ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://nv.nltu.edu.ua
https://doi.org/10.36930/40310315
Article received 06.04.2021 p.
Article accepted 29.04.2021 p.
UDC 004.[4+8]

ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

@ Correspondence author
O. S. Melnyk
oleh.melnyk.mnpz.2019@lpnu.ua

О. С. Мельник, Р. П. Базилевич
Національний університет "Львівська політехніка", м. Львів, Україна

СИСТЕМА ІДЕНТИФІКАЦІЇ ОРИГІНАЛУ ВІДЕО ЗА ЙОГО ФРАГМЕНТОМ З ВИКОРИСТАННЯМ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

Розглянуто основні сучасні та популярні підходи до вирішення задач розпізнавання ознак зображень і відео. Встановлено переваги та недоліки актуальних методів оброблення візуальної інформації, а також сучасні невирішені проблеми, пов'язані із цим сегментом робіт. Спираючись на сучасний стан досліджень з цієї предметної області, запропоновано нову систему, призначення якої "навчитись" ідентифікувати відео за його фрагментом, враховуючи характеристики зображеного у відеоряді. Першим етапом аналізу відео є його розбиття на окремі кадри, враховуючи зміну ентропії, колірної схеми та структурні відмінності сцени. Спираючись на сучасні методи, реалізовано алгоритм перетворення відео в набір кадрів. Виявлено, що компактне представлення відео у вигляді набору ключових кадрів дає змогу виділити основні контекстні характеристики. Враховуючи сучасні методи визначення характеристик зображень та ефективність машинного навчання, вирішено застосувати згорткові нейронні мережі для визначення векторних представлень. Під час вибору коректної архітектури та моделі нейронної мережі здійснено порівняльний аналіз ефективності їх роботи з використанням бази ImageNet. В наступних етапах, роботу із відео буде представлено у вигляді маніпуляції із векторами характеристик кожного кадру. Запропоновано спосіб пошуку збігу фрагментів, враховуючи оцінку кута між векторами представлень кадрів. Для покращення оптимізації пошуку розглянуто способи застосування методів індексації векторного простору кадрів. Варто застосувати цей підхід оптимізації, щоб уникнути різкої деградації ефективності пошуку із збільшенням бази. Унаслідок виконаної роботи реалізовано програмну систему у вигляді вебаплікації, яка демонструє пошук відео за його фрагментом. Проте це тільки прототип для візуалізації процесу. Під час проведення експериментів оцінено вплив та залежність довжини відео, його роздільної здатності та обсягу тестової бази від ефективності процесу пошуку. Передусім ця робота є актуальною через цінність досліджень в напрямку розвитку методів оброблення та аналізу відеоконтенту. Виявлено, що ця система має подальший розвиток та право на існування, якщо врахувати майбутні оптимізації пошуку та покращення вилучення дескрипторів.

Ключові слова: глибинне навчання; згорткова нейронна мережа; ключові кадри; вектор ознак; дескриптор зображення; коефіцієнт подібності.

Вступ

У час прогресивних технологій відео є одним з ефективних способів передачі інформації. Доступна величезна кількість різноманітних онлайн-стрімінгів та відеосервісів, які дають змогу мільйонам глядачів щодня дивитися контент на сотні терабайт даних. Це робить відео одним з основних джерел інформації у світі.

Аналіз відео та зображень – перспективна галузь штучного інтелекту і машинного навчання, яка стрімко розвивається. Проте сегмент аналізу відеоінформації містить більш складні та комплексні процеси та відрізняється з погляду побудови моделі та архітектури системи оброблення. Тому ця робота насамперед є актуальною через цінність досліджень у напрямку розвитку методів оброблення та аналізу відеоконтенту.

Переважно аналіз візуальної інформації, зокрема відеопотоки, розглядають як складову частину знаходження відповідності між двома зображеннями деякої сцени чи об'єкта. Важливим є дослідження сучасних підходів та методів аналізу відеопотоку та реалізація системи для пошуку оригіналу відео за допомогою аналізу його фрагменту, а також оцінити можливі проблеми та складності у створенні цієї системи [16].

Серед задач аналізу відео можна виділити класифікацію (визначення, до якого класу або категорії належить відео за його змістом), тегування (пошук набору тегів, що описують зображення), виявлення послідовностей (пошук характерних послідовностей кадрів на відео, які характеризують певний процес, знаходження певних дій), виявлення аномалій (пошук відео в наборі, які містять ознаки, які відрізняють їх від більшості інших подібних відео послідовностей), виокремлення ок-

Інформація про авторів:
Мельник Олег Стефанович, магістрант, кафедра програмного забезпечення. **Email:** oleh.melnyk.mnpz.2019@lpnu.ua; https://orcid.org/0000-0003-4587-6543
Базилевич Роман Петрович, д-р техн. наук, професор, кафедра програмного забезпечення. **Email:** roman.p.bazylevych@lpnu.ua; https://orcid.org/0000-0002-7949-1353
Цитування за ДСТУ: Мельник О. С., Базилевич Р. П. Система ідентифікації оригіналу відео за його фрагментом з використанням згорткових нейронних мереж. Науковий вісник НЛТУ України. 2021, т. 31, № 3. С. 94–100.
Citation APA: Melnyk, O. S., & Bazylevych, R. P. (2021). An original video fragment identification system using machine learning methods. *Scientific Bulletin of UNFU, 31*(3), 94–100. https://doi.org/10.36930/40310315

ремих об'єктів у кадрі [2]. Методи машинного навчання успішно застосовують у багатьох галузях.

Об'єкт дослідження – процес аналізу відео та зображення.

Предмет дослідження – алгоритми аналізу відео та зображень, а саме згорткові нейронні мережі, порівняння дескрипторів зображень та індексація векторного простору.

Мета роботи – розробити архітектуру та прототип системи для ідентифікації оригіналу відео за його довільним фрагментом, використовуючи методи машинного навчання та аналіз векторних дескрипторів. Також важливим є опис можливих покращень та оптимізацій складових модулів системи.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

1) дослідити поточний стан предметної області з аналізу відео та зображень;
2) вибрати оптимальний алгоритм для вилучення ключових кадрів відео;
3) експериментально визначити найефективнішу модель згорткової нейронної мережі для визначення характеристик кадрів;
4) спроектувати архітектуру та прототип веб-аплікації із застосуванням пошуку відео за ознаками ключових кадрів.

Наукова новизна отриманих результатів дослідження – вперше спроектовано та розроблено концепцію так званого "пошукового рушія" у вигляді нової системи, яка оперує відеоданими. Визначено ефективність застосування згорткових нейронних мереж, які широко застосовуються для створення інтелектуальних систем пошуку зображень, в абстракції відеопотоку кадрів. На відміну від більшості систем пошуку рекомендацій та відповідностей, де метою є виявлення низки найбільш схожих об'єктів, це дослідження сфокусоване саме на ідентифікації або відсутності фрагменту. Новим підходом є те, що вхідними даними для пошуку є не зображення або набір ключових слів, а безпосередньо фрагмент. Як результат, робота слугує поштовхом для глибшого дослідження способів аналізу відео у розрізі пошукових медіасистем.

Практична значущість результатів дослідження – розроблювана система цілком реально може використовуватись у навчальних та рекламних цілях для кращої взаємодії із різноманітним набором медіаконтенту. Тобто основне застосування полягає у використанні пошукових запитів нового рівня. Часто буває ситуація коли людина побачила тільки фрагмент деякого відео, але не має розуміння, як саме шукати оригінал за допомогою опису чи ключових слів.

Аналіз останніх досліджень та публікацій. Незважаючи на те що відео, окрім інформації про зображення, складаються з однієї додаткової розмірної інформації, тобто динамічної сцени для моделювання композиції сцени, пошук відео розглядають як складну проблему. Окрім відсутності ефективних інструментів для представлення та моделювання просторово-часової інформації, пошук відео стикається з тими ж труднощами, що й отримання зображень. Труднощі полягають у тому, що використання низькорівневих функцій для пошуку не відповідає людському сприйняттю в загальній галузі. Кластеризація завжди є рішенням для скорочення та організації вмісту відео, окрім цього, забезпечує ефективну схему індексації для пошуку відео, оскільки подібні кадри згруповані в одному кластері. Верхній рівень кластеризації згрупований за властивостями колірної схеми, тоді як нижній рівень – за функціями руху [5].

У роботі [13] запропоновано підхід до визначення збігу між візуальними сутностями (у зображеннях чи відео) базуючись на зіставленні локальних характеристик, які відображаються у вигляді компактних дескрипторів. Ці дескриптори вимірюються щільно по всьому зображенню/відео, у різних масштабах, одночасно враховуючи локальні та глобальні геометричні спотворення. Такий підхід дає можливості пошуку у складних медіаданих, де входить виявлення об'єктів у реальних "засмічених" зображеннях, використовуючи тільки грубі ескізи і межі, та виявляючи складні дії у відеоданих без попереднього вивчення та аналізу. Окрім цього, запропонований метод може зіставляти як нерухомі, так і рухомі об'єкти на відеокадрах.

Із зростанням популярності різних глибинних нейронних мереж в області штучного інтелекту, увага досліджень щодо виявлення або пошуку зображень (застосовано і для кадрів відеоряду) на підставі вмісту була перенесена з локальних функцій, таких як масштабно-інваріантне перетворення ознак (SIFT) до характеристик, отриманих від згорткових нейронних мереж (CNN). Однак CNN характеристики, безпосередньо витягнуті з цілих зображень, не підходять для виявлення невеликих повторюваних областей, тоді як особливості на підставі регіону демонструють стійкість до різноманітних модифікацій зображення, таких як масштабування та додавання шуму. Це впливає на ефективність часткового виявлення дублікатів зображення [18].

Доволі популярною в застосуванні є 3D-модель CNN для розпізнавання дій на відео. Ця модель витягує характеристики як із просторових, так і з часових вимірів, виконуючи тривимірні згортки, цим самим фіксуючи інформацію про рух, закодовану в кількох сусідніх кадрах. Розроблена модель генерує безліч каналів інформації із вхідних кадрів, а остаточне представлення ознак отримується шляхом об'єднання інформації з усіх каналів [15].

Часто трапляються дослідження, які стосуються компактного представлення відеопослідовності, так званої абстракції відео, яке може бути корисним для різних застосувань. Наприклад, забезпечення швидкого огляду вмісту на підставі відеоданих та доступу до епізодів та цілих програм у системах перегляду та пошуку відео. Послідовність попереднього перегляду зроблена для зменшення тривалості відео до короткої послідовності, яка часто використовується для того, щоб допомогти користувачеві визначити, чи варто відеопрограму переглядати повністю чи ні. Це створює враження про цілий відеовміст, бо містить тільки найцікавіші відеокадри [4].

Результати дослідження та їх обговорення

Компоненти запропонованої системи. Розробленя сучасної моделі глибокого навчання не має реальної цінності, якщо її не можна застосувати в реальному додатку. У випадку моделей глибокого навчання, переважна більшість із них фактично розгортається як веб або мобільний додаток. Остаточна мета – мати повністю функціональний сервіс, з яким можуть взаємодіяти в режимі реального часу.

Система реалізована у вигляді клієнт-серверної архітектури. Серверна частина містить окремі компоненти, які активно взаємодіють між собою:

- API модуль для надання зовнішнього інтерфейсу взаємодії із системою;
- модуль оброблення відео – розбиття на послідовність ключових кадрів;
- модуль визначення векторних дескрипторів – кожен кадр перетворюється в числовий вектор основних характеристик;
- модуль пошуку – зіставлення та знаходження збігів між вихідними характеристиками відео та опрацьованими оригіналами;
- модуль взаємодії із базою даних – читання, запис та оновлення документів у базі;
- модуль розрахунків – математичні операції над векторами та їх перетворення.

Вхідними даними для системи є невеликі фрагменти відео, а це можуть бути як коротке відзняте відео на камеру смартфону, так і завантажений короткий відеоролик. Передбачається, що тривалість вхідного фрагменту не перевищуватиме 15-20 секунд.

Першим кроком опрацювання відео є його декодування та розбиття на кадри – традиційний підхід до аналізу такого виду інформації. Відео високої якості є надто важким для опрацювання, а тому фрейми повинні відповідати розмірності, яка є прийнятною для подальшого етапу – визначення дескрипторів або представлень.

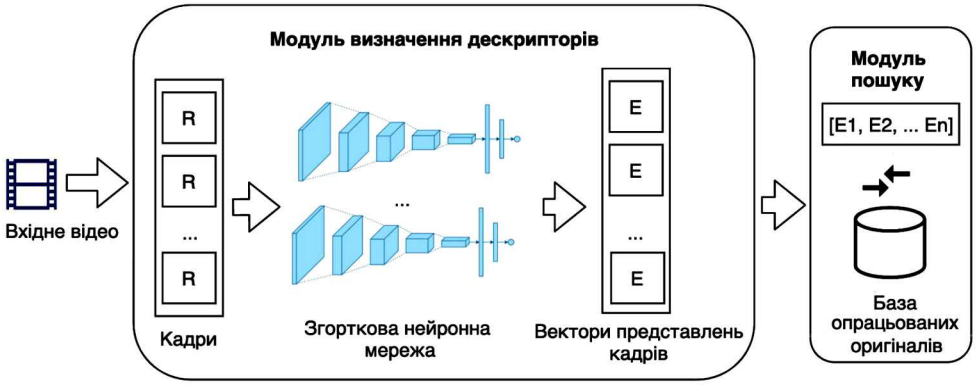

Рис. 1. Основні кроки системи ідентифікації відео

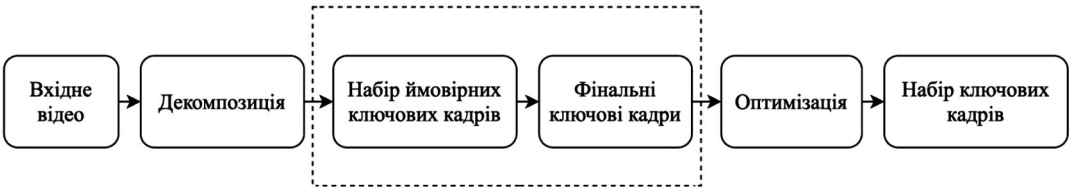

Рис. 2. Кроки вилучення ключових кадрів

Визначення ключових кадрів. Варто зазначити, що вилучення кожного кадру не є ефективним рішенням. Навпаки, необхідно фокусуватись на ключових кадрах – репрезентативних кадрах відеопотоку, що забезпечують найбільш точний і компактний підсумок відеовмісту.

Динамічне вилучення ключового кадру не повинно базуватися на тривалості поточного кадру. Чим більше змінюється поточний кадр, тим більше ключових кадрів потрібно виділити, навіть якщо поточний кадр не довгий. Навпаки, навіть при тривалому знімку, якщо картинка в основному незмінна, ми повинні витягти менше ключових кадрів [2]. Якщо поточне зображення камери повністю чорне, ми не повинні витягувати жодних ключових кадрів. Варто враховувати певні критерії під час вилучення ключових кадрів.

Витяг кадру та критерії вибору для вилучення ключових кадрів:

- кадр, який досить відрізняється від попередніх, використовуючи абсолютні відмінності в кольоровому просторі RGB;
- оцінювання яскравості вилучених кадрів;
- оцінювання ентропії/контрасту вилучених кадрів.

На рис. 2 описано основні кроки процесу вилучення ключових кадрів вхідного відео [14], а саме:

1. Здійснюється декомпозиція відеоконтенту на окремі фрейми. Причому не потрібно враховувати конкретний формат або структуру відеопотоку, тому що відео декодується перед цим.
2. Далі, вилучається ймовірна послідовність кадрів, базуючись на різниці колірної характеристики між сусідніми кадрами від вихідної послідовності.
3. Остаточна послідовність ключових кадрів отримується шляхом аналізу різниці структурних характеристик між сусідніми кадрами від ймовірної послідовності.
4. Для оптимізації методу не було застосовано сегментацію. Базуючись на кінцевому наборі кадрів, якщо їх недостатньо, тоді вибираємо відповідну кількість з оригінального відео відповідно до визначеного інтервалу.

Опираючись на зазначені вище міркування та схему (див. рис. 2), метод різниці кадрів використовується для вилучення ключових кадрів шляхом аналізу наявності просторової надмірності та тимчасової надмірності.

Модель штучної нейронної мережі. Для цього дослідження етап класифікації моделі нейронної мережі не потрібен, тому що важливо тільки отримати вектор основних характеристик кадру, що буде описувати контекст зображення. Зазвичай, для того, щоб навчити нейронну мережу необхідний великий об'єм даних. Варто зазначити, що найточніші моделі мереж є дуже вимогливими з погляду використання пам'яті та обчислювальних ресурсів, а натомість швидким та компактним моделям бракує точності. Тому для проведення експериментів було обрану наявні моделі згорткових нейронних мереж, щоб сконцентруватись на розробленні системи, а не тренуванні моделі.

Це називають передавальним навчанням, тобто використання "знань", отриманих при вирішенні однієї за-

дачі, для іншої. Проаналізовано такі популярні архітектури, як: GoogleNet, AlexNet, VGG-16, ResNet50, ResNet152, Inception_v3, EfficientNet_b7 [8]. Для тренування перерахованих вище архітектур згорткових нейронних мереж було обрано набір даних ImageNet, який налічує більше 14 млн зображень та 2048 поділених класів. Цього цілком достатньо для початкового тренування мережі, яка буде готовою розпізнати базові характеристики об'єктів.

Було проведено порівняльний аналіз [11] найпопулярніших архітектур згорткових нейронних мереж, який наведено в табл. 1.

Табл. 1. Порівняльний аналіз архітектур CNN

Архітектура CNN	К-сть шарів, шт.	К-сть параметрів, млн	Топ-1 точність, %	Топ-5 точність, %
GoogleNet	22	7	74,8	92,2
AlexNet	8	61	68,5	84,7
VGG-16	16	138	71,3	90,1
ResNet50	50	23	77,7	93,8
ResNet152	152	60	76,6	93,1
Inception_v3	48	24	77,9	93

На перший погляд, більше шарів – краще, але через проблему зникаючого градієнта ваги моделей перші шари не можуть бути коректно оновлені через зворотне розповсюдження градієнта помилок. Під час тестування та дослідження роботи різних архітектурних моделей CNN було визначено, що найкращі показники визначення характеристик зображення, а також показники ефективності виявлено для архітектури ResNet50 [2].

Визначення векторних характеристик кадру. Ознаки – це частини або візерунки об'єкта на зображенні, які допомагають його ідентифікувати. Наприклад, квадрат має 4 кути і 4 ребра, їх можна назвати ознаками квадрата, і вони допомагають нам визначити, що це квадрат. Особливості містять такі властивості, як кути, краї, області визначних місць, хребти тощо.

Традиційні екстрактори ознак можуть бути замінені згортковою нейронною мережею, оскільки вони мають сильну здатність витягувати складні функції, які виражають зображення набагато детальніше, вивчають специфічні функції завдання та є набагато ефективнішими [9].

Для зручної роботи та проведення експериментів із використанням згорткових нейронних мереж було застосовано популярну бібліотеку TensorFlow. Перед тим як працювати із зображенням, необхідно модифікувати його, відповідно до формату, з яким зручно працювати нейронній мережі.

Етапи попереднього оброблення такі:

- перетворення зображення в тензор фігури W×H×3 із типом даних цілої кількості;
- змінити розмір зображення до тензору форми 224×224×3 – розмірність, яку вимагає архітектура моделі ResNet50;
- перетворення типу даних тензора в плавучий та додавання нової осі, щоб зробити форму тензора 1×224×224×3. Це точна вхідна форма, яку очікує модель.

Для визначення характеристик кадру було використано попередньо натреновану модель згорткової нейронної мережі ResNet50 [12]. Кожен кадр після проходження через шари оброблення нейронної мережі перетворюється на вектор дійсних чисел, що є набором ключових ознак виділених та розпізнаних об'єктів зображення. Унаслідок застосування такого перетворення за допомогою штучної нейронної мережі відео репрезентується як масив векторів, які описують контекст того, що було зображено на кожному кадрі відеоряду.

Рис. 3. Схема представлення вхідного зображення

Додатково, для збільшення інформації представлення, можна використати оптичні потоки, які допомагають визначати дії та рухи на послідовних кадрах. Таке компактне представлення можна гнучко обробляти, використовувати різні моделі синтезу або застосовувати як вхідні дані якоїсь іншої моделі (рис. 3).

Порівняння кадрів. Наступний модуль системи – це модуль пошуку, який відповідає за зіставлення ключових векторів вхідного фрагменту відео із тими, що вже знаходяться в базі. Процес, що містить визначення подібності [7], використовує такі основні концепції:

- перетворення даних у вектор об'єктів;
- порівняння векторів за допомогою метрики відстані;
- класифікація відстані як подібної чи не схожої.

У базі даних зберігаються вже проаналізовані оригінали відео, тобто тільки векторне представлення їх кадрів. Усі вектори знаходяться в багатовимірному просторі, а тому можна використати функцію близькості для пошуку та зіставлення ключових кадрів.

Вибір ефективної функції подібності векторів базується на дослідженні [7], де розглянуто низку популярних методів та порівняння метрик експериментальних застосувань. Причому вхідними векторами для порівняння є результат останнього незв'язного шару згорткової нейронної мережі, що є ідентичним способом і в запропонованій системі. З-поміж методів зіставлення багатовимірних векторів, таких як *Евклідова відстань*, *косинус подібності*, *Манхеттенська метрика*, *метрика Мінковського та метрика кореляції*, найефективніший результат було отримано із використанням косинуса подібності. При цьому враховано швидкість прямого співставлення і застосування у кластеризації. Оскільки сфера застосування методів та проведені експерименти входять в область розроблюваної системи, обрано саме *метрику косинуса подібності* для реалізації пошуку збігів кадрів.

Для двох заданих векторів ознак n-вимірного простору A та B, косинус (або коефіцієнт) подібності визначають за допомогою скалярного добутку та довжини, як

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}, \qquad (1)$$

де: $\cos(\theta)$ – косинус подібності; A_i – координата i-го виміру вектора ознак A; B_i – координата i-го виміру вектора ознак B.

Встановлено мінімальний поріг схожості кадрів, що дорівнює 80, який означає, що кадри з менших коефіцієнтом схожості будуть проігноровані та відкинуті. Для пошуку збігів між фрагментом та оригіналом використовується пряме зіставлення кадрів із використанням вікна зсуву (рис. 4).

Обговорення результатів дослідження. Процес розроблення, дослідження та тестування проведено на основній робочі станції Macbook Pro late 2015 із такими характеристиками:

- CPU – Intel Core i5, 2 cores, 2,7 GHz;
- RAM – 16 ГБ 1867 MHz DDR3;
- GPU – Intel Iris Graphics 6100 1536 МБ.

На цій машині не проводились тренування нейронної мережі через обмеження в обчислювальних ресурсах. Для проведення експериментів роботи системи використано попередньо натреновану модель CNN із визначеними вагами елементів. У попередньому розділі було аргументовано вибір архітектури ResNet50, базуючись на офіційних показниках точності класифікації із використанням колекції даних ImageNet для тренування згорткових мереж. Причому на ефективність не вплинуло те, що інші моделі на виході могли мати більшу розмірність вектора.

Тестовими даними для системи слугували відносно короткі відеотрейлери до популярних фільмів – тривалістю 2-5 хвилин. Загалом було використано близько 50 відео як базу для досліджень. Вхідними даними для ідентифікації було обрано фрагменти тривалістю до 20 секунд, деякі з яких не є частиною тестової бази оригіналів. Як було зазначено вище, першим етапом аналізу відео є отримання ключових кадрів, результат процесу якого зображено на рис. 5.

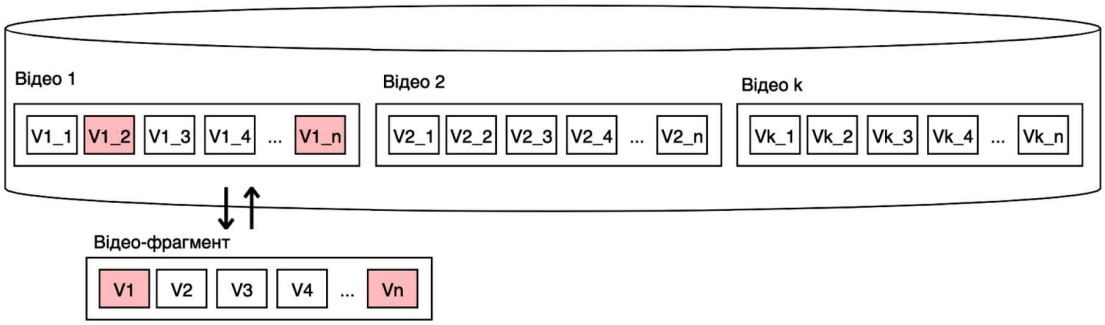

Рис. 4. Ілюстрація пошуку збігів фрагменту повним перебором

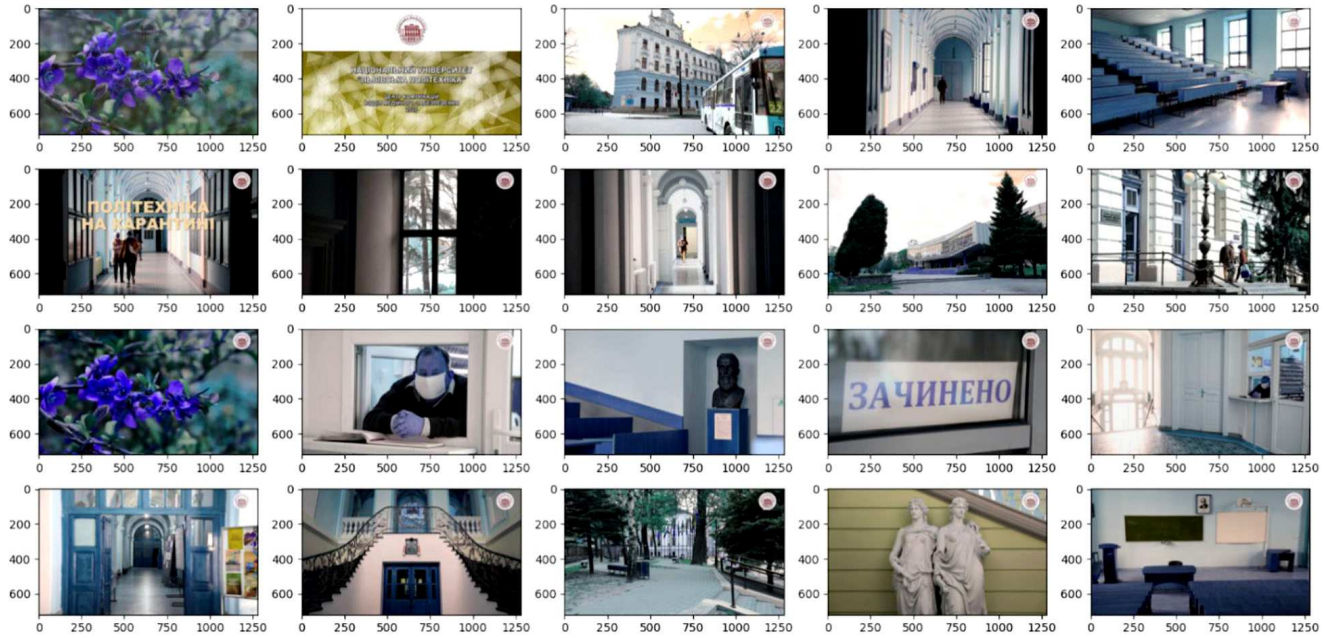

Рис. 5. Вилучені ключові кадри вхідного відеофрагменту

Описану в цій роботі систему варто порівняти із роботою [3], де проведено схоже дослідження із розроблення системи для знаходження відео, базуючись на характеристиках зображеного контенту. Так само, опираючись на поняття векторного представлення відеоряду, було реалізовано порівняння дескрипторів, але тут автори використовували такі просторові характеристики, як колірна схема та форми об'єктів. Щодо швидкості визначення характеристик відео та ідентифікації, то вони не мали істотної різниці на малих наборах даних. Для порівняння, в табл. 2 наведено приклад часових і відсоткових метрик розроблюваної системи.

Цікавими є експериментальні результати роботи [1], де описано фреймворк для визначення ймовірних дублікатів відео, що є доволі практичною задачею також. Автори поставили за мету розробити передусім швидкодіючу систему для роботи з мільйонною базою. Високої продуктивності досягнуто завдяки представленню відео у вигляді компактного мінімізованого набору характеристик та застосування кластеризації із індексацією векторного простору.

Табл. 2. Приклад часових та відсоткових метрик ідентифікації відеофрагментів різної довжини

Тривалість вхідного відео (с)	Час визначення дескриптора (с)	Частка збігу (%)	Час ідентифікції (с)
15	0,25	85,1	10
20	0,117	91,3	11
24	0,23	86,7	13
17	0,098	91,7	7
22	0,157	90,3	15
12	0,109	95,4	12
18	0,078	83,9	19
25	0,12	88,1	21

Проте зменшення розмірності дескриптору, водночас, зменшує ймовірність ідентифікації конкретного відео через те, що характеристики більш узагальнені та часто повторюються. Також варто зазначити, що пошук у базі відео здійснюється за рахунок ключових слів, а не безпосереднього відео-фрагменту.

Як довели експерименти із зміною кількості відео в базі оригіналів (рис. 6), використаний алгоритм пошуку потребує покращення. Проте система виконує поставлені завдання роботи. Реалізація ефективного пошуку та індексації векторного простору кадрів є окремою темою та не входить до завдань поточного дослідження.

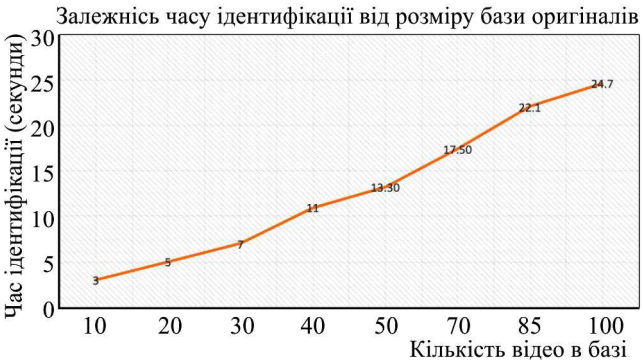

Рис. 6. Залежність швидкості ідентифікації від кількості відео в базі

На рис. 7 наведено приклад успішного результату пошуку відео за його фрагментом із можливістю перегляду ідентифікованого оригіналу, що і є кінцевою метою системи пошуку.

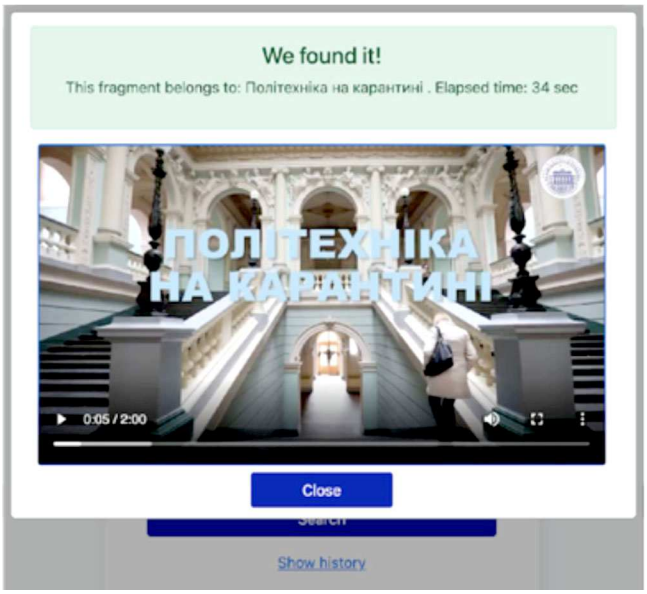

Рис. 7. Результат успішного пошуку відео за його фрагментом

Для підвищення продуктивності роботи системи потрібно застосовувати обчислювальні можливості GPU у разі роботи нейронної мережі. Графічні процесори забезпечують масивний паралелізм, коли кожне ядро орієнтоване на ефективне обчислення. Тому аплікацію варто розгортати на машині із доволі потужною графічною картою. Це можна досягти, використовуючи технологію NVIDIA CUDA, яка дає змогу запускати обчислення на GPU та паралелізувати їх.

Одним із способів покращення алгоритму пошуку збігів кадрів відео-фрагмента є застосування індексації отриманих векторних представлень кадрів. Процес пошуку векторів, близьких за схожістю до вхідного, відомий як пошук найближчих сусідів. Наївна реалізація пошуку найближчих сусідів полягає у простому обчисленні відстані між вектором запиту та кожним вектором у нашій колекції (зазвичай називають набором посилань). Однак обчислення цих відстаней методом грубої сили швидко стає нездійсненним. Клас методів, відомий як приблизний пошук найближчих сусідів [8], пропонує рішення нашої масштабної дилеми шляхом розумного розподілу векторного простору в такий спосіб, що нам потрібно тільки вивчити невелику підмножину загального набору посилань. Для реалізації може бути використано такі структури даних, як дерева, k-графи, хеш-таблиці.

Висновки

Отже, ідея пошуку відео та його аналізу є досить актуальною зараз. Існує безліч прикладів використання схожих технологій опрацювання відеопотоку даних, а тому є сенс досліджувати та розвивати цей напрям і шукати нові застосування технологій розпізнавання відеоконтенту.

Дослідження наявних рішень і наукових робіт з цієї тематики дало змогу з'ясувати сучасний стан предметної області, а також спростило пошук рішень для складових елементів системи. Встановлено, що аналіз відео передбачає розбиття відео на масив ключових кадрів, що забезпечують найбільш точний і компактний підсумок відеовмісту. Запропоновано нову систему, призначення якої ідентифікувати відео за його довільним фрагментом, враховуючи характеристики ключових кадрів відеоряду.

Реалізовано модуль представлення основних характеристик відео, а саме – набору кадрів. Для цього було застосовано згорткову нейронну мережу архітектури ResNet50. Цю модель було обрано спираючись на ефективність роботи порівняно із іншими популярними архітектурами. Такий підхід дав змогу нормалізувати відео для зручного подальшого оброблення та пошуку. Базуючись на обчисленні косинуса схожості представлень кожного із кадрів та аналізуючи найвищі коефіцієнти збігу дескрипторів, було розроблено відповідний алгоритм пошуку. У подальших дослідженнях є сенс покращити та оптимізувати цей алгоритм, оскільки такий спосіб пошуку є доволі примітивний і був застосований для демонстрації дієздатності запропонованої системи.

Наведено низку оптимізацій і покращень, які доцільно розглянути в подальших працях, щоб досягнути кращої ефективності системи. Експериментальні результати довели, що ідея такої системи є цілком реальною та практичною в застосуванні.

References

1. Cai, Y., Yang, L., Ping, W., Wang, F., Mei, T., Hua, X. S., & Li, S. (2011). Million-scale near-duplicate video retrieval system. In *Proceedings of the 19th ACM international conference on Multimedia*, 837–838. https://doi.org/10.1145/2072298.2072484
2. Gharbi, H., Bahroun, S., & Zagrouba, E. (2019). Key frame extraction for video summarization using local description and repeatability graph clustering. *Signal, Image and Video Processing, 13*(3), 507–515. https://doi.org/10.1007/s11760-018-1376-8
3. Gitte, M., Bawaskar, H., Sethi, S., & Shinde, A. (2014). Content based video retrieval system. *International Journal of Research in Engineering and Technology, 3*(06), 123–129.

4. Hanjalic, A., & Zhang, H. (1999). An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Transactions on circuits and systems for video technology*, *9*(8), 1280–1289. https://doi.org/10.1109/76.809162
5. Hanjalic, A., Lagendijk, R. L., & Biemond, J. (2001). Recent advances in video content analysis: from visual features to semantic video segments. *International Journal of Image and Graphics*, 1(01), 63–81. https://doi.org/10.1142/S0219467801000062
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90
7. Kavitha, K., & Rao, B. T. (2019). *Evaluation of distance measures for feature based image registration using alexnet*. arXiv preprint arXiv:1907.12921.
8. Kushilevitz, E., Ostrovsky, R., & Rabani, Y. (2000). Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing, 30*(2), 457–474.
9. Liu, R., Wei, S., Zhao, Y., & Yang, Y. (2018). Indexing of the CNN features for the large scale image search. *Multimedia Tools and Applications, 77*(24), 32107–32131. https://doi.org/10.1007/s11042-018-6210-3
10. Rasheed, Z., & Shah, M. (2005). Detection and representation of scenes in videos. *IEEE transactions on Multimedia, 7*(6), 1097–1105. https://doi.org/10.1109/TMM.2005.858392
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.
12. Shanmugamani, R. (2018). Deep Learning for Computer Vision: *Expert techniques to train advanced neural networks using TensorFlow and Keras*. Packt Publishing Ltd.
13. Shechtman, E., & Irani, M. (2007). Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.
14. Shi, Y., Yang, H., Gong, M., Liu, X., & Xia, Y. (2017). A fast and robust key frame extraction method for video copyright protection. *Journal of Electrical and Computer Engineering*, 20. https://doi.org/10.1155/2017/1231794
15. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3 d convolutional networks. In *Proceedings of the IEEE international conference on computer vision,* 4489–4497.
16. Tymchyshyn, R. M., Volkov, O. Ye., Hospodarchuk, O. Yu., & Bohachuk, Yu. P. (2018). Suchasni pidkhody do rozviazannia zadach kompiuternoho zoru. *Upravliaiuchi systemy ta mashyny*. [In Ukrainian].
17. Tzelepi, M., & Tefas, A. (2018). Deep convolutional learning for content based image retrieval. *Neurocomputing, 275*, 2467–2478. https://doi.org/10.1016/j.neucom.2017.11.022
18. Zhou, Z., Wu, Q. J., Wan, S., Sun, W., & Sun, X. (2020). Integrating SIFT and CNN feature matching for partial-duplicate image detection. *IEEE Transactions on Emerging Topics in Computational Intelligence, 4*(5), 593–604.

O. S. Melnyk, R. P. Bazylevych

Lviv Polytechnic National University, Lviv, Ukraine

AN ORIGINAL VIDEO FRAGMENT IDENTIFICATION SYSTEM USING MACHINE LEARNING METHODS

This paper considers the main modern and popular approaches to solving problems of image and video recognition. The advantages and disadvantages of current methods of visual information processing are identified, as well as current unresolved issues related to this segment of researches. A new system is proposed, based on the current state of research in this subject area. The purpose of the system is to learn how to identify the video by its fragment, considering the characteristics of the image in video series. The first stage of video analysis is its division into individual frames taking into account changes in entropy, colour scheme, and structural differences of the scene. An algorithm for converting video into a set of frames was implemented based on existing methods. The research has revealed that a compact representation of the video in the form of a set of keyframes allows highlighting the main contextual and temporal characteristics. Taking into consideration modern methods for determining the characteristics of images and the effectiveness of machine learning, we decided to use convolutional neural networks to extract vector representations. When choosing the correct neural network architecture and model, a comparative analysis of the effectiveness of their work was performed using the ImageNet database. In the next stages, work with video will be presented in the form of manipulation with the vector of each frame characteristics. A coincidence method for finding fragments concerning the estimate of the angle between the vectors of the frame representations is proposed. To improve search optimization, methods of applying vector frame indexing methods are considered. This optimization approach should be used to avoid a sharp degradation of search performance with increasing database. As a result of our research, a software system was implemented in the form of a web application that demonstrates search for video by its fragment. However, this is only a prototype for process visualization. During the experiments, the dependence of video length, resolution, and the size of the test base on the efficiency of the search process were evaluated. First of all, this study is relevant because of the value of research in the development of methods of processing and analysing video content. This system is revealed to have further development and the right to exist given future search optimizations and improved descriptor retrieval.

Keywords: deep learning; convolutional neural network; keyframes; feature vector; visual descriptor; similarity measure.