**А. І. Поворознюк[1]**
**В. В. Філатов[1]**
**Г. Є. Філатова[1]**

# РОЗРОБКА КЛАСИФІКАТОРА ЗОБРАЖЕНЬ ДЛЯ ШВИДКОГО ПОШУКУ У ВЕЛИКИХ БАЗАХ ДАНИХ

[1]Національний технічний університет «Харківський політехнічний інститут»

*Стрімке збільшення обсягу інформації в Інтернеті потребує розроблення дієвих методів для її оперативної обробки в інформаційних системах. Зокрема, важливим аспектом є кластеризація новинної інформації не лише з урахуванням морфологічного аналізу текстів, але й візуального контенту. Це створює актуальну задачу кластеризації зображень, що супроводжують текстову інформацію на різних веб ресурсах, таких як новинні портали, соціальні мережі, інформаційні сайти та інші. Саме кластеризація зображень дозволяє ефективніше структурувати великі обсяги даних та спростити процес пошуку потрібної інформації. Предметом цього дослідження є створення класифікатора зображень, який є малочутливим до швидкого зростання обсягу інформації в базах даних. Це особливо важливо в умовах, коли кількість інформації збільшується з високою швидкістю, і потрібно швидко та ефективно обробляти великі масиви даних. Мета дослідження полягає в тому, щоб підвищити продуктивність процесу пошуку ідентичних зображень у великих базах даних, де швидкість додавання нової інформації може сягати 10-12 тисяч зображень на добу. Це вимагає розробки спеціалізованого класифікатора зображень, який зможе забезпечити швидку та точну кластеризацію, незважаючи на інтенсивний ріст обсягу інформації. Для досягнення цієї мети використовуються різні сучасні методи, зокрема математичне моделювання, пошук зображень на основі їхнього контенту, методи обробки зображень, а також методи прийняття рішень. Одним з основних інструментів, застосованих у дослідженні, є двовимірне дискретне косинусне перетворення, яке дозволяє ефективно кодувати зображення та знижувати обсяги даних без втрати важливої інформації. Результати дослідження демонструють, що розроблений класифікатор зображень дійсно є малочутливим до збільшення кількості інформації в базах даних. Проведений аналіз властивостей класифікатора показав, що запропоноване рішення забезпечує високу швидкість обробки даних та мінімальні вимоги до обчислювальної потужності. Експерименти довели, що кластеризація зображень за допомогою даного підходу є досить швидкою і малозатратною з точки зору обсягу використовуваних ресурсів. Розроблений класифікатор може значно підвищити ефективність роботи інформаційних систем, зокрема в умовах постійного зростання інформаційних потоків, що робить його важливим інструментом для обробки великих баз даних зображень.*

**Ключові слова:** інформаційні системи, пошук зображень на основі контенту, класифікатор зображень, великі бази даних, двовимірне дискретне косинусне перетворення.

## Вступ

В даний час збору та обробці інформації приділяється велика увага. Істотний обсяг інформації містять стрічки новин електронних засобів масової інформації (ЗМІ). Одним із способів обробки такої інформації є її кластеризація за смисловим навантаженням [1]. Зазвичай кластеризація інформації проводиться з урахуванням морфологічного аналізу текстів. При цьому слід врахувати, що в стрічках новин ЗМІ практично кожна новина супроводжується графічним контентом (фотоматеріали), тому виникає можливість об'єднувати різні статті новин у кластери не тільки за їх змістом, але і за графічною супутньою інформацією (так званою «візуальною ознакою»). Кластеризація зображень відіграє роль додаткової ознаки в алгоритмі кластеризації текстів, який іноді відіграє навіть більш значущу роль, ніж кластеризація контенту з використанням ключових слів або інших інформаційних ознак (міста, персони, види спорту або інші характеристики, які можна виділити в тексті та призвести до вихідної словоформі). Тут важливо врахувати, що текст може містити більше одного зображення, що має негативний і позитивний ефект. До негативного ефекту можна віднести вставку зображень, що мало відносяться до аналізованих текстів (наприклад, логотипи компаній). При цьому позитивним ефектом є те, що якщо зображення правильно відображають текст, можуть виникнути складні перехресні правила кластеризації, які дозволять об'єднати кластери в дайджест-групи, що містять найбільш повну інформацію про контент. При такій кластеризації можна сформулювати наступну гіпотезу: чим більш унікальне зображення, тім вища ймовірність «схожості» текстового контенту. При цьому заголовки, ключові слова та інша текстова інформація можуть використовуватись як додаткові ознаки кластеризації.

Таким чином, *актуальним* є завдання кластеризації зображень, що супроводжують текстову інформацію на різних сайтах, у тому числі в стрічках новин електронних ЗМІ.

**Аналіз літератури.** Сьогодні існує багато систем для розпізнавання зображень, наприклад, TinEye, Google Similar Images, AntiDupl.NET тощо. Загальним недоліком цих систем є неможливість завантаження галереї зображень та створення власної бази даних (БД) для роботи [2]. При цьому існує велика кількість методів пошуку зображень, які відрізняються різною складністю та ефективністю. У роботах [3, 4] авторами запропоновані інваріантна модель та метод швидкого пошуку цифрового зображення у сховищах даних. Однак запропонована модель [3] має низку недоліків: модель розглядається тільки для напівтонових зображень; модель враховує лише форму гістограми, причому не доведено, що різні напівтонові зображення не можуть мати гістограми однакової форми. Також авторами зазначено, що в середньому необхідно 1,7 секунди для пошуку одного зображення у БД, що містить понад 100 тисяч зображень [4], така швидкість неприйнятна в інформаційній системі, в якій за добу треба класифікувати близько 10-12 тисяч зображень. Така класифікація може тривати понад 5 годин. Одним із сучасних підходів є пошук зображень на основі контенту (CBIR), який відіграє важливу роль у пошуку зображень, схожих на зображення запиту шляхом вилучення візуальних особливостей [5]. В основі цього підходу лежить перетворення зображень на функції низького рівня, що описують аналізовані зображення. Узагальнена структурна схема CBIR систем включає три етапи: вилучення ознак із зображення запиту, вибір ознак і потім зіставлення подібності [6]. Системою CBIR формуються багатовимірні вектори ознак, які зіставляються із векторами зображень у базі даних. За способом обробки інформації система CBIR ділиться на онлайн і офлайн підсистеми, які мають один і той же блок отримання ознак [6]. Для зіставлення векторів використовуються відповідні міри подібності або відстані. Отримана оцінка подібності порівнюється з граничним значенням, яке попередньо визначається CBIR системою. Одним з основних недоліків методів CBIR є їх низька продуктивність (тобто використання пам'яті, масштабованість, швидкість, точність) за рахунок використання багатовимірних функцій переведення вмісту візуального зображення в числову форму та складних алгоритмів прийняття рішень, наприклад, таких як машинне навчання [7] або генетичні алгоритми [8]. Авторами [6] зазначено, що найбільш широко відомі методи кластеризації, такі як k-means або метод об'єднання в пари найближчих сусідів [9], є занадто повільними для задач цифрової обробки зображень. Крім того, в цих методах необхідно заздалегідь задавати число класів, що робить їх неприйнятними для завдання кластеризації зображень у великих базах, де інформація постійно додається.

Крім методів отримання ознак і алгоритмів прийняття рішень на продуктивність систем CBIR також впливає спосіб пошуку зображень у БД, особливо у великих, що налічують десятки тисяч записів. Методи пошуку БД багатовимірного вектору ознак зазвичай діляться на два типи [6]: прямий пошук і пошук з використанням функцій. У методах прямого пошуку базі даних вектор ознак є безпосередньо індексом вихідного зображення [10]. У методах пошуку з використанням функцій виконується відображення векторів ознак великої розмірності з плаваючою комою у вектори низької розмірності або двійкові вектори, що зменшує обчислювальну складність відстані або подібності, а також вартість зберігання. У другій групі методів часто використовується хешування [11], яке дозволяє використовувати прості метрики. Слід зазначити, що у практичних реалізаціях систем CBIR часто поєднуються обидва типи методів пошуку у базі даних. Проведений аналіз показав, що існуючі методи пошуку зображень у базах даних малопридатні для вирішення задачі кластеризації зображень, що супроводжують текстову інформацію на різних сайтах, через їх низьку продуктивність у разі необхідності пошуку у великих базах даних, які постійно оновлюються.

**Постановка задачі та мета дослідження.** *Метою роботи є* підвищення продуктивності пошуку «подібних» зображень у великих базах даних, у яких швидкість додавання інформації досягає 10-12 тисяч зображень на добу.

Для реалізації поставленої мети необхідно вирішити такі завдання:

– розробити класифікатор зображень, що малочутливий до зростання кількості інформації в БД;

– виконати дослідження властивостей розробленого класифікатора зображень.

## Результати дослідження

**Розробка класифікатора зображень.** Загальну модель класифікатора зображень можна подати у вигляді кортежу

$$IC = \langle I, C, S, R, O \rangle, \quad (1)$$

де $I = \{I_1, I_2, \ldots, I_P\}$ — множина зображень, які необхідно класифікувати (колекція зображень); $C = \{C_1, C_2, \ldots, C_K\}$ — множина кластерів (класів зображень), причому $C_i \cap C_j = \varnothing \quad \forall i \neq j$; $S = \{\overrightarrow{S_1}, \overrightarrow{S_2}, \ldots, \overrightarrow{S_L}\}$ — множина сигнатур зображень; $R \subset C \times S$ — відношення між кластерами та сигнатурами; $O : I \to C$ — операція кластеризації, яка полягає у перетворенні зображень, після яких, або зображення $I_n \in I$ з сигнатурою $\overrightarrow{S_l} \in S$ відноситься до існуючого кластера $C_k \in C$, або робиться висновок про необхідність створення нового кластера $C_{K+1} \in C$, до якого можна віднести це зображення, при цьому одне зображення може бути віднесене тільки до одного кластера.

Сигнатура зображення – це вектор значень ознак, що використовуються для однозначної класифікації зображення. Відношення *R* має таку властивість: $\forall C_i \in C \exists \overrightarrow{S_j} \in S : \left(C_i, \overrightarrow{S_j}\right) \in R$.

З запропонованої моделі (1) видно, що у швидкість роботи класифікатора насамперед впливає метод обчислення сигнатури зображення. У роботі пропонується використовувати двовимірне дискретне косинусне перетворення (ДКП) для обчислення сигнатури зображення. Двовимірне ДКП має низку корисних властивостей, які можна використовувати при розрахунку сигнатури зображення. Крім того, існує швидкий алгоритм обчислення ДКП.

У стрічках новин електронних ЗМІ використовуються як кольорові, так і чорно-білі (напівтонові) зображення, тому в роботі пропонується використовувати наступну модель зображення

$$Img = \langle \mathbf{A}_{gs}, \mathbf{A}_r, \mathbf{A}_g, \mathbf{A}_b \rangle, \quad (2)$$

де $\mathbf{A}_{gs}, \mathbf{A}_r, \mathbf{A}_g, \mathbf{A}_b$ — квадратні матриці напівтонової та колірних складових зображення відповідно (колірна модель RGB). Для напівтонових зображень матриці колірних складових $\mathbf{A}_r, \mathbf{A}_g, \mathbf{A}_b$ заповнюються нулями. Для кольорових зображень матриця напівтонової складової $\mathbf{A}_{gs}$ обчислюється з урахуванням матриць колірних складових $\mathbf{A}_r, \mathbf{A}_g, \mathbf{A}_b$.

Позначимо двовимірне ДКП кожної складової вихідного зображення як $\mathbf{F}_{gs}, \mathbf{F}_r, \mathbf{F}_g, \mathbf{F}_b$. З урахуванням властивостей коефіцієнтів ДКП у роботі пропонується виконати квантування коефіцієнтів з метою аналізу лише частотних компонент, які перевищили заданий поріг. Задамо для кожної складової $\mathbf{F}_{gs}, \mathbf{F}_r, \mathbf{F}_g, \mathbf{F}_b$ матриці квантування коефіцієнтів ДКП $\mathbf{Q}_{gs}, \mathbf{Q}_r, \mathbf{Q}_g, \mathbf{Q}_b$. Тоді квантування виконується шляхом поелементного поділу кожної матриці $\mathbf{F}_{gs}, \mathbf{F}_r, \mathbf{F}_g, \mathbf{F}_b$ на відповідну матрицю $\mathbf{Q}_{gs}, \mathbf{Q}_r, \mathbf{Q}_g, \mathbf{Q}_b$, значення елементів якої зростають у міру віддалення від лівого верхнього кута і наближення до правого нижнього кута $F_{(q)}[u,v] = F[u,v] / Q[u,v]$.

З урахуванням розміщення частотних компонентів у матриці $\mathbf{F}_{(q)}$ перетворимо кожну з матриць $\mathbf{F}_{(q)gs}, \mathbf{F}_{(q)r}, \mathbf{F}_{(q)g}, \mathbf{F}_{(q)b}$ у відповідні одновимірні вектори $\vec{f}_{gs}, \vec{f}_r, \vec{f}_g, \vec{f}_b$ (зигзагоподібне сканування матриці, починаючи з лівого верхнього кута). З кожної частотної компоненти можна скласти наступний вектор $\vec{f}_i = (f_{gs}[i], f_r[i], f_g[i], f_b[i])$, $i = \overline{0, M-1}$, де $M \square N^2$, тобто пропонується враховувати лише перші *M* квантованих частотних компонентів. Тоді довжина вектору $\vec{f}_i$ обчислюється як

$$\left|\vec{f}_i\right| = \sqrt{f_{gs}^2[i] + f_r^2[i] + f_g^2[i] + f_b^2[i]}.$$

Тому що орієнтація вихідного зображення невідома (зображення може бути дзеркально відображено і/або повернене на 90°), то з урахуванням властивостей ДКП для обчислення сигнатури, інваріантної до поворотів і дзеркальним відображенням зображення по будь-якій осі, в роботі пропонується вектор $\vec{s} = (s_0, s_2, \ldots s_{M-1})$, елементи якого обчислюються так

$$s_i = \frac{\left|\vec{f}_i\right| + \left|\vec{f'}_i\right|}{2}, \; i = \overline{0, M-1}. \quad (3)$$

Тоді сигнатура $\overrightarrow{S_l} = (S_{l0}, S_{l2}, \ldots, S_{lM-1})$ зображення $I_n$ з моделлю (2) може бути обчислена як

$$S_{li} = round\left(\frac{w_i S_i}{|\vec{s}|}\right),\ i = \overline{0, M-1}, \tag{4}$$

де $w_i$ — вагові коефіцієнти.

Як міра подібності двох векторів $\vec{a}$ і $\vec{b}$ пропонується метрика міських кварталів

$$d(\vec{a},\vec{b}) = \sum_i |a_i - b_i|. \tag{5}$$

Тоді нове зображення $I_n \in I$ з сигнатурою $\overrightarrow{S_n} \in S$ відноситься до існуючого кластера $C_k \in C$, якщо існує така сигнатура $\overrightarrow{S_l}$, відстань до якої мінімальна серед усіх сигнатур і менше заданого порога $\varepsilon_k$. Інакше формується новий кластер $C_{K+1}$ із зображенням $I_n \in I$. Таким чином, вирішальне правило класифікатора має такий вигляд

$$I_n \in \begin{cases} C_k, & \text{якщо } \exists\left(C_k, \overrightarrow{S_l}\right): d(\overrightarrow{S_n}, \overrightarrow{S_l}) = \min\limits_{\substack{p \neq n,\\ p \in \overline{1,P}}} d(\overrightarrow{S_n}, \overrightarrow{S_p}) < \varepsilon_k, 1 \le k \le K; \\ C_{K+1}, & \text{в іншому випадку.} \end{cases} \tag{6}$$

**Дослідження властивостей розробленого класифікатора.** Для розробки швидкого алгоритму класифікації зображень за допомогою запропонованого методу розглянемо деякі його властивості. Для експериментів враховуватимемо всі коефіцієнти ДКП, тобто матриці квантування $\mathbf{Q}_{gs}, \mathbf{Q}_r, \mathbf{Q}_g, \mathbf{Q}_b$ це матриці одиниць. Характеристики набору даних для експериментів: кількість зображень – 804724; розміри зображень різні; всі зображення кольорові; число кластерів зображень – 445999, при цьому 172344 (38,64%) кластерів містить по 2 зображення, решта кластерів – від 3 до 6828 зображень.

Для використання швидкого способу розрахунку ДКП необхідно, щоб зображення були квадратними і мали однаковий розмір. Дослідження показали, що при зменшенні зображення до розмірів $64 \times 64$ пікселів (методом кубічної інтерполяції) сигнатури вихідного та зменшеного зображень не змінюються. У всіх експериментах довжина сигнатури $M = 10$. Аналіз значень вектору $\vec{s} = (s_0, s_2, \ldots s_9)$ показав, що нульове значення на порядок перевищує значення всіх інших елементів вектору, тому як вагові коефіцієнти у формулі (4) використовувався наступний вектор $\vec{w} = (0.1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1)$. Експерименти показали, що подальше зменшення розмірів зображень призводить до того, що сигнатури починають трохи відрізнятися. Тому на першому етапі попередньої обробки зображення в роботі пропонується приводити всі зображення до одного розміру $64 \times 64$ пікселів. При цьому сигнатури зберігають свою здатність ідентифікувати конкретні зображення. Якщо у вирішальному правилі (6) $\varepsilon_k = 0$, тоді одні й самі кластери будуть зібрані повністю однакові зображення або зображення, які візуально не можна відрізнити. Аналіз зображень показав, що збільшення порога понад 40 призводить до різкого збільшення помилок кластеризації зображень, при цьому кількість кластерів різко починає зменшуватися (тобто кластери починають об'єднуватися), а при порозі $\varepsilon_k < 10$ до одного кластера потрапляють схожі зображення (рис. 1).


(46, 31, 31, 45, 9, 45, 20, 23, 23, 20) (46, 30, 30, 45, 9, 45, 20, 24, 24, 20) (46, 30, 30, 45 , 9, 45, 21, 23, 23, 21)

Рис. 1. Приклади схожих зображень, які можуть бути об'єднані в один кластер, та їх сигнатури

Такі зображення у разі нечіткої кластеризації може бути об'єднані до одного кластеру за допомогою розробленого класифікатора, тому що надалі кластеризація текстового контенту здійснюється не тільки на підставі зображень, але й інших параметрів текстів, таких як дата новини, заголовок, ключові слова тощо.

**Швидкий пошук у великих базах даних.** При надходженні нового зображення в інформаційну систему (наприклад, портал новин) для класифікації виконуються такі основні кроки:

1) зменшення розмірів зображення до $64 \times 64$;

2) розрахунок сигнатури за виразом (4) з урахуванням (3);

3) класифікація згідно з вирішальним правилом (6) з урахуванням метрики (5).

4) запис нового зображення в БД із зазначенням існуючого кластера, якщо такий знайдено, або нового кластера, якщо зображення унікальне.

Для виконання кроку 3 спочатку потрібно спробувати знайти вже існуючий кластер БД. Розглянемо випадок, як у вирішальному правилі (6) для всіх кластерів $\varepsilon_k = 0$. Сучасні реляційні СУБД дозволяють виконувати швидкий пошук індексованих полів. Тоді для реалізації швидкого пошуку необхідно в реляційній БД виконати індексацію поля, що містить сигнатуру.

Згідно з запропонованою моделлю класифікатора (1), сигнатура – це вектор, що містить кілька значень, тому необхідно перетворити цей вектор в одне поле, за яким буде проіндексована БД, а надалі проводитиметься пошук. В роботі пропонується виконати перетворення вектору $\overrightarrow{S_l} = (S_{l0}, S_{l2}, \ldots, S_{lM-1})$ в рядок наступного виду $S_{l0} : S_{l2} : \ldots : S_{lM-1}$, тобто послідовно записати елементи вектору, розділені двокрапкою. Наприклад, значення рядкового поля для сигнатури зображення може мати наступний вигляд 47:42:42:36:23:36:12:22:22:12. Таке уявлення дозволяє записувати сигнатуру в одному полі, незалежно від значення *M* (кількість елементів у векторі). Тоді номер кластера Nklaster із зображеннями, у яких сигнатура imsig збігається із сигнатурою нового зображення newimsig, можна отримати простим запитом до таблиці news_images:

```
SELECT Nklaster FROM news_images WHERE imsig = newimsig
```

Такий пошук буде здійснено за мінімально можливий час для таблиць з індексним полем. Експерименти показали, що у середньому швидкість класифікатора становила 0,03 сек/зображення. Таким чином, при надходженні на добу до 10-12 тисяч зображень на їх кластеризацію витрачається до 5-6 хвилин.

## Висновки

Розроблено загальну модель класифікатора зображень на основі контенту, в основі якої лежить перетворення зображень до його сигнатури – вектору значень ознак, що використовуються для однозначної класифікації зображення. Виходячи із запропонованої моделі, показано, що у швидкість роботи класифікатора насамперед впливає метод обчислення сигнатури зображення. У роботі пропонується використовувати двовимірне ДКП для обчислення сигнатури зображення. На основі дослідження властивостей ДКП при аналізі напівтонової та колірних складових зображень відповідно (колірна модель RGB) отримані математичні вирази обчислення компонентів сигнатури, інваріантної до поворотів та дзеркальних відображень зображення по будь-якій осі. Розроблено метод класифікації зображень шляхом порівняння їх сигнатур при використанні метрики міських кварталів як міру подібності. Отриманий метод є науковою новизною цього дослідження.

Виконано дослідження властивостей розробленого класифікатора зображень, у результаті якого обґрунтовано перетворення всіх зображень одного розміру $64 \times 64$ пікселів. При цьому їх сигнатури зберігають свою здатність ідентифікувати конкретні зображення, а матриця коефіцієнтів ДКП розраховується 1 раз для всіх зображень, що суттєво знижує трудомісткість методу. Проведені експерименти показали, що кластеризація інформації за зображеннями виявилася досить швидкою та маловитратною з погляду обсягів інформації та вимог до обчислювальної потужності.

Подальші дослідження спрямовано пошук оптимальних параметрів запропонованого класифікатора, і навіть вивчення можливості використання класифікатора для нечіткої кластеризації, і навіть нечіткого пошуку у базі даних.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] О. А. Амонс, Ю. О. Янов, та І. О. Безпалий, "Кластеризація документів на основі статистичної близькості термів", *Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техніка*, №49, 2008, с. 55-62.

[2] О. М. Верес, Я. П. Кісь, В. А. Кугівчак, та І. В. Рішняк, "Вибір методів для пошуку однакових або схожих зображень",

*Вісник Національного університету Львівська політехніка, Серія: Інформаційні системи та мережі*, № 887, 2018, с. 43-50.
[3] К. С. Смеляков, А.С. Чуприна, Д. Л. Сандркін, Є. В. Вакулік, та Є. М. Дроб, "Розробка інваріантної моделі цифрового зображення для швидкого пошуку у сховищах даних", *Збірник наукових праць Харківського національного університету Повітряних Сил*, № 2(68), 2021, с. 108-15.
[4] К. С. Смеляков, Д. Л. Сандркін, Д. О. Товчиречко, Є. В. Вакулік, та Є. М. Дроб, "Розробка методу швидкого пошуку цифрового зображення у сховищах даних", *Системи обробки інформації*, № 2(165), 2021, с. 54-63.
[5] F. A. A. Salih, and A. A. Abdulla, "An Efficient Two-layer based Technique for Content-based Image Retrieval", *UHD Journal of Science and Technology*, № 5(1), 2021, p. 28-40.
[6] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma, "Recent developments of content-based image retrieval (CBIR)", *Neurocomputing*, Volume 452, 2021, p. 675-689.
[7] A. Sezavar, H. Farsi, and S. Mohamadzadeh, "Content-based image retrieval by combining convolutional neural networks and sparse representation", *Multimedia Tools and Applications*, vol. 78(15), 2019, p. 20895-20912.
[8] M. K. Alsmadi, "Content-based image retrieval using color, shape and texture descriptors and features", *Arabian Journal for Science and Engineering*, vol. 45(4), 2020, p. 3317–3330.
[9] N. Sampathila, and R. J. Martis, "Computational approach for content-based image retrieval of K-similar images from brain MR image database", *Expert Syst.*, vol. 39, 2022, e12652.
[10] Junjie Cai, Qiong Liu, Francine Chen, Dhiraj Joshi, and Qi Tian, "Scalable Image Search with Multiple Index Tables". in *Proceedings of International Conference on Multimedia Retrieval (ICMR '14), Association for Computing Machinery*, New York, NY, USA, 2014, p. 407-410.
[11] S. Cheng, L. Wang, and A. Du, "An Adaptive and Asymmetric Residual Hash for Fast Image Retrieval," in *IEEE Access*, vol. 7, 2019, p. 78942-78953.

***Поворознюк Анатолій Іванович*** — д-р техн. наук, професор, професор кафедри комп'ютерної інженерії та програмування, e-mail: anatolii.povorozniuk@khpi.edu.ua;
***Філатов Валерій Володимирович*** — аспірант кафедри комп'ютерної інженерії та програмування.
***Філатова Ганна Євгенівна*** — д-р техн. наук, професор, професор кафедри комп'ютерної інженерії та програмування.
Національний технічний університет «Харківський політехнічний інститут», Харків.

**A. I. Povoroznyuk**[1]
**V. V. Filatov**[1]
**A. E. Filatova**[1]

# Designing an Image Classifier for Fast Search in Large-Scale Databases

[1]National Technical University "Kharkiv Polytechnic Institute"

*The rapid increase in the volume of information on the Internet necessitates the development of effective methods for its real-time processing in information systems. In particular, an important aspect is the clustering of news information, which considers the morphological analysis of texts and visual content. This creates a relevant challenge in clustering images accompanying textual information on various web resources, such as news portals, social networks, and informational websites. Image clustering allows for more efficient structuring of large data sets and simplifies searching for the necessary information. The subject of this study is the creation of an image classifier that is resistant to the rapid growth of information in databases. This is especially important in conditions where the amount of information is increasing at a high rate, and there is a need to quickly and efficiently process large data arrays. The study aims to enhance the performance of searching for identical images in large databases, where the rate of new information addition can reach 10–12 thousand images per day. This requires the development of a specialized image classifier that can provide fast and accurate clustering despite the intense growth in the volume of information. To achieve this goal, various modern methods are used, including mathematical modeling, content-based image retrieval, image processing techniques, and decision-making methods. One of the main tools applied in the study is the two-dimensional discrete cosine transform, which allows for efficient image encoding and data compression without loss of important information. The research results demonstrate that the developed image classifier is indeed resistant to the increase in the amount of information in databases. The analysis of the classifier's properties showed that the proposed solution ensures high data processing speed and minimal computational requirements. Experiments have proven that image clustering using this approach is relatively fast and resource-efficient. The developed classifier can significantly improve the efficiency of information systems, particularly in the context of the constant growth of information flows, making it an important tool for processing large image databases.*

**Keywords**: Information systems, content-based image retrieval, image classifier, large databases, two-dimensional discrete cosine transform.

***Povoroznyuk Anatolii Ivanovych*** — Doctor of Science, Professor, Professor of the Computer Engineering and Programming Department, e-mail: anatolii.povorozniuk@khpi.edu.ua;

***Filatov Valerii Volodyrymovuch*** — Post-Graduate Student of the Computer Engineering and Programming Department;

***Filatova Anna Evgenivna*** — Doctor of Science, Professor, Professor of the Computer Engineering and Programming Department