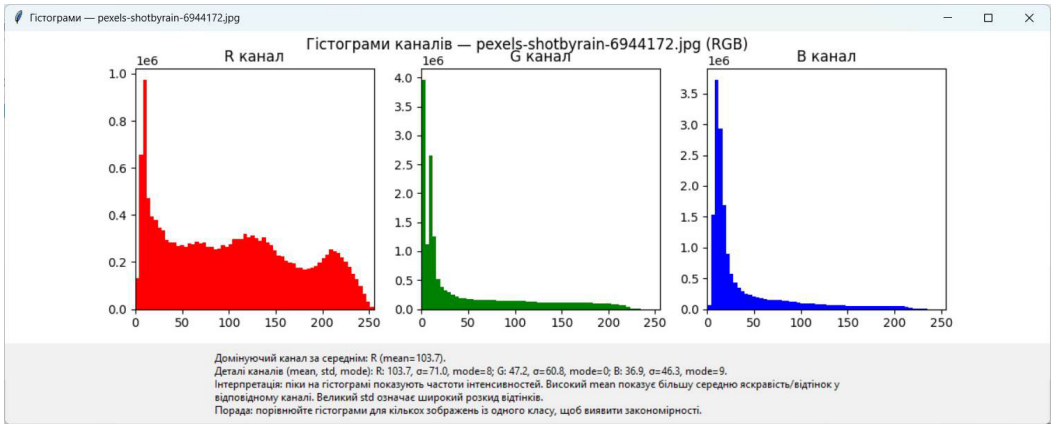
4. Отримані результати

Для експериментального дослідження використано десять зображень різних категорій (гірський пейзаж, природа, фрукти тощо). Усі зображення завантажено з відкритих джерел (Pexels, Pixabay та ін.).

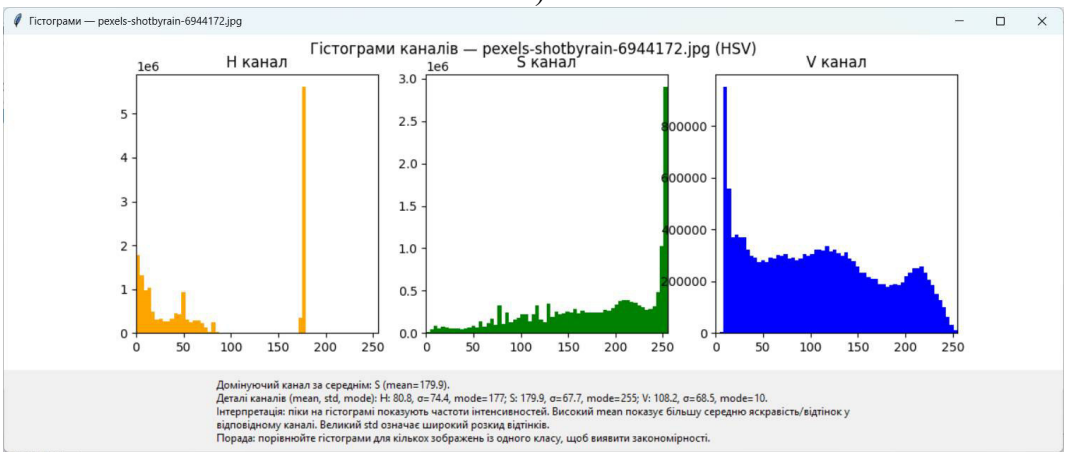
Статистичний аналіз у просторі RGB показав, що середні значення та дисперсії компонент R, G і B тісно пов'язані між собою, оскільки зміна яскравості зображення

одночасно впливає на всі три канали. Це ускладнює розрізнення зображень зі схожим рівнем освітленості, але з різними кольоровими характеристиками.

Гістограми компонент у просторі HSV продемонстрували більш виразні відмінності між зображеннями різних категорій. Компонента H (відтінок) чітко вирізняє «теплі» (схід сонця, фрукти) і «холодні» (природа, вода, небо) зображення, а компонента S (насиченість) дозволяє відокремити яскраво забарвлені зображення від монохромних. Наприклад, для зображення №7 (де представлено ягоди полуниці) на рис. 2 наведено отримані гістограми у колірних просторах RGB і HSV.



а)



б)

Рис. 2. Гістограми кольорових каналів зображення №7 у просторах RGB (а) та HSV (б)

4.1. Кольоровий простір RGB

Для визначення оптимальної кількості кластерів у колірному просторі RGB застосовано Elbow Method (при N = 6). Отриманий графік SSE представлено на рис.3а. Також виконано обчислення Silhouette Score, результат наведено на рис. 3б. Найбільше значення даного показника отримано при k = 2.

Розподіл профілів зображень за кластерами у просторі RGB за допомогою розробленого програмного простору подано на рис. 4. Також користувач отримує окремі пояснення та інтерпретацію обчислених значень. А саме:

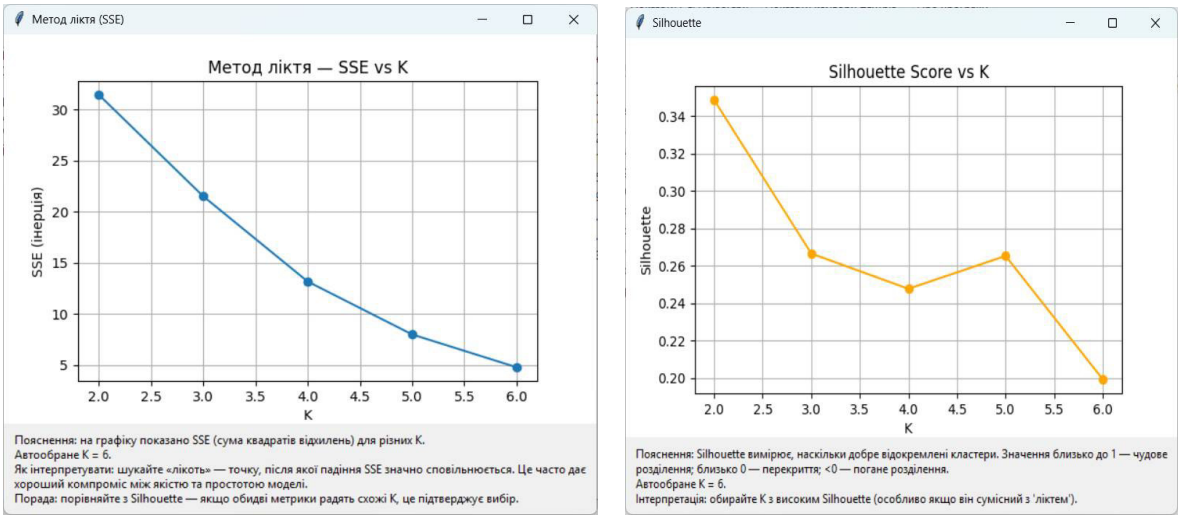
```
Результати аналізу:
   –  Завантажено зображень: 10
   –  Простір кольорів: RGB
   –  Знайдено кластерів: 2
```

Ключові метрики:

- SSE (сума квадратів відхилень): 31.45
- Silhouette: 0.349 свідчить про задовільне розділення кластерів.

Результати ANOVA (mean-канали):

- R_mean: F=14.08, p=0.0056 — середні значення статистично відрізняються між кластерами (p < 0.05).
- G_mean: F=18.33, p=0.0027 — середні значення статистично відрізняються між кластерами (p < 0.05).
- B_mean: F=1.29, p=0.2897 — немає статистично значущої різниці (p ≥ 0.05).



а) б)

Рис. 3. Результат застосування Elbow Method (а) і Silhouette Score (б) у колірному просторі RGB

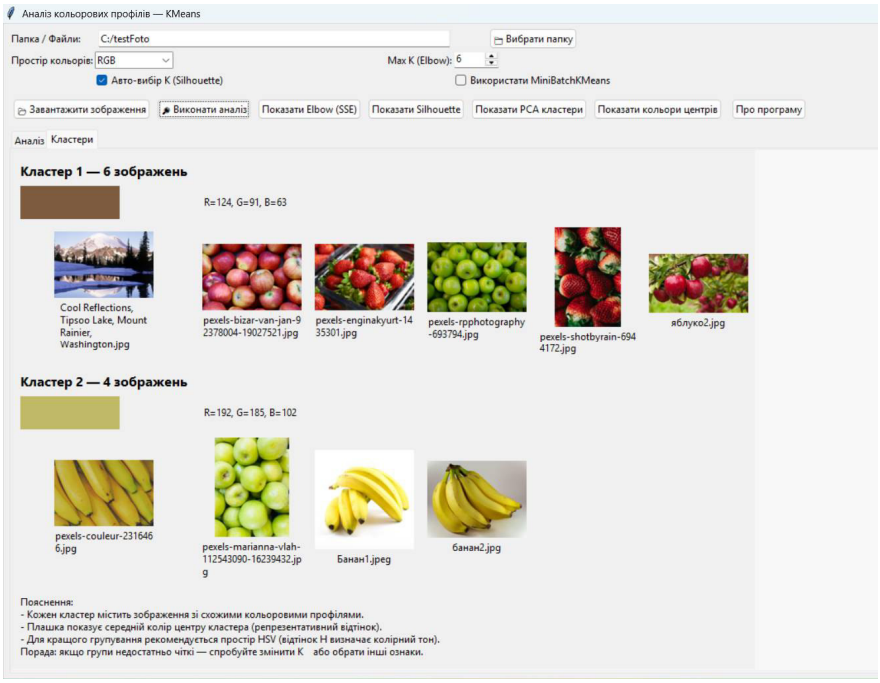


Рис. 4. Розподіл профілів зображень за кластерами у колірному просторі RGB

На рис. 5 представлено візуалізацію результатів кластеризації у просторі RGB з використанням PCA-проєкцій, де зображення центроїдів позначено хрестиком, а середні значення профілів зображень – кружечками з номером зображення.

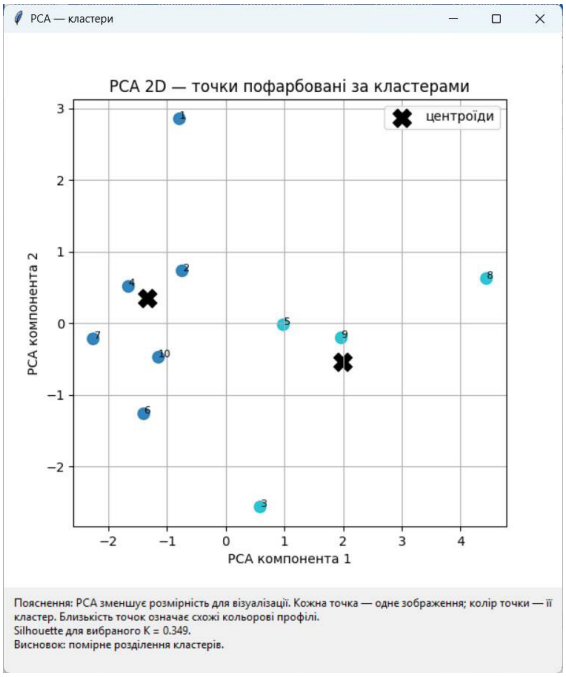
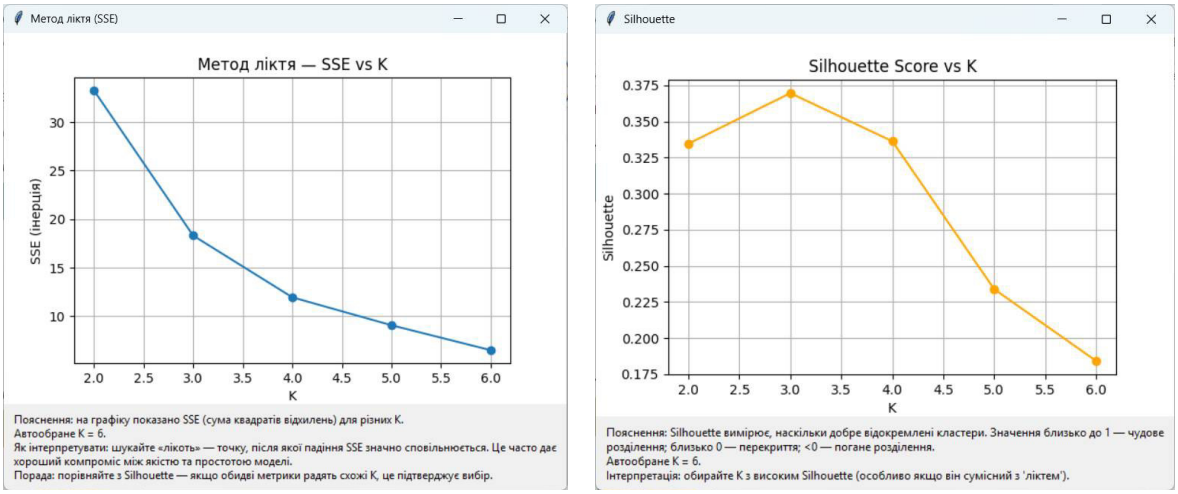


Рис. 5. Візуалізація результату кластеризації у кольоровому просторі RGB при k=2

Також було здійснено кластеризацію профілів зображень у просторі RGB при k = 3.

4.2. Кольоровий простір HSV

Результати застосування Elbow Method (при N = 6) та Silhouette Score у колірному просторі HSV для визначення оптимальної кількості кластерів наведено на рис. 6.



а) б)

Рис. 6. Результат застосування Elbow Method (а) і Silhouette Score (б) у колірному просторі HSV

Визначено оптимальне число кластерів k = 3 та отримано наступні результати за допомогою програмного комплекса:

```
Результати аналізу:
   –  Завантажено зображень: 10
   –  Простір кольорів: HSV
   –  Знайдено кластерів: 3

Ключові метрики:
   –  SSE (сума квадратів відхилень): 18.32
   –  Silhouette: 0.370 свідчить про задовільне розділення кластерів.

Результати ANOVA (mean-канали):
   –  H_mean: F=18.99, p=0.0015 → середні значення статистично
      відрізняються між кластерами (p < 0.05).
   –  S_mean: F=4.10, p=0.0663 → немає статистично значущої різниці (p ≥
      0.05).
   –  V_mean: F=4.32, p=0.0600 → немає статистично значущої різниці (p ≥
      0.05).
```

Візуалізацію результатів кластеризації у просторі HSV з використанням PCA-проєкцій подано на рис. 7.

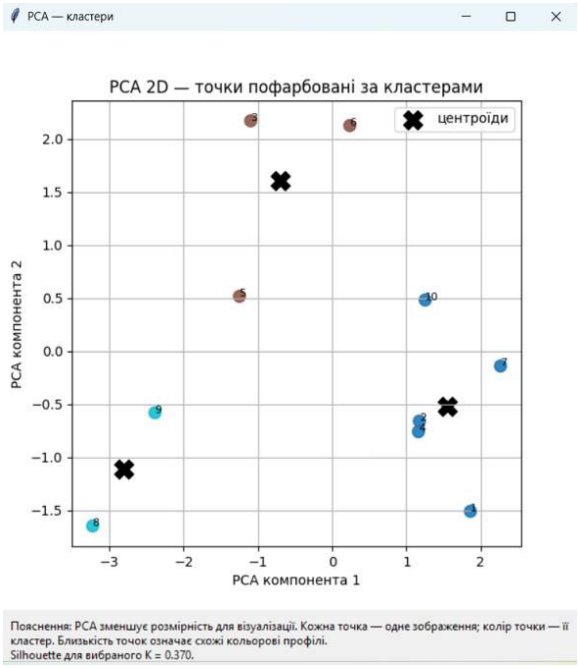


Рис. 7. Візуалізація результату кластеризації у кольоровому просторі HSV при k = 3

Також було здійснепно кластеризацію профілів зображень у просторі HSV при k = 4.

5. Аналіз результатів дослідження

Порівняльний аналіз кластеризації у просторах RGB та HSV показав суттєві відмінності як у кількості оптимальних кластерів, так і у якості групування. Простір RGB є зручним для базового числового аналізу та програмної реалізації, проте має обмежені можливості для якісного розрізнення кольорових профілів: компоненти R, G, B кодують яскравість і колір одночасно, що ускладнює інтерпретацію статистичних характеристик.

Простір HSV забезпечує більш наочну інтерпретацію та вищу якість кластеризації завдяки явному розподілу інформації між відтінком, насиченістю та яскравістю.

Значення Silhouette Score у HSV виявилися вищими, що свідчить про більш чітке розмежування між кластерами. Результати ANOVA підтвердили статистичну значущість відмінностей між кластерами в обох просторах, однак з нижчим p-значенням у HSV, що вказує на вищу дискримінантну здатність цього простору.

Наукова новизна роботи полягає у систематизації та адаптації методики застосування алгоритму k-середніх для кластеризації кольорових профілів цифрових зображень. Було проведено порівняльний аналіз ефективності кластеризації у різних колірних просторах, що дозволяє визначити найбільш придатний простір для задачі групування зображень. Запропоновано алгоритмічний підхід, що поєднує статистичний аналіз кольорових характеристик із методами оцінки якості кластеризації (SSE, Silhouette Score, ANOVA) для підвищення обґрунтованості вибору параметрів алгоритму.

Отримані результати узгоджуються з відомими теоретичними та прикладними дослідженнями у галузі комп'ютерного зору, що підтверджує коректність обраних методів та здійсненої програмної реалізації.

Висновки

У статті розроблено та апробовано методику статистичного аналізу кольорових профілів цифрових зображень із застосуванням алгоритму k-середніх. На основі проведеного дослідження можна зробити такі висновки:

1. Цифрове зображення є повноцінним об'єктом статистичного аналізу: множина піксельних векторів утворює статистичну вибірку, для якої можна обчислювати числові характеристики кольорових каналів і будувати кольорові профілі.

2. Алгоритм k-середніх у поєднанні з методами оцінки якості (Elbow Method, Silhouette Score, ANOVA) забезпечує обґрунтований вибір оптимальної кількості кластерів та надійне групування зображень за кольоровими характеристиками.

3. Колірний простір HSV є більш придатним для кластеризації кольорових профілів, ніж RGB: у HSV отримано вищі значення Silhouette Score та більшу кількість семантично значущих кластерів (k = 3 проти k = 2), що підтверджує кращу відповідність HSV особливостям людського зорового сприйняття.

4. Розроблений програмний комплекс мовою Python реалізує повний цикл аналізу: від завантаження зображень до автоматичної кластеризації та інтерпретації результатів. Практичне значення методики – застосування у системах комп'ютерного зору, автоматизованому сортуванні фотографій, медичній діагностиці та аналізі супутникових знімків.

Список використаної літератури:

1. Коломієць А. С., Марченко О. Г. Основи кластерного аналізу та його застосування. Київ : Національний університет, 2015. 210 с.
2. Коваленко І. В. Аналіз методів кластеризації у системах обробки даних. Журнал сучасних інформаційних технологій. 2017. № 3 (5). С. 75–81.
3. Скляр В. М. Методи кластерного аналізу у машинному навчанні. Математичні дослідження. 2020. № 1 (4). С. 112–119.
4. Цимбал А. Вступ до кластерного аналізу: основні підходи та методи. Харків : Наукова думка, 2002. 186 с.
5. Власенко О. А. Метрики для кластерного аналізу даних: теоретичні основи та практичні аспекти. Київ : Видавництво КНЕУ, 2016. 148 с.
6. Демченко О. Б. Кластеризація як метод обробки даних у сучасних дослідженнях. Інформаційні системи та технології. 2018. № 2 (8). С. 50–57.
7. Швець К. В. Дослідження моделей еволюції кластерів в задачах розпізнавання образів [Електронний ресурс]. URL: https://openarchive.nure.ua/server/api/core/bitstreams/4b7cf3f6-1df4-4f0f-b180-4beb0ca94454/content (дата звернення: 30.10.2025).

8. Jain A. K. Data clustering: 50 years beyond k-means. Pattern Recognition Letters. 2010. Vol. 31, No. 8. P. 651–666.
9. Марченко О. О., Россада Т. В. Актуальні проблеми Data Mining : навч. посіб. для студентів факультету комп'ютерних наук та кібернетики. Київ, 2017. 150 с.
10. Дудко О. В., Поляков С. В. Основи кластерного аналізу: методи та алгоритми. Наукові записки НаУКМА. Комп'ютерні науки. 2018. № 1. С. 23–32.
11. Дем'янчук І. С. Кластеризація даних: огляд сучасних методів та підходів. Вісник Київського національного університету імені Тараса Шевченка. 2019. № 4. С. 45–56.
12. Тараненко А. І. Метрики у задачах кластеризації: огляд та рекомендації. Проблеми прикладної математики та інформатики. 2020. № 2. С. 5–15.
13. Гончаренко С. У., Злепко С. М. Комп'ютерне розпізнавання образів. Вінниця : ВНТУ, 2015. 272 с.
14. Панкратова Н. Д., Яшин С. Н. Колірний аналіз зображень у комп'ютерних системах обробки інформації. Київ : Наукова думка, 2006. 198 с.

References:

1. Kolomiiets, A. S., & Marchenko, O. H. (2015). Osnovy klasternoho analizu ta yoho zastosuvannia. Kyiv: National University.
2. Kovalenko, I. V. (2017). Analiz metodiv klasteryzatsii u systemakh obrobky danykh [Analysis of clustering methods in data processing systems]. Journal of Modern Information Technologies, 3(5), 75–81.
3. Skliar, V. M. (2020). Metody klasternoho analizu u mashynnomu navchanni [Methods of cluster analysis in machine learning]. Mathematical Research, 1(4), 112–119.
4. Tsymbal, A. (2002). Vstup do klasternoho analizu: osnovni pidkhody ta metody [Introduction to cluster analysis: basic approaches and methods]. Kharkiv: Naukova Dumka.
5. Vlasenko, O. A. (2016). Metryky dlia klasternoho analizu danykh: teoretychni osnovy ta praktychni aspekty [Metrics for cluster data analysis: theoretical foundations and practical aspects]. Kyiv: KNEU Publishing House.
6. Demchenko, O. B. (2018). Klasteryzatsiia yak metod obrobky danykh u suchasnykh doslidzhenniakh [Clustering as a method of data processing in modern research]. Information Systems and Technologies, 2(8), 50–57.
7. Shvets, K. V. (2024). Doslidzhennia modelei evoliutsii klasteriv v zadachakh rozpiznavannia obraziv [Study of cluster evolution models in pattern recognition tasks]. Retrieved October 30, 2024, from https://openarchive.nure.ua/server/api/core/bitstreams/4b7cf3f6-1df4-4f0f-b180-4beb0ca94454/content
8. Jain, A. K. (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8), 651–666.
9. Marchenko, O. O., & Rossada, T. V. (2017). Aktualni problemy Data Mining [Current problems of Data Mining]. Kyiv.
10. Dudko, O. V., & Poliakov, S. V. (2018). Osnovy klasternoho analizu: metody ta alhorytmy [Fundamentals of cluster analysis: methods and algorithms]. Scientific Notes of NaUKMA. Computer Science, 1, 23–32.
11. Demianchuk, I. S. (2019). Klasteryzatsiia danykh: ohliad suchasnykh metodiv ta pidkhodiv [Data clustering: review of modern methods and approaches]. Bulletin of Taras Shevchenko National University of Kyiv, 4, 45–56.
12. Taranenko, A. I. (2020). Metryky u zadachakh klasteryzatsii: ohliad ta rekomendatsii [Metrics in clustering problems: review and recommendations]. Problems of Applied Mathematics and Informatics, 2, 5–15.
13. Honcharenko, S. U., & Zlepko, S. M. (2015). Kompiuterne rozpiznavannia obraziv [Computer pattern recognition]. Vinnytsia: VNTU.
14. Pankratova, N. D., & Yashyn, S. N. (2006). Kolirnyi analiz zobrazhen u kompiuternykh systemakh obrobky informatsii [Color image analysis in computer information processing systems]. Kyiv: Naukova Dumka.

ZABOLOTNII Yurii,
Lecturer of Specialized Disciplines, Talne Construction and Economic Professional College of Uman National University, Ukraine
LEVCHENKO Alla,
Mathematics Teacher, Talne Secondary School I–III Grades No. 2 of Talne City Council, Cherkasy Region, Ukraine
TYSIACHNA Kateryna,
Student, Talne Secondary School I–III Grades No. 2 of Talne City Council, Cherkasy Region, Ukraine
анізація
STATISTICAL ANALYSIS OF DIGITAL IMAGE COLOR PROFILES AND THEIR CLUSTERING USING THE K-MEANS ALGORITHM

Summary. Introduction. *The rapid growth of visual data volumes necessitates the development of effective methods for automated digital image processing and classification. This paper addresses the problem of grouping images by color similarity through statistical analysis of color profiles and unsupervised clustering. A digital image is formalized as a statistical sample — a set of pixel vectors in a color space – over which numerical characteristics are computed. Two color models, RGB and HSV, are compared with respect to the informativeness of their statistical descriptors and the quality of k-means clustering results.*

Purpose. *The aim of the study is to develop and validate a methodology for statistical analysis of digital image color profiles using the k-means clustering algorithm, and to conduct a comparative analysis of the influence of the color space choice (RGB vs HSV) on the informativeness of statistical features and the quality of image clustering.*

Results. *For each image, the mean, variance, standard deviation, and mode were computed for each color channel in both RGB and HSV spaces. Color profiles were represented as numerical feature vectors used as input to the k-means algorithm. The optimal number of clusters was determined using the Elbow Method (SSE analysis) and the Silhouette Score. Statistical significance of inter-cluster differences was verified with ANOVA. The experimental study on a set of ten test images showed that in the RGB space the optimal number of clusters is k = 2, while in the HSV space k = 3, with higher Silhouette Score values and lower p-values in ANOVA, indicating clearer cluster separation. The HSV space separates hue, saturation, and brightness information explicitly, which leads to more semantically meaningful clusters and better interpretability of statistical characteristics. The software system was implemented in Python using OpenCV, NumPy, scikit-learn, Matplotlib, Pandas, and Tkinter libraries.*

Conclusion. *Statistical analysis of color profiles is an effective tool for quantitative image description. The k-means algorithm combined with quality evaluation metrics (SSE, Silhouette Score, ANOVA) provides reliable clustering of images by color features. The HSV color space is more suitable for color profile clustering than RGB, owing to its better alignment with human visual perception. The developed methodology can be applied in computer vision systems, automated photo sorting, medical diagnostics, and satellite image analysis.*

Keywords: *statistical image analysis, color profile, RGB color space, HSV color space, clustering, k-means clustering algorithm.*

Одержано редакцією 31.10.2025 р.
Прийнято до публікації 17.12.2025 р.