«Системні технології» 3 (164) 2026 «System technologies»
DOI 10.34185/1562-9945-3-164-2026-14
УДК 004.93

І.Р. Ніколаєв, Ю.О. Олійник

МЕТОД ТА ПРОГРАМНИЙ ЗАСІБ РОЗПІЗНАВАННЯ ТА АНАЛІЗУ ОБ'ЄКТІВ НА ВІДЕО

Анотація. Актуальність дослідження зумовлена стрімким розвитком технологій штучного інтелекту та комп'ютерного зору, що дозволяють автоматизувати процеси аналізу великих обсягів відеоданих у реальному часі. Однак існує вагома проблема: сучасні алгоритми генерують значні масиви «сирих» неструктурованих даних, які не мають єдиної багатовимірної структури, що суттєво ускладнює їх гнучкий аналіз та агрегування. Метою роботи є підвищення швидкості аналізу надвеликих масивів відеоданих шляхом розробки методу автоматизованої екстракції структурованих даних та їхньої інтеграції у багатовимірну OLAP-модель. Основними методами дослідження є каскадна фільтрація з використанням нейромережі YOLOv26m для швидкої детекції об'єктів, застосування мультимодальної VLM-моделі Moondream2 для семантичного збагачення опису сцен, а також побудова ROLAP-моделі у межах розподіленої мікросервісної архітектури. У результаті експериментів доведено високу точність детекції об'єктів (mAP50 = 0.814 при швидкості 81.03 FPS), збалансовану швидкодію семантичного аналізу (0.422 с/кадр) та здатність системи до лінійного прискорення при масштабуванні. Розроблений комплексний інструмент успішно трансформує великі обсяги неструктурованих відеоданих у формат, придатний для глибокого багатовимірного аналізу та ефективного прийняття рішень у високонавантажених системах відеомоніторингу.

Ключові слова: комп'ютерний зір, відеоаналітика, виявлення об'єктів, відстеження об'єктів, багатовимірна модель, OLAP, YOLO, VLM, військова техніка.

Постановка проблеми. У сучасному світі однією з найважливіших галузей використання ШІ є комп'ютерний зір, що дозволяє автоматизувати процеси аналізу зображень та відео, включаючи задачі виявлення об'єктів. Проте відео, як джерело даних, є складним для обробки, оскільки містить велику кількість інформації у часовій послідовності кадрів.

Проблема полягає у недостатній ефективності та масштабованості наявних підходів до перетворення великих неструктурованих відеоданих (результатів виявлення та відстеження об'єктів) у форму, придатну для глибокого аналітичного аналізу та прийняття рішень відеоаналітиками в реальному часі. Сучасні методи комп'ютерного зору генерують значні обсяги "сирих" даних, які складно агрегувати за часом, територією чи типом об'єкта, і вони не мають єдиної багатовимірної структури, що обмежує можливо-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

сті гнучкого виконання аналітичних операцій. Тому необхідно розробити та дослідити ефективність методу та програмного засобу, що забезпечують безперервний процес екстракції та семантичного збагачення даних з відеопотоку, збереження отриманих даних у структурі, яка підтримує складні аналітичні операції.

Аналіз останніх досліджень і публікацій. Задача виявлення, відстеження та аналізу об'єктів у відеопотоці є однією з ключових у галузі комп'ютерного зору та активно досліджується протягом останніх років.

У дослідженні[1] розглянуто класичні методи розпізнавання об'єктів на основі згорткових нейронних мереж (CNN), які дозволяють ефективно виділяти ознаки зображень. Проте такі підходи, особливо двоетапні моделі, наприклад, Faster R-CNN, мають недостатню швидкодію для задач реального часу. Це обмежує їх застосування у високонавантажених системах відеоаналізу.

Альтернативою є одноетапні методи детекції, зокрема алгоритм YOLO[2]. Його ключовою перевагою є можливість одночасного визначення координат об'єктів та їх класів за один прохід нейронної мережі, що забезпечує високу швидкість обробки. Це робить YOLO одним із найбільш популярних рішень для задач відеоаналітики в реальному часі.

У роботі[3] досліджуються сучасні підходи до відстеження об'єктів, зокрема парадигма Tracking-by-Detection, яка поєднує детекцію та подальшу асоціацію об'єктів між кадрами. Розглянуто алгоритми SORT, DeepSORT та ByteTrack. Автор відзначає, що використання додаткових ознак зовнішнього вигляду підвищує точність трекінгу, проте збільшує обчислювальні витрати. Водночас метод ByteTrack демонструє високу ефективність за рахунок оптимізованої асоціації навіть низьковпевнених детекцій.

Використання мультимодальних моделей для семантичного аналізу зображень[4] дозволяє отримувати більш глибоке розуміння сцени, включаючи опис середовища, взаємодію об'єктів та контекст подій. Однак їх використання в потокових системах обмежується високою обчислювальною складністю.

Окремим напрямом досліджень є аналіз та агрегування результатів відеообробки, зокрема із застосуванням OLAP-технологій для багатовимірного аналізу даних[5, 6]. На основі цих технологій розроблено спеціалізовані моделі представлення відеоданих, такі як SurvCube та VideoCube, які інтегрують результати відеоаналізу у багатовимірні структури. Незважаючи на ефективність запропонованих рішень, вони мають обмеження, пов'язані з недостатньою гнучкістю, відсутністю механізмів автоматичного оновлення семантичних класів, а також низьким рівнем семантичного формалізму, тобто відсутністю ієрархій.

Таким чином, аналіз існуючих досліджень показав, що сучасні методи забезпечують високу точність виявлення та відстеження об'єктів, однак проблема інтеграції результатів у зручну для аналізу форму залишається актуальною. Це обґрунтовує доцільність розробки підходу, який поєднує методи комп'ютерного зору з багатовимірними моделями даних для забезпечення ефективної аналітики відеопотоків.

Мета дослідження. Метою дослідження є підвищення швидкості аналізу надвеликих масивів відеоданих шляхом розробки методу та програмного засобу автоматизо-

ваної екстракції структурованих даних та їхньої інтеграції у багатовимірну OLAP-модель, що підтримує паралельну обробку та семантичне збагачення інформації.

Основна частина. Процес збору та обробки даних у запропонованій системі базується на принципах каскадної фільтрації та семантичного збагачення інформації. Система забезпечує паралельну обробку множинних вхідних відеопотоків, трансформуючи неструктуровані візуальні дані у структуровані записи для їх подальшої інтеграції в аналітичну OLAP-модель. На рисунку 1 наведено загальну схему розробленого методу.

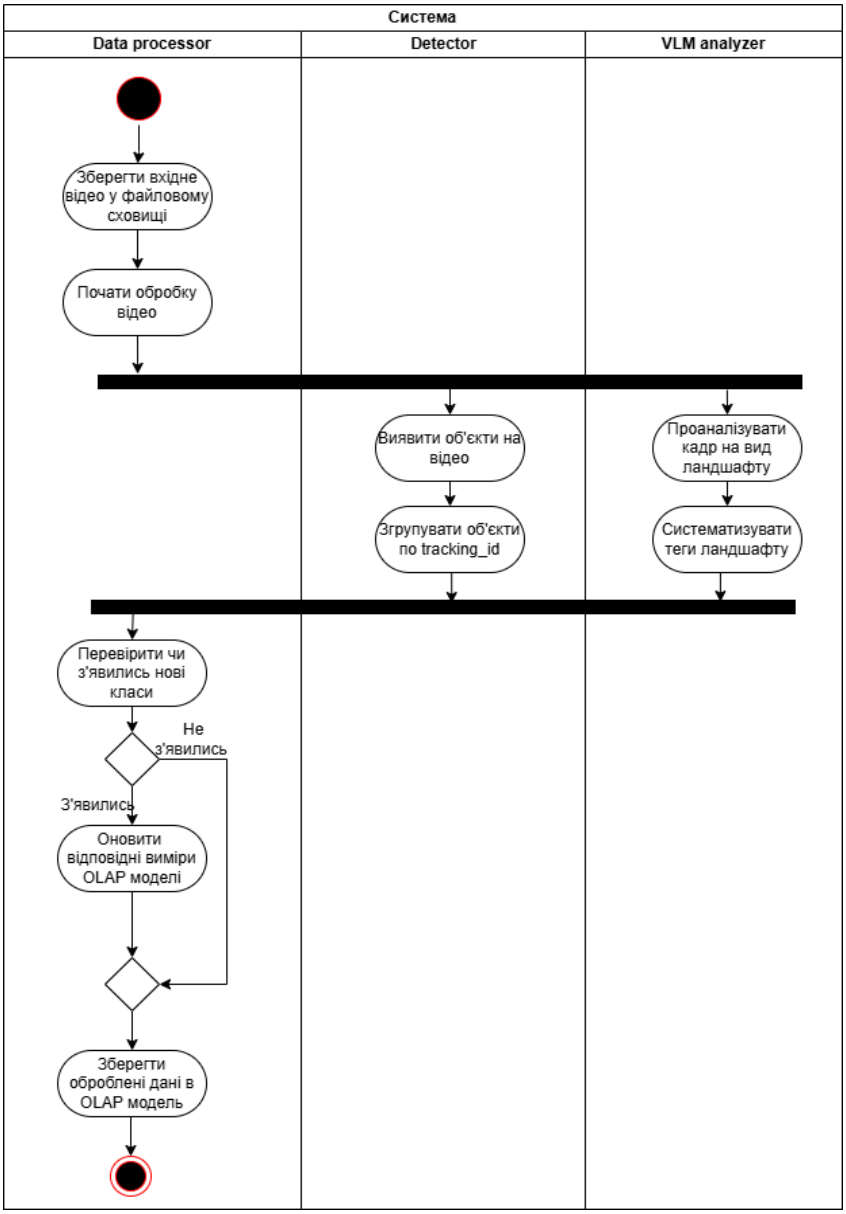

Рисунок 1 – Загальна схема розробленого методу розпізнавання та аналізу об’єктів на відео

Попередня обробка

Вхідним сигналом для системи є сукупність відеопотоків від розподілених джерел, наприклад, IP-камер та камер з дронів тощо. На цьому етапі вирішуються два ключові завдання:

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

- екстракція метаданих: з відеопотоку зчитуються часова мітка геокоординати, роздільна здатність, тривалість і частота кадрів;

- оптимізації навантаження: з метою забезпечення швидкодії в умовах високо навантаженої системи, аналіз проводиться не для кожного кадру, а використовується метод проріджування кадрів, де для детекції обирається кожен n-й кадр, наприклад, 2-5 кадрів на секунду, що дозволяє суттєво зменшити навантаження на обчислювальні потужності без втрати точності моніторингу динамічних об'єктів

Детекція, класифікація та трекінг об'єктів

На етапі виявлення об'єктів ключові кадри передаються до нейромережі YOLO, яка виконує функцію первинного фільтра та класифікатора. У разі відсутності цільових об'єктів на кадрі процес обробки завершується автоматично. При успішній детекції модель повертає параметри домену та класу об'єкта, координати обмежувальної рамки та рівень впевненості. Для підвищення достовірності результатів дані, що не досягають встановленого порогу впевненості, відсікаються.

Важливою частиною цього етапу є механізм об'єктного трекінгу. Кожному виявленому об'єкту присвоюється унікальний ідентифікатор object_id, що забезпечує унікальність даних та дозволяє відстежувати переміщення об'єкта в межах поля зору камери, мінімізуючи надлишковість обчислень та обсяг зберіжених даних.

Семантичне збагачення даних за допомогою мультимодальних моделей

Паралельно з детекцією відбувається семантичне збагачення даних за допомогою мультимодальної моделі Moondream2, яка швидко генерує наповнені відповіді. Процес аналізу поєднує візуальний вхід (кадр або область об'єкта) та текстовий запит-постановку задачі. Це забезпечує гнучкість системи, дозволяючи адаптувати моніторинг під різні сценарії (наприклад, тип локації, погодні умови, супутні об'єкти) лише зміною тексту запиту, без необхідності перенавчання нейромережі. На відміну від YOLO, модель VLM виконує складні когнітивні завдання, аналізуючи взаємозв'язки між об'єктами. Оскільки цей процес ресурсомісткий, його застосовують лише до найбільш репрезентативних кадрів. Отримана структурована відповідь дозволяє автоматично виділяти нові типи середовищ та подій для додавання до багатовимірної моделі.

Для отримання структурованої відповіді, емпірично було підібрано наступний промпт до VLM моделі: "Describe landscape, structures and people". З отриманої відповіді можна виділити тип середовища, об'єкти та події, зображені на кадрі. Якщо було отримано новий тип, він буде доданий до багатовимірної моделі.

Багатовимірна модель представлення та аналізу даних

Для систематизації та збереження інформації використано багатовимірну модель Кодда, яка орієнтована на аналітичні запити та агрегування даних у вигляді інформаційних кубів. На відміну від класичних реляційних моделей, цей підхід дозволяє користувачеві розглядати дані одночасно у кількох вимірах.

Математично багатовимірний куб даних у запропонованій системі можна виразити як функцію, що відображає декартовий добуток множин вимірів у простір числових мір:

$$F: T \times L \times O \times E \times ET \rightarrow R^m, \quad (1)$$

де $m_i \epsilon R$ – числові міри, наприклад, кількість виявлених об’єктів

У межах дослідження ієрархічну структуру куба формують такі виміри:

- Час (Т) – агрегування результатів детектування за ієрархією «день – місяць – рік»;
- Простір – агрегування результатів детектування за ієрархією «населений пункт – область»
- Тип об’єкту – класифікація за структурою «домен – клас»
- Оточення – опис середовища за ієрархією “домен – тип»
- Тип події – класифікація подій на відео за структурою «домен – тип»

Для проведення аналітичних досліджень над моделлю визначено базові OLAP операції, які дозволяють отримувати специфічні зрізи даних. Зокрема, операція Slice фіксує значення одного виміру та формує відповідну підмножину даних. Наприклад, вибір даних для певного дня $d \epsilon T_1$ визначається як:

$$F_d = \{ f \epsilon F \mid h^T_{(0,1)}(\pi_T(f)) = d \}, \quad (2)$$

де $\pi_T(f)$ – проекція на вимір часу. Операція Dice дозволяє виконувати більш складну селекцію за декількома умовами одночасно. Наприклад, для обчислення кількості військових об'єктів у визначених областях за конкретний місяць (де $t_1 \epsilon T_1, l_1 \epsilon L_1, domain = \text{"військовий"}$), агрегована міра виражається наступним відношенням:

$$M = \text{card}\{f \in F \mid h^T_{(0,1)}(\pi_T(f)) = t_1 \wedge h^L_{(0,1)}(\pi_L(f)) = l_1 \wedge \pi_O(f) = \text{військовий}\} \quad (3)$$

Логічна структура системи реалізована через три основні фактові таблиці: Object_Tracking_Fact для обліку унікальних об'єктів, Video_Environment_Fact для фіксації типів середовища та Video_Event_Fact для збереження даних про події.

Логічна структура системи реалізована через три основні фактові таблиці: Object_Tracking_Fact для обліку унікальних об'єктів, Video_Environment_Fact для фіксації типів середовища та Video_Event_Fact для збереження даних про події.

Основними аналітичними мірами моделі є кількість виявлених об’єктів (на основі object_id), що вказує на інтенсивність руху, та тривалість перебування об’єкта в зоні спостереження. Використання такої сукупності мір дозволяє проводити поглиблений аналіз динаміки подій, поєднуючи дані про ідентифіковані об'єкти з детальним описом умов, у яких вони були зафіксовані.

Розробка програмного забезпечення

Архітектура та функціональні компоненти системи

Запропонований метод покладено в основу програмного комплексу, архітектура якого наведена на рис. 2. Вона базується на мікросервісному підході з кластерними обчисленнями, що забезпечує високу доступність та масштабування. Система поділена на

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

рівні аналізу, координації та збереження даних. Центральний елемент – розподілена мережа вузлів, об'єднаних у кластери:

– Detection Cluster. Ідентифікація об’єктів у реальному часі за допомогою моделі YOLO, з паралельним розподілом навантаження.

– Semantic Analyze Cluster: Глибинний семантичний аналіз сцени за допомогою мультимодальних моделей (VLM), що формують описи контексту.

Для незалежності та гарантованої доставки результатів впроваджено брокер повідомлень (MessageBroker). Асинхронний обмін даними за протоколом AMQP дозволяє накопичувати завдання в чергах, що критично при пікових навантаженнях.

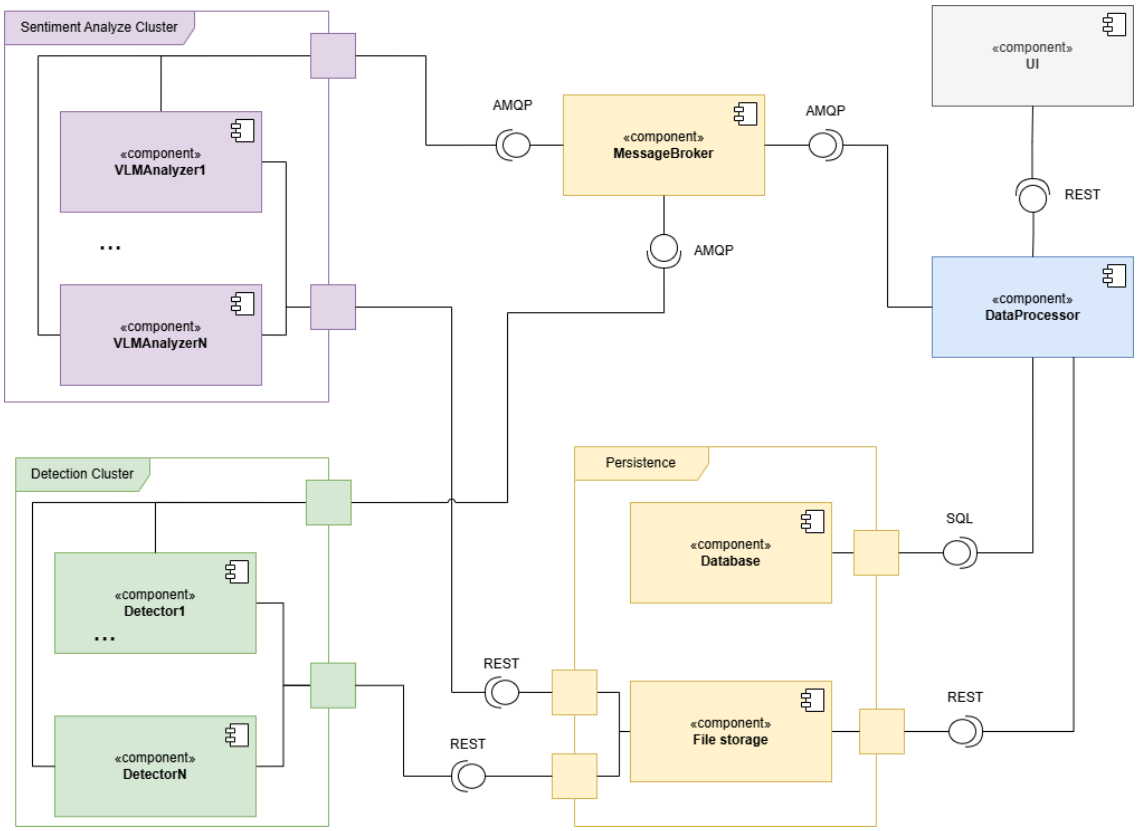

Рисунок 2 – Схема архітектури розробленого програмного забезпечення

Координація та збереження даних

Роль координатора робочих процесів виконує компонент Data Processor. Він агрегує результати з брокера повідомлень, виконує їх фінальну структуризацію та забезпечує наповнення аналітичного сховища. Взаємодія з клієнтським інтерфейсом та отримання медіафайлів реалізовані через REST API.

Система використовує гібридну модель збереження даних для розділення “важкого” контенту та структурованої аналітики: файлове сховище (MinIO) для збереження об’ємних відеофайлів та кадрів та реляційну базу даних (PostgreSQL) зі сховищем даних ROLAP за схемою «зірка».

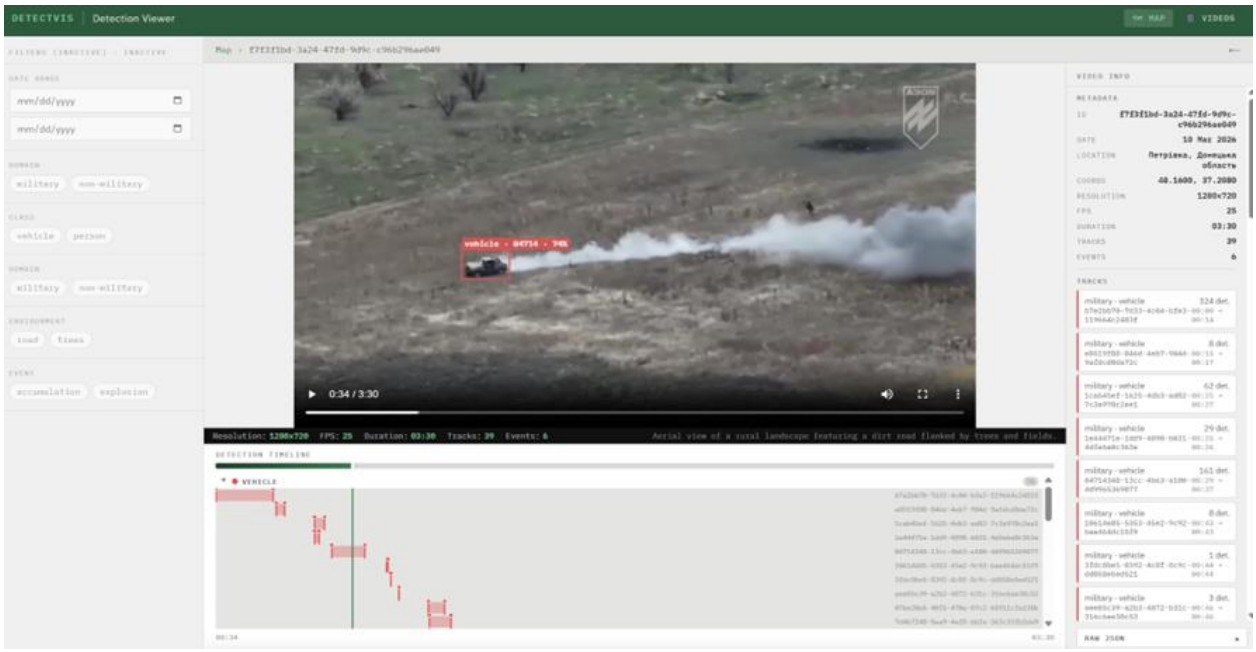

Рисунок 3 – Приклад відображення зібраної інформації про відео

Результати експериментальних досліджень. Для оцінки ефективності запропонованої системи було проведено серію експериментів, що охоплює ключові аспекти продуктивності та ресурсоємності. Дослідження фокусуються на чотирьох основних метриках: швидкості та точності роботи моделі YOLO залежно від розміру моделі, продуктивності VLM-моделей з урахуванням запитів і відповідей, загальній швидкості всього алгоритму обробки, а також обсязі даних, що генеруються при інтеграції відео в OLAP-модель. Тестування проводилося на апаратній платформі з GPU NVIDIA RTX 4070, 8 ГБ RAM та процесором Intel Core i7, з використанням підготовленого датасету військової техніки та відео з дронів.

Швидкість і точність YOLO залежно від розміру моделі

Для визначення найбільш ефективної конфігурації нейромережі проведено порівняльний аналіз моделей сімейства YOLOv26 варіантів Nano, Medium та Large. Експерименти проводилися на відео роздільною здатністю 1080p з частотою 30 FPS. Результати вимірювань, наведені у таблиці 1, свідчать, що найбільш раціональним рішенням для поставленої задачі є модель YOLOv26m. Дана конфігурація забезпечує високий рівень точності (map-50 = 0.814, precision = 0.869) при збереженні продуктивності на рівні 81.03 FPS. Таке співвідношення характеристик дозволяє системі виконувати стабільну детекцію та відстеження об’єктів у режимі, близькому до реального часу, мінімізуючи затримки при обробці динамічних сцен.

Оцінка точності та швидкодії VLM моделей

Для визначення найефективнішого рішення для задачі генерації описів до зображень було проведено порівняльне експериментальне дослідження чотирьох сучасних візуально-мовних моделей (VLM): Moondream2, Qwen2.5, PaliGemma та BLIP-2. Експерименти виконувалися з використанням обчислювальних потужностей графічного прискорювача NVIDIA A100. Оцінювання проводилося на випадковій підвибірці з 200 зо-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

бражень зі спеціалізованого датасету RSICD (Remote Sensing Image Captioning Dataset). Кількісна оцінка якості згенерованих текстів здійснювалася за допомогою стандартних метрик: BLEU-1, METEOR, Jaccard Index, а також CIDEr. Швидкодія моделей вимірювалася як середній час генерації опису для одного зображення. Усі метрики якості наведено у масштабі від 0 до 100. Зведені результати експериментів представлені у Таблиці 2.

Таблиця 1

Порівняння швидкості та точності моделей YOLO

Модель	Кількість епох	map50	map50-95	Precision	Recall	FPS
YOLOv26n	20	0.794	0.423	0.819	0.72	98.03
YOLOv26m	20	0.814	0.448	0.869	0.723	81.03
YOLOv26l	30	0.821	0.4682	0.871	0.741	66.31

Таблиця 2

Порівняння швидкості та точності VLM моделей

Модель	BLEU-1	METEOR	Jaccard Index	CIDEr	Швидкість обробки, (с)
paligemma-3b-pt-224	32.48	18.27	18.84	11.54	0.321
blip2-opt-2.7b	17.91	18.6	18	10.34	1.318
qwen2.5-VL-3B-Instruct	45.92	24.22	23.15	14.94	0.851
moondream2	38.01	25.73	21.6	12.55	0.422

Аналіз отриманих даних показує, що моделі Qwen2.5 та Moondream2 є беззаперечними лідерами за якістю генерації тексту. Хоча Qwen2.5 демонструє найвищі показники точного збігу слів (BLEU-1: 45.92, Jaccard: 23.15) та специфічності (CIDEr: 14.94), архітектура Moondream2 виявляється найкращою та найбільш збалансованою моделлю для практичного застосування. По-перше, Moondream2 демонструє найвищі результати за метрикою METEOR (25.73). Це свідчить про те, що модель краще зберігає загальний семантичний зміст та структуру речень, навіть якщо використовує синоніми. По-друге, Moondream2 працює більш ніж удвічі швидше за Qwen2.5 (0.422 секунди на зображення проти 0.851), поступаючись у швидкості лише PaliGemma, яка при цьому має значно гіршу якість генерації.

Варто також відзначити важливий якісний аспект генерації, який складно повною мірою відобразити кількісними метриками. Моделі Moondream2 та Qwen2.5 схильні генерувати дуже детальні, розгорнуті описи, які містять глибокий аналіз контексту зо-

браження. Натомість еталонні описи у датасеті RSICD здебільшого є короткими та лаконічними. У результаті класичні метрики (такі як BLEU та CIDEr) "штрафують" ці моделі за генерацію додаткових деталей, яких немає в еталоні. На відміну від PaliGemma та BLIP-2, які генерують простіші конструкції, Moondream2 надає значно багатший контекст, що робить її найефективнішим вибором для задач, де потрібне глибоке візуальне розуміння сцени при збереженні високої швидкості обробки даних.

Оцінка масштабованості та продуктивності системи

Для перевірки здатності розробленої архітектури ефективно розподіляти навантаження та працювати з великими масивами даних було проведено тестування масштабованості. Експеримент виконувався з використанням обчислювальних потужностей графічного прискорювача NVIDIA RTX 4070.

Тестовий набір даних складався з 100 відеороликів тривалістю 210 секунд кожен. Базова кадрова частота відео становила 25 FPS, проте для оптимізації обчислювальних ресурсів без втрати інформативності з відеопотоку вилучався кожен 5-й кадр.

Оцінювання проводилося для двох ключових етапів пайплайну: детекції та семантичного аналізу. Щоб виміряти ефективність горизонтального масштабування системи, обробка імітувалася з використанням 1, 5 та 10 паралельних сервісів. Зведені результати експерименту наведено у таблиці 3.

Таблиця 3

Залежність часу обробки від кількості підключених сервісів

Кількість сервісів	Час обробки 1 відео в середньому, с		Час обробки 100 відео, с	
	Детекція	Семантичний аналіз	Детекція	Семантичний аналіз
1	36.5	3.9	4351.2	397.3
5	35.34	3.78	935.74	85.64
10	36.45	3.86	486.16	45.93

Аналіз загального часу обробки масиву зі 100 відео демонструє високу здатність системи до масштабування. Збільшення пулу обробників з 1 до 5 дозволило скоротити загальний час детекції з 4351.2 с до 935.74 с, тобто прискорення у 4.65 разів. При підключенні 10 паралельних сервісів загальний час детекції впав до 486.16 с, що забезпечило прискорення у 8.95 разів порівняно з одним сервісом. Аналогічна тенденція спостерігається і для етапу семантичного аналізу, де розпаралелювання на 10 воркерів зменшило час обробки зі 397.3 с до 45.93 с.

Незначне відхилення показників від ідеального математичного лінійного прискорення, наприклад, 486.16 с замість теоретичних 435.12 с для 10 сервісів, є очікуваним. Воно пояснюється накладними витратами на диспетчеризацію завдань, мережеву взаємодію між мікросервісами, черги повідомлень та операції читання і запису відеофайлів.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Результати підтверджують, що запропонована архітектура не має внутрішніх блокувань при паралельній роботі і дозволяє лінійно нарощувати пропускну здатність системи шляхом додавання нових обчислювальних вузлів.

Висновки. У роботі розроблено метод та масштабований програмний засіб для автоматизованої екстракції, семантичного збагачення та багатовимірного аналізу відеоданих. Наукова новизна роботи полягає у поєднанні каскадної обробки відеопотоків із використанням моделі YOLOv26m та мультимодальної моделі Moondream2 з ROLAP-поданням даних у багатовимірному кубі, що забезпечує безперервний перехід від неструктурованого відео до аналітичних зрізів у режимі, близькому до реального часу. Спроєктована мікросервісна архітектура з використанням брокера повідомлень забезпечує паралельну обробку та відмовостійкість системи, що підтверджено серією експериментальних досліджень.

У частині детекції та відстеження об'єктів (зокрема, військової техніки) оптимальною визначено конфігурацію нейромережі YOLOv26m. Вона забезпечує високу точність (mAP50 = 0.814) при стабільній роботі в реальному часі (81.03 FPS). Для етапу семантичного збагачення найкращий баланс аналізу та швидкості продемонструвала мультимодальна модель Moondream2 (METEOR: 25.73, 0.422 с/кадр), яка генерує детальні просторові описи без необхідності донавчання базисної архітектури.

Оцінка продуктивності програмного комплексу довела його здатність до лінійного прискорення при масштабуванні. Розширення пулу обчислювальних воркерів з 1 до 10 дозволило скоротити загальний час обробки масиву відео в 8.95 разів (з 4351.2 с до 486.16 с), що підтверджує відсутність внутрішніх блокувань у системі обміну повідомленнями.

Створено комплексний інструмент для високонавантажених систем відеомоніторингу, який успішно перетворює великі обсяги неструктурованих відеоданих у формат, придатний для гнучкого багатовимірного аналізу та швидкого прийняття рішень.

ЛІТЕРАТУРА

1. Mahmud A., Setiawan A. A. A. A survey of convolutional neural networks in object detection // Int. J. Adv. Comput. Sci. Appl. – 2021.

2. Redmon J., Divvala S., Girshick R., Farhadi A. You Only Look Once: Unified, Real-Time Object Detection // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016.

3. Abouelyazid M. Comparative Evaluation of SORT, DeepSORT, and ByteTrack for Multiple Object Tracking in Highway Videos // International Journal of Sustainable Infrastructure for Cities and Societies. – 2023. – Vol. 8, No. 11. – P. 42–52.

4. Din M. U., Akram W., Bakht A. B., Hussain I. LLM-VLM Fusion Framework for Autonomous Maritime Port Inspection using a Heterogeneous UAV-USV System. – Текст. – arXiv:2601.13096 [cs.RO]. – 2026. – DOI: 10.48550/arXiv.2601.13096.

5. Hansung L., Sohee P., Jang-Hee Y. A Data Cube Model for Surveillance Video Indexing and Retrieval // SIGMAP 2013: International Conference on Signal Processing and Multimedia Applications. – 2013. – URL: https://www.scitepress.org/Papers/2013/46121/46121.pdf.

6. Wu Y., Zhang C., Lu Y., Su Y., Jiang X., Xiang Z., Li Z. VideoARD: An Analysis-Ready Multi-Level Data Model for Remote Sensing Video // Remote Sens. – 2025. – Vol. 17, No. 22. – Art. 3746. – DOI: 10.3390/rs17223746.

REFERENCES

1. Mahmud, A., & Setiawan, A. A. A. (2021). A survey of convolutional neural networks in object detection. *Int. J. Adv. Comput. Sci. Appl.*

2. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

3. Abouelyazid, M. (2023). Comparative evaluation of SORT, DeepSORT, and ByteTrack for multiple object tracking in highway videos. *International Journal of Sustainable Infrastructure for Cities and Societies*, *8*(11), 42–52.

4. Din, M. U., Akram, W., Bakht, A. B., & Hussain, I. (2026). *LLM-VLM fusion framework for autonomous maritime port inspection using a heterogeneous UAV-USV system*. arXiv. https://doi.org/10.48550/arXiv.2601.13096

5. Hansung, L., Sohee, P., & Jang-Hee, Y. (2013). A data cube model for surveillance video indexing and retrieval. In *SIGMAP 2013: International Conference on Signal Processing and Multimedia Applications*. https://www.scitepress.org/Papers/2013/46121/46121.pdf

6. Wu, Y., Zhang, C., Lu, Y., Su, Y., Jiang, X., Xiang, Z., & Li, Z. (2025). VideoARD: An analysis-ready multi-level data model for remote sensing video. *Remote Sens.*, *17*(22), Art. 3746. https://doi.org/10.3390/rs17223746

Received 07.04.2026.
Accepted 10.04.2026.
Published 30.04.2026

A method and software tool for video object recognition and analysis

The modern scientific community is paying significant attention to the automation of video analytics. Traditional methods based on convolutional neural networks (CNNs), such as Faster R-CNN, demonstrate high accuracy but have limited processing speed for real-time streams. In contrast, single-stage algorithms in the YOLO family achieve the required performance. Research in object tracking (Tracking-by-Detection) highlights DeepSORT and ByteTrack as the most effective algorithms for associating detections across frames. The application of multimodal vision-language models (VLMs) opens up new possibilities for semantic scene description, although their implementation in high-load systems is hindered by computational complexity. Additionally, multidimensional data analysis (OLAP) technologies are considered, specifically the SurvCube and VideoCube models, which integrate video processing results into structured cubes; however, they often offer limited flexibility for creating new semantic hierarchies.

The objective of the research is to increase the analysis speed of ultra-large video datasets by developing a method and software tool for the automated extraction of structured data and its subsequent integration into a multidimensional OLAP model that supports parallel processing and semantic information enrichment.

A method for cascaded video data processing is proposed, combining rapid object detection with the YOLOv26m architecture and dynamic semantic scene enrichment using the

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Moondream2 multimodal model. During the preprocessing stage, metadata extraction (coordinates, time) and load optimization are performed by downsampling the frame rate to 2–5 frames per second. The YOLOv26m model is used for object detection and classification, achieving an mAP50 of 0.814 at 81.03 FPS. Tracking is implemented by assigning a unique object_id, which minimizes data redundancy. Semantic context enrichment (landscape type, events) is carried out by the multimodal Moondream2 model using text prompts, allowing the system to adapt to new scenarios without retraining the network. The data is integrated into a multidimensional OLAP model using a star schema, where the dimensions are time, space, object type, environment, and event type. The software features a microservices architecture utilizing a message broker (RabbitMQ/AMQP) for asynchronous communication between the detection and semantic analysis clusters.

In the course of this work, a scalable software complex was developed that successfully transforms unstructured video streams into the format of analytical cubes. It was experimentally confirmed that the use of the Moondream2 model provides the best balance between description quality (METEOR: 25.73) and processing speed (0.422 s/frame). Scalability testing demonstrated the architecture's capability for linear speedup: increasing the number of parallel workers to 10 reduced the total video dataset processing time by a factor of 8.95. The proposed solution is effective for high-load monitoring systems and rapid decision-making based on deep analytics.

Keywords: computer vision, video analytics, object detection, object tracking, multidimensional model, OLAP, YOLO, VLM, Moondream2, semantic enrichment.

Ніколаєв Іван Романович – магістрант кафедри інформатики та програмної інженерії Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського».
ORCID: https://orcid.org/0009-0004-7779-8329
Олійник Юрій Олександрович – доцент кафедри інформатики та програмної інженерії Національного технічного університету України «КПІ ім. Ігоря Сікорського».
ORCID: https://orcid.org/0000-0002-7408-4927

Nikolaiev Ivan - Master's student at Department of Computer Science and Software Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.
ORCID: https://orcid.org/0009-0004-7779-8329
Oliinyk Yurii - Ph.D, associate professor at department of computer science and software engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.
ORCID: https://orcid.org/0000-0002-7408-4927