УДК 621.39; 004.9; 004.92

О. С. Прокопенко; С. В.Смеляков, д-р техн. наук, проф.

РОЗРОБКА ЕФЕКТИВНОЇ СИСТЕМИ ПОШУКУ ЗОБРАЖЕНЬ НА ОСНОВІ ЗМІСТУ, ЗАСНОВАНОГО НА ВИЯВЛЕНИХ ОБ'ЄКТАХ У СХОВИЩАХ ВЕЛИКИХ ДАНИХ

Об'єктом дослідження є пошук зображень на основі вмісту. Предметом дослідження є моделі та методи пошуку зображень на основі вмісту (CBIR), а також керування великими обсягами медіаконтенту в масштабних системах зберігання зображень. Метою роботи є розроблення ефективної системи пошуку зображень на основі вмісту з використанням сучасних моделей комп'ютерног озору для виявлення об'єктів. Результати виявлення об'єктів використовуються для побудови бази даних, що містить дескриптори зображень. Запропонована система пошуку зображень на основі вмісту спрямована на підвищення ефективності та точності процесів пошуку та керування зображеннями. Розроблено CBIR-систему, яка здійснює пошук зображень на основі об'єктів, виявлених за допомогою сучасних моделей машинного навчання; проведено серію експериментів для оцінювання ефективності та якості пошуку у великих сховищах зображень із використанням запропонованої CBIR-системи. Експерименти порівняли її продуктивність з наявними методами, підкресливши її сильні сторони та обмеження, а саме: швидше створення дескрипторів, швидше порівняння дескрипторів порівняно з хеш-базованими, ручними та глибинними дескрипторами; ефективне фільтрування даних у сховищі зображень на основі вмісту об'єктів, що дає змогу здійснювати цільовий пошук; вища якість і швидкість пошуку зображень у великих депозитаріях даних порівняно з аналогами. Водночас ефективність системи значною мірою залежить від якості моделі та даних, використаних для виявлення об'єктів, оскільки зображення без виявлених об'єктів не з'являтимуться в результатах пошуку, що потенційно може обмежувати повноту вибірки.

Ключові слова: *обробка зображень, класифікація та детекція, машинне навчання, дескриптор зображення, пошук зображень на основі вмісту (CBIR), великі дані, обробка зображень, зберігання зображень, оптимізація пошуку зображень, інформаційні технології.*

Вступ

Люди отримують більшість інформації через зір. З розвитком комп'ютерних технологій значна увага була зосереджена на вилученні, формалізації та використанні візуальних даних для аналізу та ухвалення рішеньу системах комп'ютерного зору, моніторингу, підтримки прийняття рішень та штучного інтелекту [1].

Насиченість візуальної інформації створює високі вимоги до пам'яті та обчислювальних ресурсів, що ускладнює обробку в режимі реального часу. Тому зображення мають бути представлені у вигляді дескрипторів – компактних векторів ознак, які зменшують обсяг даних, водночас зберігаючи здатність розпізнавати зображення в базі даних [2]. Оскільки час і увага людини є фундаментально обмеженими ресурсами, лише невелика частина завантажених зображень буде фактично переглянута згодом [3].

Сучасним трендом у пошуку зображень за контентом є використання дескрипторів зображень на основі глибокого навчання, що використовують можливості згорткових нейронних мереж (CNN) та інших нейронних архітектур для вилучення високодискримінативних ознак із зображень. На відміну від вручну створених дескрипторів, які спираються на заздалегідь визначені алгоритми, моделі глибокого навчання навчаються безпосередньо на даних, захоплюючи як низькорівневі деталі (наприклад, краї, текстури), так і високорівневу семантичну інформацію (наприклад, об'єкти, сцени). Такі дескриптори генеруються шляхом пропускання зображення через навчену мережу та вилучення карт ознак або вбудовувань (embeddings) з проміжних шарів [4].

Зокрема, у задачах пошуку зображень низка підходів безпосередньо використовує активації мережі як ознаки зображення та успішно виконує пошук. Популярні підходи

включають використання попередньо навчених моделей, таких як VGG, ResNet та EfficientNet, як екстракторів ознак, або навчання спеціалізованих мереж, адаптованих під конкретні задачі. Дескриптори на основі глибокого навчання перевершують традиційні методи в задачах пошуку зображень, розпізнавання об'єктів та семантичної сегментації, оскільки вони здатні добре узагальнюватися на різноманітних наборах даних і адаптуватися до складних візуальних закономірностей [5]. Водночас вони часто потребують значних обчислювальних ресурсів для навчання та інференсу, а їхня ефективність сильно залежить від якості та різноманітності навчальних даних. Незважаючи на ці виклики, здатність витягувати багаті ієрархічні ознаки робить їх основою сучасних систем комп'ютерного зору. Вони знаходять застосування в багатьох галузях науки і техніки.

Постановка задачі. Незважаючи на високу ефективність пошуку зображень за змістом з використанням дескрипторів зображень на основі глибокого навчання, високовимірні вектори ознак потребують значних обчислювальних ресурсів для зберігання та часу для обробки даних і порівняння зображень.

У той час коли часові обмеження не грають великої ролі у невеликих користувацьких репозиторіях зображень, використання пошуку зображень за змістом у корпоративних додатках, що оперують великими сховищами даних або за наявності обмежених ресурсів, наприклад з IoT, мають значно суворіші вимоги до використовуваних ресурсів, часу побудови дескрипторів та виконання пошукових запитів за зображеннями.

Зважаючи на наведені вище недоліки, враховуючи стрімке зростання сфери штучного інтелекту та іновацій у цій галузі, виникає припущення про можливість розробки нових моделей та методів створення дескрипторів зображень, що використовують гібридний підхід з обмеженою кількістю даних про об'єктний склад зображення, що на відміну від домінуючих на сьогодні високорівневих векторів ознак дозволять інженерам будувати системи пошуку зображень за змістом, що зберігатимуть високорівневу контекстну інформацію, проте матимуть швидкість обробки притаманну найпростішим хеш-дескрипторам.

Мета та підхід

Метою роботи є підвищення швидкодії оброблення запитів і точності видачі результатів під час пошуку зображень за рахунок розроблення та експериментальної перевірки системи пошуку зображень за змістом для великих сховищ даних. У цій роботі описано розробку CBIR-системи, що використовує дескриптори, отримані на основі об'єктів, виявлених на зображеннях за допомогою моделей на кшталт YOLO. Дескриптори кодують тип об'єкта, його розмір і просторове розташування, що дає змогу виконувати завдання, такі як пошук за тегами та пошук за зображенням. Такий підхід забезпечує швидке обчислення з мінімальним обсягом даних, водночас надаючи багату контекстну інформацію для точного порівняння зображень. Завдяки фокусуванню на об'єктах як інформаційно насичених одиницях система досягає кращої точності пошуку, ніж традиційні методи, і працює швидше, ніж моделі з високорозмірними ознаками глибинного навчання. Найкращі результати можуть додатково уточнюватися за допомогою більш обчислювально затратних методів для підвищення точності.

Важливою інновацією дескриптора є поєднання ефективності низькорівневих дескрипторів ознак із високорівневою семантичною інформативністю, яку забезпечують CNN, що створює масштабовану та надійну CBIR-систему з високою швидкістю та точністю. На відміну від наявних високорівневих векторів ознак такі дескриптори займають значно менший обсяг пам'яті і час на порівняння зображень, проте все ще є стійкими до багатьох трансформацій, зміни кутів та освітлення.

Підтримувальні алгоритми використовують ці об'єктно-орієнтовані дескриптори, зокрема пошук за схожістю на основі типів об'єктів і їхніх просторових розташувань,

інвертований індекс для ефективних запитів за тегами, а також просторову верифікацію для уточнення результатів шляхом порівняння розміщення об'єктів. Групування подібних дескрипторів додатково прискорює пошук у масштабних наборах даних, забезпечуючи масштабованість як для пошуку за зображенням, так і за тегами.

Для досягнення цієї мети необхідно виконати такі завдання:

– Визначити поля даних для збережених об'єктів, щоб забезпечити достатню інформацію для ефективного пошуку;

– Розробити гнучку структуру зберігання дескрипторів, адаптовану до різних моделей детекції об'єктів та налаштовувану для конкретних застосувань;

– Обрати відповідну модель детекції об'єктів для проведення експериментів, що забезпечить справедливе порівняння з наявними дескрипторами;

– Розробити програмне забезпечення для побудови бази даних дескрипторів та реалізації ефективних алгоритмів порівняння, що підтримують як пошук за тегами, так і пошук за зображенням.

З реалізацією дескриптора та підтримувального програмного забезпечення CBIR-система забезпечує точний та ефективний пошук у великих наборах зображень, використовуючи обрану модель детекції об'єктів і набір даних, що дозволяє отримувати універсальний високопродуктивний пошук зображень для широкого спектра застосувань.

Вибір моделі для детекції об'єктів

Детекція об'єктів є важливим компонентом комп'ютерного зору та знаходить застосування у відеоспостереженні [6], медичній візуалізації [7] та навігації роботів [8]. Поширені методи включають віднімання фону, часові різниці, оптичний потік, фільтрацію Калмана, SVM та порівняння контурів [9].

Для цієї роботи обирається нейронна мережа, спеціалізована на детекції об'єктів, для вилучення інформації зображень. Вибрана модель повинна відповідати наступним критеріям:

– бути широко використовуваною та часто застосовуваною для задач детекції об'єктів, забезпечуючи перевірену ефективність у різних умовах;

– бути зручною у використанні, здатною виявляти багато об'єктів і легко перенавчатися для адаптації до нових класів або спеціалізації на певних типах об'єктів;

– пропонувати кілька версій з різними параметрами, швидкістю та точністю, що дозволяє впровадження як на пристроях з обмеженими ресурсами (наприклад, мобільних), так і на високопродуктивних системах.

Виходячи з цих критеріїв, обраноYOLOv8 для інтеграції у систему керування великими сховищами зображень. YOLOv8 є провідною моделлю для детекції об'єктів, пропонуючи ефективний баланс між швидкістю, точністю та гнучкістю. Головною перевагою є швидкість – у 2 – 2,5 рази швидше за популярні моделі, такі як Faster R-CNN, SSD та RetinaNet – за збереження порівнянної або вищої точності. На відміну від двоетапних моделей R-CNN, YOLOv8 використовує одноетапну архітектуру детекції.

Методологія пошуку

На першому етапі використовується нейронна мережа для детекції об'єктів, щоб ідентифікувати об'єкти на кожному зображенні в репозиторії та побудувати базу данихд ескрипторів зображень на основі об'єктів.

Далі застосовуються розроблені метрики та алгоритми порівняння для вимірювання схожості між дескрипторами, підтримувані пошуковим алгоритмом, який використовує ці метрики для пошуку зображень.

Загальний процес пошуку виглядає наступним чином:

1. Налаштування пошуку: Визначення критеріїв пошуку та встановлення параметрів (наприклад, пороги схожості, типи об'єктів).

2. Побудова дескриптора: Генерація дескриптора для зображення-запиту.

3. Виконання пошуку:

– Ітеративне проходження по дескрипторах у репозиторії;

– Порівняння кожного дескриптора з дескриптором зображення – запиту за допомогою метрики схожості;

– Сортування зображень за показником схожості.

В ідеалі відсортовані результати пошуку спершу відображатимуть зображення-запит (якщо воно присутнє), за ним – його варіації або трансформації (наприклад, змінені за розміром, обрізані абов ідредаговані), а потім інші зображення репозиторію, ранжовані за ступенем схожості із запитом.

Нажаль така система містить і свої недоліки, оскільки є залежною від моделі детекції об'єктів, адже зображення без знайдених об'єктів будуть невидимі для неї, а у випадку коли модель на трансформованому зображенні знаходить об'єктний склад відмінний від оригіналу, таке зображення може не потрапити до найкращих результатів. Незважаючи на це, при правильному підборі моделі до наявних даних, вона дозволятиме фільтрувати і шукати зображення у сховищах зі швидкістю притаманною хеш-дескрипторам та точністю дескрипторів на основі глибокого навчання.

Моделі дескрипторів

Дескриптор зображення – це компактне подання, яке зберігає основну інформацію про зображення в репозиторії. Це включає: розташування зображення та його формальні параметри (наприклад, роздільна здатність, формат) та інформацію про об'єкти, виявлені на зображенні, а саме: їхні координати (координати обмежувальної рамки), розміри та площі, а також рівень впевненості моделі комп'ютерного зору, що обмежувальна рамка містить об'єкт певного типу (наприклад, тип і).

Кожен об'єкт представлено вектором із семи чисел, як показано в Таблиці 1.

Таблиця 1

Модель дескриптора об'єкта

Назва	Опис
x	Відносна координата центру обмежувальної рамки об'єкта по осі x (діапазон 0–1)
y	Відносна координата центру обмежувальної рамки об'єкта по осі y (діапазон 0–1)
w	Відносна ширина об'єкта (діапазон 0–1)
h	Відносна висота об'єкта (діапазон 0–1)
area	Відносна площа об'єкта
ratio	Відношення меншої сторони до більшої сторони обмежувальної рамки об'єкта
conf	Впевненість нейронної мережі, що рамка містить об'єкт класу і (діапазон 0–1)

Дескриптор класу представляє всі об'єкти певного класу на зображенні, містячи окремі дескриптори об'єктів та додаткову інформацію: кількість об'єктів, загальну площу та центр мас групи (Таблиця 2).

Попереднє обчислення цих даних дозволяє уникнути зайвих розрахунків під час порівнянь, скорочуючи час обробки та підвищуючи ефективність.

Таблиця 2

Модель дескриптора класу

Назва	Опис
class	Ідентифікатор класу
number	Кількість об'єктів цього класу на зображенні
area	Сума відносних площ об'єктів цього класу на зображенні
center	Арифметичний центр групи об'єктів цьогокласу на зображенні
objects	Колекція дескрипторів об'єктів цього класу на зображенні

Дескриптор зображення містить основну інформацію для ідентифікації та опису зображення, включаючи його розташування, розміри та дескриптори класів. Його структура є гнучкою, що дозволяє за потреби додавати додаткову інформацію про зображення або об'єкти (Таблиця 3).

Таблиця 3

Модель дескриптора зображення

Назва	Опис
path	Шлях до файлу зображення
width	Ширина зображення
height	Висота зображення
classes	Дескриптори класів знайдених на зображенні

Метрика схожості дескрипторів

Пропонується використовувати відстань між зображеннями як метрику: ідентичні зображення мають нульову відстань, тоді як більші відстані свідчать про нижчу схожість. Метод повинен бути стійким до поширених трансформацій зображень – таких як зміни яскравості, кольору, стиснення та масштабування – спираючись на відносні, а не абсолютні розміри зображень та об'єктів.

Щоб досягти цього, спочатку враховуються пропорції зображення, оскільки масштабування або зміна розміру не впливає на них. Пропорція визначається як відношення меншоїсторони до більшої. Перший компонент метрики, позначений як Δp, є абсолютною різницею цих пропорцій, де h – менша сторона, а w – більша.

$$\Delta \mathrm{p} = \left| \frac{h_1}{w_1} - \frac{h_2}{w_2} \right| \quad (1)$$

Наступний компонент метрики порівнює параметри об'єктів, зокрема відношення площ об'єктів та їх центроїди. Центроїд об'єктів у класі обчислюється як арифметичне середнє координат усіх об'єктів i-го класу. Кожне зображення може містити нуль або більше об'єктів класу i, і для порівняння груп між двома зображеннями використовуються центроїди x_{ij}, y_{ij} класу і у зображенні j.

Відстань d_i обчислюється між центроїдами об'єктів і-го класу на двохзображеннях за допомогоюевклідовоївідстані. Цеформуєдругий компонент метрики, Δo_i, що відображає різницю зображень на основі цього класу. Він обчислюється шляхом множення відстані між центроїдами на різницю загальних площ об'єктів і додавання штрафу, якщо два зображення містять різну кількість об'єктів і-го класу.

$$d_i = \sqrt{(x_{i1} - x_{i2})^2 + (y_{i1} - y_{i2})^2} \quad (2)$$

$$\Delta o_i = d_i * |area_{i1} - area_{i2}| * (|num_{i1} - num_{i2}| + 1) \quad (3)$$

Формула (2) обчислює відстань між центроїдами об'єктів і-го класу для двох зображень, тоді як формула (3) визначає відповідну відстань між зображеннями на основі цього класу.

Об'єднання всіх компонентів дає метрику, яка дозволяє виявляти трансформовані версії зображення та ранжувати інші зображення за схожістю. Кожен компонент у формулі (4) має діапазон від 0 до 1, що забезпечує уніфіковане порівняння; вплив будь-якого компонента можна регулювати шляхом додавання або множення на константу за потреби.

Остаточна формула виглядає так:

$$diff = \left(\Delta p + \frac{\sum_{i=0}^{n} \Delta o_i}{n}\right) \qquad (4)$$

Було розроблено формулу для швидкого обчислення схожості між зображеннями за допомогою дескрипторів із моделі комп'ютерного зору. Вона використовує загальні параметри зображення та властивості груп об'єктів – пропорції зображення, центроїди груп об'єктів, загальну площу об'єктів і кількість об'єктів – замість деталей окремих об'єктів. Це дозволяє виконувати пошук у репозиторіях із сотнями тисяч зображень менш ніж за 3 секунди.

Час порівняння за цією методикою лінійно залежить від кількості класів об'єктів, навчених нейронною мережею, і не залежить від фактичної кількості виявлених об'єктів. Для більш високої точності можна застосувати вторинне порівняння за розмірами та позиціями окремих об'єктів для підмножини зображень (наприклад, топ-20 або топ-100), хоча це виходить за межі поточного дослідження і є темою для майбутніх робіт.

Метод розроблений для великих репозиторіїв, де зображення можуть зазнавати поширених трансформацій: різних форматів файлів, змін яскравості або контрасту, різкості, квантування, стиснення або масштабування. Більш складні зміни, такі як обертання, віддзеркалення, додавання/видалення об'єктів або вставка зображень, на цьому етапі не враховуються.

Метрика якості пошуку

Якість пошуку q визначається порядком зображеньу результатах, відсортованих за збільшенням відстані від оригіналу. Нехай у репозиторії є $n1$ схожих зображень (включаючи оригінал та його трансформації), а max – позиція останнього зображення цієї групи у відсортованому списку. Далі, нехай $n2$ – кількість зображень, що не належать до цієї групи, але з'являються перед зображенням на позиції max. Наявність таких зображень свідчить про те, що схожі зображення не є суміжними та включають нетипові результати, що відображає нижчу якість пошуку. Якість пошуку оцінюється на основі цих значень.

$$q = 1 - \frac{n2}{n1 + n2} \qquad (5)$$

Коли $n2 = 0$, першi$n1$зображеньу результатах належать до бажаної групи та йдуть підряд без пропусків. Така ситуація вважається ідеальною $q = 1$.

Опис даних

Для експериментів використовується датасет COCO 2017, який містить понад 163 000 зображень, анотованих 80 класами об'єктів. Його великий обсяг дозволяє оцінити швидкість пошуку в репозиторіях із сотнями тисяч зображень.

Датасет переважно містить повсякденні об'єкти, людей, тварин, міські пейзажі, спортивні події, пристрої, предмети та аксесуари. Окрім оригінальних зображень, додаються трансформовані копії для тестування стійкості пошуку. Трансформації включають: зменшення яскравості на 20 %, стиснення до 70 %, збільшення контрасту, збільшення зображення на 20 %, зменшення зображення на 20 %, конверсію у відтінки сірого, квантування до 128 кольорів, збільшення насиченості на 30 %, підвищення різкості зображення. Включаючи оригінальні зображення, ці варіації очікується бачити серед верхніх результатів пошуку для кожного тестового зображення.

Існують інші види трансформацій такі, як ротації, віддзеркалення, тощо. Вони не розглядаються у поточній роботі, хоча представлена модель може бути адаптована і для

роботи з ними. У випадку складних трансформацій зображення існує вірогідність, що деякі або всі об'єкти з оригіналу не будуть знайдені і трансформоване зображення буде втрачене при пошуку. Проте для будь-якого методу пошуку можна підібрати набір трансформацій, що залишить таке зображення поза результатами пошуку, в такому випадку може постати питання чи залишають трансформації зображення схожим на оригінал з людської точки зору. У цій роботі фокус саме на типових трансформаціях зображень, що використовуються для оптимізації зберігання у великих сховищах даних, стиснення, квантування, зміна масштабу, тощо. Саме вони дають можливість оцінити ефективність використання запропонованих моделей і методів у великих сховищах зображень.

Підготовка експериментів

Експерименти оцінюють швидкість виконання системи, споживання ресурсів та якість пошуку зображень. Основна увага приділяється наявності трансформованих зображень серед верхніх результатів пошуку, які повертає система.

Тестування пошуку зображень включає виконання пошуку у репозиторії для десяти трансформованих зображень, вимірювання часу виконання, якості пошуку та середньої точності на 20. Фіксується кількість оригінальних і трансформованих зображень, що йдуть підряд на початку результатів, що надає практичне уявлення про ефективність розроблених дескрипторів, метрик та моделей пошуку.

Розроблені дескриптори, метрики та моделі пошуку порівнювалися з Perceptual Hashes та дескрипторами на основі глибинного навчання, згенерованими сучасною моделлю Vision Transformer, представленою наприкінці 2020 року Алексєєм Досовіцьким, Лукасом Бейєром, Александром Колесніковим та іншими [10]. Векторні представлення всіх зображень були попередньо обчислені та збережені в базі даних. Під час пошуку вектор цільового зображення обчислюється та порівнюється із збереженими векторами за допомогою косинусної подібності. Це дозволяє здійснити пряме порівняння запропонованого методу з сучасними моделями, зосереджуючись на якості пошуку та часу виконання.

Пошук зображень за дескрипторами

У цьому експерименті тестові зображення разом із їхніми трансформаціями шукаються з метою ідеального повернення топ-10 результатів пошуку. Пошук передбачає виявлення об'єктів на зображенні та побудову дескрипторів; після цього зображення без спільних об'єктів відфільтровуються, а решта порівнюється за допомогою розробленого алгоритму.

Декілька скріншотів результатів пошуку наведено на рис. 1, 2. Усі скріншоти результатів пошуку, отриманих із використанням дескрипторів та перцептивних хешів, а також скріншоти розробленого програмного забезпечення, доступні у нашому репозиторії: https://github.com/alex-prokopenko-nure/image-search-results.

Таблиця 4 узагальнює результати експерименту. Як показано, для всіх 10 тестових зображень оригінал був отриманий першим, а всі трансформації з'явилися у топ-15 результатів. Сім зображень досягли ідеальної якості пошуку, із усіма трансформаціями в топ-10. Найнижча якість була у зображення №6-76,92 %, через появу трьох несумісних зображень на позиціях 9, 11 і 12. Середня якість пошуку по всіх тестах склала 95,12 %. mAP@20 = 0,9941.

Рисунок1. Результати пошуку зображень для тестового зображення№3

Рис. 2. Результати пошуку зображень для тестового зображення №6

Таблиця 4

Результат пошуку зображень за дескрипторами

	Час пошуку	Максимальний індекс трансформованого зображення, max	Якість пошуку, q	AP@20
1	144 мс	10	1	1
2	122 мс	12	0.8333	0.9833
3	123 мс	10	1	1
4	130 мс	10	1	1
5	154 мс	10	1	1
6	132 мс	13	0.7692	0.9669
7	132 мс	10	1	1
8	131 мс	11	0.9091	0.9909
9	155 мс	10	1	1
10	142 мс	10	1	1

Пошук зображень за перцептивнимихешами

У цьому експерименті тестові зображення та їхні трансформації були знайдені за

допомогою перцептивних хешів, порівнюючи зображення за сумою квадратів різниць між хешами. Результати цього пошуку на основі хешів узагальнено в Таблиці 5.

Таблиця 5

Результат пошуку зображень за перцептивними хешами

	Час пошуку	Максимальний індекс трансформованого зображення, max	Якість пошуку, q	AP@20
1	214 мс	>1000	<0.01	0.8
2	221 мс	>1000	<0.01	0.7
3	230 мс	>1000	<0.01	0.7888
4	242 мс	>1000	<0.01	0.7
5	215 мс	>1000	<0.01	0.7177
6	226 мс	>1000	<0.01	0.6
7	223 мс	>1000	<0.01	0.7
8	252 мс	>1000	<0.01	0.7
9	235 мс	>1000	<0.01	0.8562
10	228 мс	>1000	<0.01	0.7

Пошук на основі хешів виконується за приблизно той самий час, але у топ-результатах з'являються лише 6 – 8 трансформацій. Чорно-білі версії всіх тестових зображень розташовані поза топ-1000, а для 8 із 10 зображень 1 – 2 трансформації, що зазнали змін гамми кольору, також потрапляють поза топ-1000. mAP@20 = 0,7463.

Пошук зображень з використанням моделіVisionTransformer

У цьому експерименті тестові зображення та їхні попередньо згенеровані трансформації були знайдені за допомогою ознак Vision Transformer, при цьому схожість зображень оцінювалась через косинусну схожість між векторами ознак. Результати пошуку з використанням Vision Transformer наведено в Таблиці 6.

Таблиця 6

Результат пошуку зображень з використанням моделі VisionTransformer

	Час пошуку	Максимальний індекс трансформованого зображення, max	Якість пошуку, q	AP@20
1	69.7 с	10	1	1
2	62.6с	>20	<0.5	0.9
3	63.2с	18	0.5556	0.9555
4	66.4 с	>20	<0.5	0.9
5	58.1 с	15	0.6667	0.975
6	68с	12	0.8333	0.9651
7	61.7с	10	1	1
8	64.9с	11	0.9091	0.9909
9	62.7с	10	1	1
10	62.8с	>20	<0.5	0.9

Пошук із використанням векторів ознак Vision Transformer значно повільніший за інші методи. Подібно до перцептивних хешів, чорно-білі трансформації часто потрапляють поза топ-результати, тоді як інші трансформації зазвичай знаходяться серед перших

десяти. Якщо не враховувати чорно-білі версії, деякі зображення демонструють вищу якість пошуку, ніж запропонований метод, але за рахунок значно більшого часу виконання. Цей підхід можна застосовувати для більш точного порівняння на попередньо відфільтрованому підмножині зображень, спочатку обмеженій методом на основі запропонованих дескрипторів. mAP@20 = 0,9586.

Обговорення

Отримані результати свідчать, що розроблений алгоритм успішно досягає поставлених цілей, а саме – пошуку схожих зображень у великих базах даних. Слід відзначити, що розроблені методи та алгоритми дозволяють порівнювати зображення значно швидше, ніж очікувалося, завдяки використанню бази даних дескрипторів. Простота обчислень, закладених в алгоритмі, забезпечує швидке обчислення відстаней між зображеннями.

Середній час пошуку становить близько 150 мілісекунд для понад 164 000 зображень, що демонструє придатність алгоритму для масштабних репозиторіїв і дозволяє отримувати майже миттєві результати навіть для мільйонів зображень.

Розроблений алгоритм стабільно повертає всі 10 трансформаційних зображень серед перших 15 результатів для всіх тестових зображень. Для 7 із 10 зображень усі трансформації знаходяться у топ-10 без появи нерелевантних зображень. У випадках, коли не всі трансформації потрапляють у топ-10, вони залишаються серед 15 найближчих збігів, що дозволяє проводити подальше порівняння та уточнення за допомогою інших методів. Середнє середнє значення точності на 20 (mAP@20) для розробленого дескриптора вищий, ніж у перцептивних хешів та дескрипторів на основі глибинного навчання.

Важливим напрямом досліджень є поєднання запропонованого методу з іншими техніками пошуку. Його висока швидкість дозволяє застосовувати більш точні порівняння, наприклад, на основі ключових точок або методів глибинного навчання – до верхньої частини рейтингу (наприклад, топ-100 зображень). У цій відфільтрованій групі можна додатково аналізувати розташування об'єктів або додаткові дескриптори для уточнення та підвищення точності пошуку.

Ефективність розробленого дескриптора зображень та методу порівняння повністю залежить від якості результатів нейронної мережі. Прикладом неточності системи у такому випадку може слугувати зображення №2 з тестової вибірки, а також його зменшена до масштабу 0,8 від оригіналу **версія, що наведені на рис. 3 і 4.**

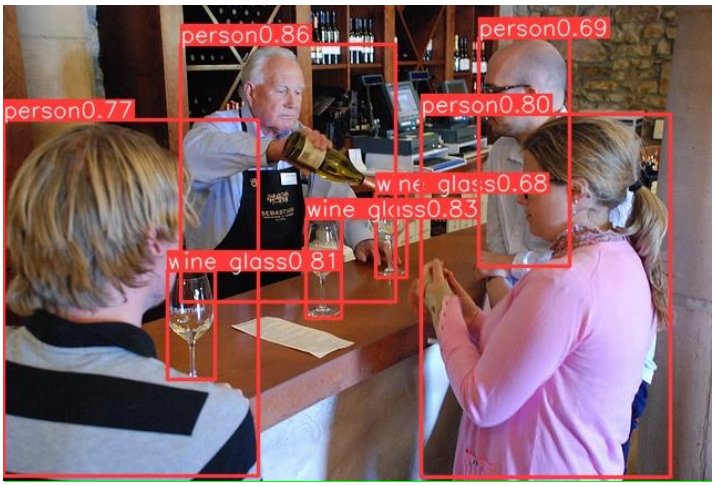

Рис. 3. Тестове зображення №2

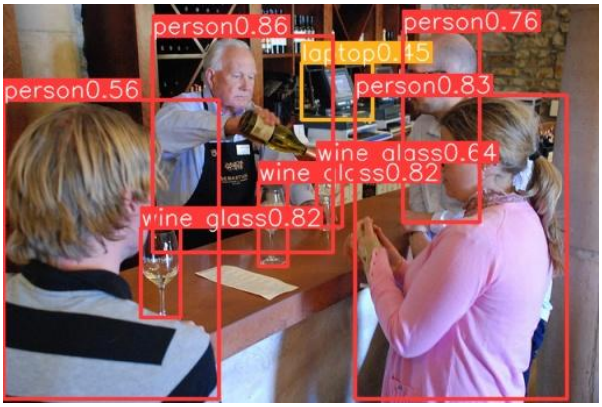

Рис. 4. Зменшене тестове зображення №2

Для зображення №2 на зменшеній трансформації модель помилково ідентифікує ноутбук як знайдений об'єкт, якого не було в оригінальному наборі об'єктів зображення, тому його відстань трохи більша, ніж у інших перетворених зображень, і воно не потрапляє до першої десятки, а опиняється позаду кількох зображень, що не є трансформацією оригіналу.

Інший приклад, продемонстрований на рис. 5 – 6, – це тестове зображення №4 слонів з двома людьми на задньому плані та його чорно-біла трансформація. Оскільки людські об'єкти мають невеликий розмір і розташовані в затіненому кутку знімка, на чорно-білому зображенні нейронна мережа не виявляє ці об'єкти (достовірність нижче порогового значення).

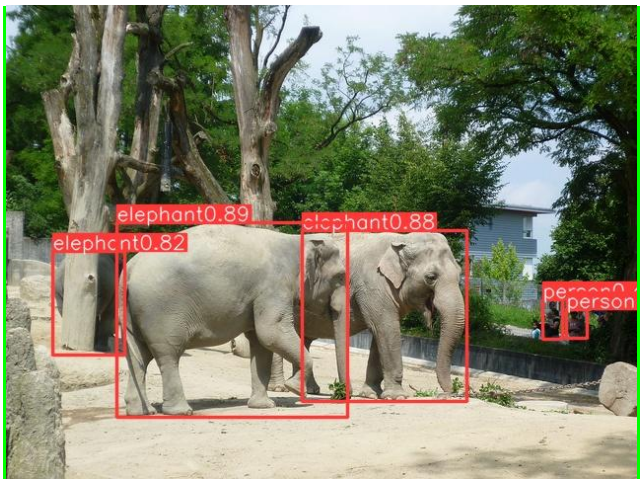

Рис. 5. Тестове зображення №4

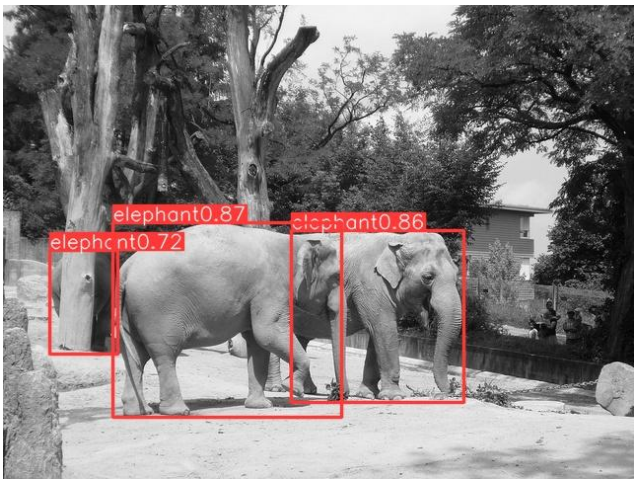

Рис. 6. Чорно-біле тестове зображення №4

Якщо об'єкти на зображенні не будуть виявлені, таке зображення не можна буде знайти. Тому критично важливим є правильний вибір моделі детекції та, за потреби, її донавчання за допомогою transfer learning з урахуванням характеристик конкретного репозиторію зображень.

Природним напрямом покращення системи є використання гібридних моделей, які поєднують кілька дескрипторів, підвищуючи якість пошуку та зменшуючи обмеження окремих моделей.

На завершення, розроблений дескриптор зображень та метод порівняння ефективно знаходять схожі зображення на високій швидкості, зосереджуючись на об'єктах, що зображені, а не на формальних параметрах зображення. Окрім функціонування як самостійного методу пошуку, він може служити попереднім фільтром, зменшуючи простір пошуку для застосування більш складних та обчислювально затратних методів порівняння.

Висновки

Ця робота представляє нову систему CBIR, яка поєднує взаємопов'язані моделі дескрипторів зображень та об'єктів. Кожен дескриптор об'єкта кодує його клас та просторове розташування, формуючи надійну та гнучку структуру для представлення як зображень, так і об'єктів, що на них зображені.

Кожне зображення репозиторію представлено дескриптором зображення, який зберігається в базі даних та завантажується у пам'ять для швидкого доступу. Метрика схожості оцінює зображення на основі зображених об'єктів: нуль означає ідентичні зображення, а більші значення відображають більші відмінності.

Було проведено огляд наявних моделей детекції об'єктів і обрано оптимальну для побудови ефективної CBIR-системи для великих репозиторіїв, що підтримує пошук та фільтрацію за вмістом зображень. Запропонований метод порівняння стійкий до поширених трансформацій, таких як зміни яскравості, кольору та стиснення, проте його продуктивність залежить від моделі детекції: якщо об'єкти не виявлені, це може збільшити відстань за метрикою та знизити позицію зображення у рейтингу.

Для перевірки дескрипторів, метрики схожості та CBIR-системи було проведено експерименти на наборі даних із понад 160 000 зображень. Дескриптори були згенеровані для всіх зображень, а продуктивність пошуку порівнювалась із моделлю Vision Transformer та алгоритмом на основі перцептивних хешів.

Експерименти показали, що розроблена CBIR-система працює ефективно, перевершуючи за часом виконання, використанням пам'яті, точністю пошуку та mAP@20, що робить її придатною для репозиторіїв із мільйонами зображень. Метрика схожості надійно ранжувала майже всі трансформовані зображення у верхній частині результатів, при цьому всі 10 трансформацій потрапляли в топ-15. Порівняно з моделлю Vision Transformer система досягла вищої якості пошуку та значно більшої швидкості виконання, а також перевершила пошук на основі перцептивних хешів за точністю за аналогічної швидкості.

Експерименти також визначили напрямки майбутніх досліджень, такі як застосування порівнянь окремих об'єктів для покращення ранжування, використання паралельних або розподілених обчислень для більших наборів даних та удосконалення модулів системи для задоволення специфічних потреб репозиторію або користувача.

У цілому, розроблена CBIR-система робить крок у побудові гібридних архітектур управління великими масивами зображень, забезпечуючи високу якість пошуку, стійкість до типових трансформацій та швидке виконання пошуку у великих сховищах даних.

СПИСОК ЛІТЕРАТУРИ

1. Rafael C. Gonzalez, Richard E. Woods Digital Image Processing, 4th. ed., Pearson/Prentice Hall, 2018. 1168 p. DOI/ISBN:9780133356724.

2. Holistic Descriptors of Omnidirectional Color Images and Their Performance in Estimation of Position and Orientation / F. Amorós et al. IEEE Access. 2020. Vol. 8. P. 81822 – 81848. DOI:

10.1109/ACCESS.2020.2990996.

3. Prior-Based Quantization Bin Matching for Cloud Storage of JPEG Images / X. Liu et al. IEEE Transactions on Image Processing. July 2018. Vol. 27, №7. P. 3222 – 3235. DOI: 10.1109/TIP.2018.2799704.

4. A NewApproach to Descriptors Generation for Image Retrievalby Analyzing Activations of Deep Neural Network Layers / P. Staszewski et al. IEEE Transactionson Neural Networks and Learning Systems. Dec. 2022. Vol. 33, №12. P. 7913–7920. DOI: 10.1109/TNNLS.2021.3084633.

5. Learning Enriched Feature Descriptor for Image Matching and Visual Measurement / Y. Rao et al. IEEE Transactions on Instrumentation and Measurement. 2023. Vol. 72. P. 1 – 12. Artno. 5008512. DOI: 10.1109/TIM.2023.3249237.

6. Gajjar V., Khandhediya Y., Gurnani A. Human Detection and Tracking for Video Surveillance: A Cognitive Science Approach, 2017. IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy, 2017. P. 2805–2809. DOI: 10.1109/ICCVW.2017.330.

7. In Search of Big Medical Data Integration Solutions - A Comprehensive Survey / H. Dhayne et al. IEEE Access. 2019. Vol. 7. P. 91265–91290. DOI: 10.1109/ACCESS.2019.2927491.

8. Truong X. -T., Yoong V. N., Ngo T. -D. RGB-D and laser data fusion-based human detection and tracking for socially aware robot navigation framework. 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO). Zhuhai, China. 2015. P. 608–613. DOI: 10.1109/ROBIO.2015.7418835.

9. Object Detection Using Convolutional Neural Networks / R. L. Galvez et al. TENCON 2018 – 2018 IEEE Region 10 Conference. Jeju, Korea (South), 2018. P. 2023–2027. DOI: 10.1109/TENCON.2018.8650517.

10. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale / A. Dosovitskiy et al.

11. International Conference on Learning Representations. 2021. DOI: 10.48550/arXiv.2010.11929.

Стаття надійшла до редакції 06.03.2026.
Стаття пройшла рецензування 18.03.2026.
Стаття опублікована 31.03.2026.

Прокопенко Олександр Сергійович– аспірант кафедри Програмної Інженерії, e-mail: oleksandr.prokopenko1@nure.ua, ORCID: 0000-0003-0489-6820.

Смеляков Сергій Вячеславович – доктор фізико-математичних наук, професор кафедри Програмної Інженерії, e-mail: serhii.smeliakov@nure.ua, ORCID: 0000-0002-5791-2479, Scopus Author ID: 24527617600.
Харківський національний університет радіоелектроніки.