COMPUTER VISION AND PATTERN RECOGNITION
КОМП'ЮТЕРНИЙ ЗІР ТА РОЗПІЗНАВАННЯ ОБРАЗІВ

https://doi.org/10.15407/intechsys.2025.05.022
УДК 004.51

В.М. КИЙКО, канд. техн. наук, старш. наук. співроб., пров. наук. співроб.,
Інститут інформаційних технологій та систем НАН України,
просп. Акад. Глушкова, 40, м. Київ, 03187, Україна
https://orcid.org/0009-0005-6605-0339
kiiko@gmail.com

ВИЗНАЧЕННЯ МАСШТАБУ ТА КУТА ПОВОРОТУ ДЛЯ ДОВГОСТРОКОВОГО ВІДСТЕЖЕННЯ ОБ'ЄКТА У ВІДЕО

Надійність відстеження у відео значною мірою залежить від ефективності (точності та порівняно малої обчислювальної складності) задіяних алгоритмів визначення масштабу та кута повороту об'єкта відстеження на зображеннях. Пропонується алгоритм для оцінки цих параметрів на основі пошуку відповідних ключових точок (КТ) на кожному кадрі до КТ у моделі об'єкта М, що складається із КТ об'єкта та навколишнього фону. Алгоритм переважно може бути задіяний в умовах, коли зміни масштабу та кута повороту головним чином є наслідком змін руху камери або дій оператора і в значній мірі корелюють зі змінами на фоні, що зазвичай відповідає відеоспостереженню з літального апарата, зокрема БПЛА. Переваги алгоритму полягають у тому, що є порівняно більш стійким до наявності помилок в визначенні відповідних пар КТ, а також може бути використаний під час тривалої відсутності об'єкта у відео для оцінки масштабу та кута повороту шляхом пошуку КТ на зображенні, що відповідають до КТ фону у М. Це виконується з метою оновлення моделі об'єкта і його детектування після появи у відео зі значно зміненими параметрами.

Наведено приклади використання алгоритму для довгострокового відстеження із застосуванням пропонованого критерію наявності об'єкта в полі зору камери, а також двох способів оновлення М за його присутності або відсутності на зображеннях.

Ключові слова: *відстеження об'єктів, BRISK ключові точки, KCF алгоритм відстеження, HOG ознаки, масштаб та кут повороту об'єкта на зображенні, детектування та розпізнавання об'єктів.*

Цитування: Кийко В.М. Визначення масштабу та кута повороту для довгострокового відстеження об'єкта у відео. Information Technologies and Systems, Київ, 2025, Том 5 (5), 22–38. https://doi.org/10.15407/intechsys.2025.05.022

Вступ

Задача відстеження полягає у безперервному визначенні положення довільного об'єкта (цілі) у відео після того як його детектував оператор або інші засоби на початковому кадрі. Алгоритми (трекери) для виконання відстеження розділяють на такі, що забезпечують короткочасне [1–8] або більш складне довгострокове (тривале) відстеження у відео [9–17]. Довгострокове відстеження стосується визначення місцезнаходження довільної цілі у відносно довгому відео в умовах змін подання на зображеннях, а також можливого порівняно тривалого зникнення з поля зору камери. Ключовою проблемою довгострокового відстеження є відновлення цілі після періоду відсутності або збоїв відстеження [15], тому що за цей час можуть значно змінитись не тільки координати, а також масштаб та кут повороту цілі, знання яких збільшує точність та надійність детектування. Внаслідок цього після тривалого зникнення пошук цілі необхідно виконувати не локально, а в межах всього зображення і значних інтервалів можливих змін масштабу та кута повороту.

Основні складнощі задачі відстеження полягають у різноманітті об'єктів та їх подання внаслідок зміни масштабу, орієнтації, освітлення, перекриття або зникнення з поля зору, а також вимогою відстеження у реальному часі. Наразі не існує єдиного методу подолання всіх цих складнощів, а скоріше сукупність алгоритмів та засобів для зменшення впливу на результат відстеження кожного з цих факторів. У цій статті переважно розглядаються та пропонуються алгоритми для забезпечення надійності детектування та відстеження за порівняно значних змін масштабів та орієнтації об'єктів. Незважаючи на наявність таких алгоритмів, поки що розпізнавання об'єктів у значно різних масштабах залишається однією з недостатньо вирішених проблем комп'ютерного зору [29]. Надалі задача відстеження розглядається в поширених на практиці умовах, коли об'єкт зміщується на зображеннях у відео переважно повільно, а зміни його масштабу та кута повороту головним чином є наслідком змін параметрів руху камери та дій оператора і значною мірою корелюють зі змінами відповідних параметрів фону. Такі умови зазвичай відповідають відеоспостереженню з літального апарата, зокрема БПЛА, що знаходиться достатньо високо над поверхнею Землі.

Найбільш відомими є такі три різновиди алгоритмів для відстеження цілі у відео: 1) дискримінаційні кореляційні фільтри [1–3, 9–13, 22, 23]; 2) на основі навчання на нейронних мережах [4, 5, 14–19]; і 3) через детектування та пошук відповідності так званих «ключових точок» (КТ) на порівнюваних зображеннях [6–8]. Надійність відстеження з допомогою кожного з цих алгоритмів значною мірою залежить від точності визначення масштабу та кута повороту об'єкта на зображенні, а також спроможності забезпечувати інваріантність на-

дійного відстеження від цих параметрів. Зазначені алгоритми відстеження застосовують різні підходи і засоби, в тому числі для визначення масштабу та кута повороту, мають свої області для кращого використання і можуть сумісно використовуватись, взаємно доповнюючи один одного [8–12, 17].

Привабливі властивості трекерів на основі визначення відповідності КТ полягають зокрема у можливості відстеження не тільки так званих «жорстких» об'єктів, а й таких, що складаються з кількох частин, які можуть незалежно один від одного змінюватись з часом або зникати з поля зору. Алгоритми [6–8] мають також наступні недоліки. Оцінювання змін масштабу та повороту на основі відповідності тільки КТ об'єкта без врахування КТ фону не є достатньо надійним, особливо в умовах малої кількості КТ. Зазначені алгоритми не призначено для визначення змін масштабу та кута повороту на зображеннях за відсутності на них об'єкта відстеження і тому можуть забезпечувати відстеження в умовах лише короткочасної відсутності об'єкта, під час якої масштаб та кут повороту мало змінюються.

Метою дослідження є розроблення алгоритму визначення змін масштабу та кута повороту об'єкта, який долає вказані недоліки для отримання надійніших результатів відстеження об'єктів у складних умовах.

Отримано такі основні результати:

• розроблено алгоритм визначення змін масштабу та кута повороту об'єкта у відео на основі відповідних КТ об'єкта і фону за наявності значної кількості помилок у знайденій відповідності КТ на порівнюваних зображеннях;

• запропоновано спосіб визначення масштабу та кута повороту зображень у відео за відсутності об'єкта на цих зображеннях з метою оновлення моделі об'єкта і його детектування після появи на зображеннях із суттєво зміненими значеннями цих параметрів.

Далі у статті наведено огляд відомих алгоритмів для відстеження та визначення масштабу об'єктів на зображеннях на основі використання дискримінаційних кореляційних фільтрів, нейронних мереж та пошуку відповідності КТ на зображеннях, подано розроблені алгоритми визначення змін масштабу та кута повороту об'єкта, що використовуються в алгоритмі для довгострокового відстеження об'єкта у відео, що базується також на моделі об'єкта *M* і застосованих алгоритмах пошуку відповідності КТ на кожному кадрі до КТ у *M* та оновлення *M* в процесі відстеження.

Огляд споріднених алгоритмів

Дискримінаційні кореляційні фільтри використовують швидке перетворення Фур'є (ШПФ) при навчанні та відстеженні об'єктів у відео. Навчання фільтра *W* виконується на основі лінійної або нелінійної

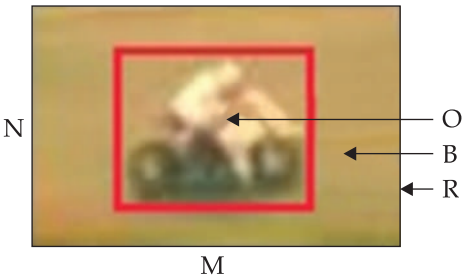

Рис. 1. Ділянки зображень об'єкта (*O*) та навколишнього фону (*B*) під час навчання фільтрів.

регресії шляхом вирішення відповідно однієї з двох наступних задач пошуку мінімуму цільової функції як суми квадратичних відхилень кореляцій від заданих значень [1, 2]:

$$W = \arg\min_{w} \sum_{m,n} |\langle x_{m,n}, w\rangle - y(m,n)|^2 + \lambda |w|^2, \quad (1)$$

$$W = \sum_{m,n} \alpha(m,n) \cdot \varphi(x_{m,n}) = \arg\min_{w} \sum_{m,n} |\langle \varphi(x_{m,n}), w\rangle - y(m,n)|^2 + \lambda |w|^2, \quad (2)$$

де $x = x_{0,0}$ — зображення ділянки $M \times N$ на початковому кадрі, що містить об'єкт в оточенні навколишнього фону (Рис. 1); $x_{m,n}$, ($m = 0, 1 \ldots M - 1$, $n = 0, 1 \ldots N - 1$) — сукупність прикладів об'єкта для навчання шляхом застосування циклічного оператора циркулянтної матриці для зсувів еталонного зображення x у горизонтальному та вертикальному напрямках; $y(m, n)$ — мітки, що відповідають $x_{m,n}$ і дорівнюють відстаням по функції Гауса між x та $x_{m,n}$; φ — невідома функція відображення первинних ознак на зображенні у новому просторі, що індукована ядром k (Гаусова функція близькості двох зображень з первинними ознаками як яскравості або коди кольору пікселів) і задовольняє умові $\langle \varphi(f), \varphi(g)\rangle = k(f, g)$; α — матриця коефіцієнтів (двоїстих змінних), що визначається внаслідок вирішення задачі (2); $\lambda \geq 0$ — параметр регуляризації, для протидії так званому «перенавчанню» та діленню на нуль за обчислень α за формулою (3).

Задача навчання (1) вирішується з використанням первинних ознак зображення об'єкта, а (2) — більш інформативних ознак, у тому числі *HOG* ознак [21], що є результатом нелінійного перетворення цих первинних ознак. Коефіцієнти α в результаті вирішення (2) визначаються за формулою

$$A = F(\alpha) = \frac{F(y = \{y(m,n) \,|\, (m,n) \in \{0,1,\ldots,M-1\} \times \{0,1,\ldots,N-1\}\}}{F\{k(x_{m,n}, x)\} + \lambda}, \quad (3)$$

якщо $k(f_{m,n}, g_{m,n}) = k(f, g)$ для всіх m, n, f, g, що виконується зокрема для Гаусова ядра k у алгоритмі [2] (*KCF — kernelized correlation filter*).

Після навчання пошук об'єкта виконується на ділянці z у наступному кадрі з розмірами $M \times N$ і центральною точкою, що збіга-

ється з положенням об'єкта у попередньому кадрі, шляхом обчислення матриці значень відгуків фільтра за формулою

$$\overline{y} = F^{-1}(A \circ F(\varphi(z) \cdot \varphi(z))), \quad (4)$$

де $\overline{x}$ — подання об'єкта з допомогою ознак, що оновлюється під час відстеження; $\circ$ — поелементне множення двох матриць. Після цього визначаються координати об'єкта, що відповідають положенню максимального значення відгуку на матриці $\overline{y}$.

Модель об'єкта складається з матриці коефіцієнтів A та подання $\overline{x}$ об'єкта за допомогою *HOG* (*Histogram of Oriented Gradients*) ознак [21] і оновлюється під час відстеження за формулами:

$$\begin{aligned} \overline{x}(t) &= (1-\beta)\,\overline{x}(t-1) + \beta\,\overline{x}(t), \\ \overline{A}(t) &= (1-\beta)\overline{A}(t-1) + \beta\overline{A}(t), \end{aligned} \quad (5)$$

де t — індекс поточного кадра і β — коефіцієнт оновлення моделі при навчанні.

Гіпотетично *KCF* може використовуватись не тільки для визначення нового положення цілі на кожному кадрі, а також змін масштабу, якщо застосовувати в ньому не двовимірні, а тривимірні фільтри із значеннями масштабу у третьому вимірі. Однак, таке рішення не є ефективним унаслідок значного зростання обчислювальної складності алгоритму. Тому було розроблено порівняно швидкі алгоритми [10, 23, 24] для визначення змін масштабу після визначення положення цілі шляхом побудови піраміди зображень з різним масштабом і ознаками на кожному рівні, обчислення значень кореляції на всіх рівнях піраміди і визначення нового масштабу, що відповідає максимальному з цих значень.

У [10] використовуються два фільтри, що навчаються на основі вирішування задачі нелінійної регресії. Перший фільтр R_C застосовується для визначення положення об'єкта, використовує двовимірну сукупність зображень об'єкта та фону з розмірами $M \times N$ для навчання (приклад на Рис. 1), а другий фільтр R_T — масштабу об'єкта на кожному кадрі з використанням порівняно малої одномірної сукупності прикладів зображень тільки об'єкта з різними масштабами. Хай $P \times Q$ — розміри зображення z_t об'єкта (в межах ділянки o на Рис. 1) на поточному t-му кадрі, d — коефіцієнт зміни масштабу і N_S — кількість рівнів піраміди і оцінюваних масштабів $S = \left\{ d^n \mid n = \left\lfloor -\frac{N_S - 1}{2} \right\rfloor, \left\lfloor -\frac{N_S - 3}{2} \right\rfloor, \ldots, \left\lfloor \frac{N_S - 1}{2} \right\rfloor \right\}$ відносно масштабу об'єкта на попередньому кадрі. На першому етапі алгоритму для кожного з масштабів $s \in S$ на зображенні виділяється ділянка J_S з розмірами $sP \times sQ$ і центром в попередньо визначеній точці положення об'єкта, після чого розміри J_S змінюються на $P \times Q$. Далі обчислюється

подання x_S зображення J_S з допомогою ознак з подальшим обчисленням карти відгуків Y_S фільтра R_T за формулою (4). На заключному етапі визначається масштаб $s_t = \arg\max_s \left(\max \left(Y_s \right) \middle| \; s = 1, \; ..., \; S \right)$. Оновлення обох фільтрів відбувається тільки за умови, що максимальні значення відгуків фільтрів перевищують попередньо задані порогові значення. У [23] на відміну від [10] використовується лінійна регресія під час навчання фільтрів, розмір зображень J_S не змінюється на $P \times Q$, і оновлення фільтрів відбувається на кожному кадрі. У [24] на відміну від [10, 23] використовується тільки один фільтр з розмірами $M \times N$ для визначення координат і масштабу об'єкта на кожному кадрі.

Для прикладу, якщо d =1,02 і зміна масштабу на сусідніх кадрах може бути від 0 до 6%, то N_S =7 і S={0,94, 0,96, 0,98, 1, 1,02, 1,04, 1,06}. Це є цілком прийнятним для визначення масштабу на сусідніх кадрах і пошуку об'єкта в межах малої ділянки зображення. Але якщо об'єкт був відсутній на значній кількості кадрів, то, по-перше, необхідно виконувати пошук на всьому зображенні, а, по-друге, його масштаби перед зникненням і після появи можуть значно відрізнятись між собою, що значно збільшує обчислювальну складність пошуку. Наприклад, якщо припустити, що масштаби можуть відрізнятись в 6 разів, то N_S = |S| = 90, що практично виключає доцільність використання кореляційних фільтрів за таких умов довгострокового відстеження об'єкта.

Отримання карти відгуків для визначення масштабу може бути реалізоване також шляхом обчислення нормалізованих кореляцій подань зображень z_t і J_S з допомогою ознак. Це потребує менше операцій порівняно з використанням фільтра R_T, але може знизити точність визначення масштабу.

Трекери на основі глибинного навчання на нейронних мережах (глибинні трекери), які попередньо навчаються на великій кількості даних, демонструють порівняно високу надійність відстеження об'єктів. Однак, хоча глибинні трекери демонструють порівняно високу ефективність на відео з відомих баз даних (зокрема, *VOT* [20]), вони часто не підходять для середовищ з обмеженими обчислювальними ресурсами, таких як безпілотні літальні апарати (БПЛА). Один із напрямів виходу з таких ситуацій полягає у застосуванні полегшених варіантів глибинних трекерів [19], але завдяки зменшенню точності та надійності результатів відстеження.

Ідентифікація об'єктів різних розмірів на зображенні в глибинних трекерах часто виконується шляхом побудови та використання піраміди ознак, що поєднує геометричні та семантичні ознаки з різних шарів нейронної мережі. Основою для стандартного, але не завжди ефективного, рішення такої задачі є побудова піраміди ознак на піраміді зображень, тобто піраміди зображень з додаванням на кожному рівні відповідних ознак. Ці піраміди є інваріантними за

масштабом в тому сенсі, що дають змогу виявляти об'єкти у широкому діапазоні масштабів шляхом ковзного сканування їх моделей як за положеннями (координатами), так і за рівнями піраміди. Піраміди зображень з додаванням відповідних ознак і щільною дискретизацією масштабу широко використовувалися раніше в методах, що застосовували розроблені людиною ознаки. Наприклад, ознаки *HOG* [21] і *SIFT* [22] після щільного обчислювання по всіх пірамідах зображень використовувалися в численних роботах для класифікації зображень, виявлення об'єктів, оцінки пози людини тощо. Також виявлено значний інтерес до швидкого обчислення пірамід зображень з ознаками. Так, в [25] розглянуто алгоритм для швидкого обчислення пірамід на основі обчислення спочатку рідко дискретизованої (в масштабі) піраміди, а потім інтерполяції відсутніх рівнів.

З наступною появою глибоких згорткових мереж (*ConvNets*) [26] виявилось, що кількість рівнів у пірамідах ознак може бути зменшена, тому що *ConvNets* є більш стійкими до змін масштабу, але навіть з цією стійкістю багатомасштабне виявлення все ще працює краще порівняно з ознаками одного масштабу, особливо для малих об'єктів. Однак, збільшення рівнів піраміди зображень має очевидні обмеження внаслідок значного зростання часу виведення під час навчання, що робить цей підхід непрактичним для реальних застосувань. Більше того, мережі для глибинного навчання, що об'єднуються від початку до кінця на піраміді зображень, потребують надто значних затрат пам'яті, тому фактично такі піраміди можуть використовуватись лише під час тестування [27], що створює невідповідність між висновками під час навчання / тестування. З цих причин на основі досліджень [28] рекомендовано в *Fast/Faster R-CNN* використовувати ознаки, обчислені в одному масштабі, оскільки це є хорошим компромісом між точністю та швидкістю.

Крім піраміди зображень існують також інші способи обчислення багатомасштабного подання ознак. Глибока мережа обчислює ієрархію ознак шар за шаром, що має властиву багатомасштабну пірамідальну форму. Ця ієрархія ознак всередині мережі створює карти ознак з різною просторовою роздільною здатністю, але з тим недоліком, що різним рівням цієї піраміди відповідають різні можливості семантичного подання даних про зображення. Цей недолік може бути усуненим з допомогою алгоритму [29], результатом якого є піраміда ознак, яка має сильну семантику на всіх рівнях і швидко будується з одного вхідного масштабу зображення.

Основні алгоритми виявлення об'єктів, зокрема для відстеження, можна розділити на два типи: двоетапні та одноетапні. Двоетапні алгоритми виявлення (серія *R-CNN* [28]: *Faster R-CNN, Mask R-CNN, Cascade R-CNN, R-FCN*) спочатку генерують області-кандидати, а потім класифікують та регресують ці області. Ці алгоритми мають порівняно високу точність, особливо у виявленні малих об'єктів, але

вимагають генерації численних областей-кандидатів, що призводить до значної обчислювальної та часової складності.

Одноетапні алгоритми (серія *YOLO, SSD, RetinaNet*) розглядають виявлення об'єктів як загальну проблему регресії про їх місцезнаходження та розпізнавання. Серед них методи *YOLO* є наразі найбільш ефективними в галузі виявлення об'єктів, але як і інші поки не забезпечують достатньо надійний пошук малих за розмірами (менше 32 × 32 пікселів) та багатомасштабних об'єктів. Ознаки отримуються в *YOLO* з вхідного зображення з допомогою кількох операцій згортки в небагатьох масштабах, кількість яких, наприклад, дорівнює трьом (80 × 80, 40 × 40 і 20 × 20) в *YOLOv*8 [30]. Подальша побудова ефективної мережі піраміди ознак має визначальне значення для забезпечення ефективності багатомасштабного виявлення об'єктів.

Визначення масштабу та кута повороту об'єкта на основі пошуку відповідних КТ на зображеннях. В [6–8] на кожному *t*-му кадрі спочатку знаходиться сукупність $p_A^t = \left\{p_i^t, p_i^1\right\}_{i=1}^{N_A}$ пар відповідних КТ, після чого обчислюються масштаб s_t і кут повороту θ_t об'єкта за формулами:

$$s_t = med\left(\left\{\frac{\left\|p_i^t - p_j^t\right\|}{\left\|p_i^1 - p_j^1\right\|}\right\}, i \neq j\right), \tag{6}$$

$$\theta_t = med\left(\left\{a\tan 2\left(p_i^t - p_j^t\right) - a\tan 2\left(p_i^1 - p_j^1\right)\right\}, i \neq j\right), \tag{7}$$

де N_A — кількість пар відповідних КТ, *med* — медіана, p_i^t — *i*-та КТ на поточному кадрі і p_i^1 — КТ на першому кадрі, що є відповідною до p_i^t.

Пошук відповідних КТ і обчислення (6, 7) виконуються з використанням відповідних КТ об'єкта без врахування КТ фону, що забезпечує високу швидкість відстеження, але може призводити до зниження як точності значень s_t і θ_t, так і надійності відстеження, особливо якщо об'єкт має малу кількість КТ і може зникати на тривалий час з поля зору камери. Крім того, обчислення s_t і θ_t як медіанних значень за (6, 7) може забезпечити дійсні дані, тільки якщо кількість p_{Atrue}^t правильно знайдених пар відповідних КТ більше кількості помилкових пар — $p_{Atrue}^t > N_A * 0{,}5$.

Детектування та опис *BRISK* КТ на зображеннях

Відомо кілька алгоритмів визначення КТ і локальних інформативних областей в околі цих точок, кожному з яких відповідає певний тип КТ і спосіб опису областей. Цей опис має бути інформативним і, бажано, інваріантним до різних перетворень зображення, зокрема

до зміщення, повороту та масштабування, для пошуку відповідності КТ на зображеннях.

В [23] показано, що найбільшу точність мають *SIFT* КТ, але потребують значних обчислювальних затрат. Внаслідок цього при вирішенні багатьох задач активно використовуються *BRISK* (*Binary Robust Invariant Scalable Keypoints*) КТ [22]. Масштаб цих КТ визначається під час детектування шляхом застосування піраміди масштабованих представлень зображення. Після цього в околі КТ з розміром, відповідним знайденому масштабу, обираються 60 точок напівтонового зображення, які рівномірно розташовані на концентричних колах з центром у КТ, і на основі локальних градієнтів яскравості в цих точках визначається кут повороту КТ. На заключному етапі обчислюється бінарний дескриптор КТ в результаті порівняння яскравості в 512 парах точок в околі, який є 512-м бітовим числом. Таке бінарне подання забезпечує подальше швидке обчислення відстані Хемінга (кількості відповідних бітів з різними значеннями) між кожними двома дескрипторами з допомогою поширених *SSE* інструкцій, а також пошук відповідності між КТ на зображеннях. Дескриптори *BRISK* КТ є інваріантними до повороту та зміни масштабу на зображеннях.

Модель об'єкта

Модель $M = T \cup B$ об'єкта складається з множини T ключових точок на зображенні об'єкта, а також множини B ключових точок навколо цього об'єкта:

$$T = \left\{\left(d_i^T, p_i^T\right)\right\}_{i=1}^{N_T}, B = \left\{\left(d_i^B, p_i^B\right)\right\}_{i=1}^{N_B}, \quad (8)$$

де $d_i^T \left(d_i^B\right) \in Z^{512}$ — 512-вимірний бінарний дескриптор *BRISK* КТ об'єкта (фону); $p_i^T \left(p_i^B\right) \in R^2$ — відповідні координати КТ на зображенні.

Модель об'єкта M формується після виділення на початковому кадрі I_1 мінімального обмежувального об'єкт прямокутника R на цьому кадрі. Кожна знайдена в межах R точка вважається належною до КТ об'єкта, якщо відстань дескриптора цієї КТ до множини дескрипторів КТ навколишнього фону більше заданого порогового значення. Інші КТ в межах R, а також певна кількість відібраних в околі R належать до КТ фону в M.

На відміну від [6] у M використовуються не тільки КТ об'єкта, а і фону, що дає змогу: 1) виконувати більш точний пошук відповідних КТ на кожному кадрі I_t; 2) надійніше порівняно з [6] визначати масштаб та кут повороту об'єкта на I_t; 3) відновлювати пошук об'єкта після тривалого зникнення з поля зору і появи знову у відео з можливими значними змінами його масштабу та кута повороту.

Пошук КТ, відповідних до моделі об'єкта на зображеннях

Пошук на кожному кадрі I_t КТ, відповідних до КТ у M, відбувається шляхом виконання таких трьох етапів. На першому етапі визначається зміщення на поточному кадрі I_t знайдених відповідних до моделі КТ на попередньому I_{t-1} кадрі шляхом обчислення розрідженого оптичного потоку від I_{t-1} до I_t [33]. Це зміщення визначається достатньо точно, якщо рух об'єкта у відео є порівняно стабільним і повільним. Для підвищення надійності далі може бути виконана додаткова перевірка знайдених КТ на I_t через визначення їх положення на попередньому кадрі I_{t-1} шляхом застосування [33] для обчислення зворотного оптичного потоку від I_t до I_{t-1} [6–9].

Описані умови для локального пошуку відповідних КТ не завжди задовольняються, тому на другому етапі виконується так званий «глобальний» пошук КТ, відповідних до КТ у M, на основі обчислення відстані між дескрипторами по евклідовій метриці, що має відмінності від [6–8]. Спочатку для кожної КТ p_i^t на I_t знаходиться перелік із m КТ у M, які є найближчими по відстані між дескрипторами до p_i^t і упорядковані за зменшенням цієї відстані: $\left\{p_i, n_i, k_i, d_i\right\}_{i=1}^{m}$, де p_i, n_i, k_i, d_i — відповідно координати, ідентифікатор, клас (O — об'єкт, B — фон) і відстань i-ї точки у переліку до p_i^t, а p_0, n_0, k_0, d_0 — подібні дані про КТ у M, що є найближчою до p_i^t за відстанню між дескрипторами і належить об'єкту або фону. При цьому зазвичай, $m = 2$ і якщо виконуються умови $k_0 = O$, $d_0 < T_1^O$ і $d_0 / d_1 < T_2^O$, то для точки p_i^t на I_t відповідною є КТ об'єкта у моделі з ідентифікатором n_0. Інакше, якщо виконуються умови: $k_0 = k_1 = B$, $d_0 < T_1^B$ і $d_0 / d_1 < T_2^B$, тоді відповідною для p_i^t буде КТ фону з порядковим номером n_0 у M.

У наведених умовах застосовуються 4 порогові значення для прийняття рішень про відповідність p_i^t до КТ об'єкта $\left(T_t^O,\ T_2^O\right)$ або фону $\left(T_t^B, T_2^B\right)$ у M. Після завершення глобального пошуку відповідних КТ визначаються координати центральної точки об'єкта з допомогою алгоритма кластеризації [7].

На третьому етапі для кожної p_i^t на I_t виконується локальний пошук відповідної КТ у моделі об'єкта M у такий спосіб [6–8]. Спочатку обчислюється проєкція pr_i^t на сукупність КТ у M за формулою:

$$pr_i^t = H_t^{-1}\left(p_i^t - c_t\right) + c_{t1}, \tag{9}$$

де H_i^{-1} — зворотне перетворення з попередньо визначеними значеннями масштабу та повороту відносно визначеного центра об'єкта c_t на I_t; c_{t1} — центральна точка об'єкта на зображенні I_{t1}, яке використовувалось для створення ($I_{t1} = I_1$) або оновлення M.

Після цього визначається сукупність КТ у M, що є близькими за значенням координат з точністю до порогового значення до проєкції pr_i^t.

Далі серед КТ у цій сукупності знаходиться найближча p_j^{dm} до p_i^t за відстанню d між дескрипторами i, якщо $d\left(p_j^{dm}, p_i^t\right) < T_1^O$, тоді p_j^{dm} вважається відповідною до p_i^t.

Визначення масштабу та кута повороту об'єкта на основі пошуку відповідних КТ

Результатом пошуку відповідності КТ на поточному кадрі I_t до моделі об'єкта M є сукупність $p_A^t = \left\{p_i^t, p_i^M\right\}_{i=1}^{N_A}$, що складається із N_A пар відповідних КТ, де p_i^t – i-та КТ на I_t, а p_i^M – відповідна їй точка (об'єкта або фону) у M. M є сукупністю КТ об'єкта та фону, і при визначенні s_t та θ_t за (6, 7) обчислюються не по одному, а по два значення — обчисленні першого використовуються КТ об'єкта у M, а другого — КТ фону у цій моделі. Таким чином, під час обчислення не використовуються пари КТ у M, одна з яких належить об'єкту, а друга — фону, тому що довжина відрізка лінії з такими кінцевими точками може змінюватись в процесі руху об'єкта і не відповідати дійсному значенню масштабу. Після цього з обчислених двох значень як вихідне використовується те, що менш відрізняється від відповідного середнього значення, обчисленого на останніх кількох кадрах.

Було виявлено, що обчислення масштабу та кута повороту на основі відповідних КТ фону є більш точним порівняно з КТ об'єкта внаслідок того, що, по-перше, кількість КТ фону зазвичай значно більше, ніж об'єкта і, по-друге, КТ фону розташовані на більших відстанях одна від одної, що знижує вплив помилок у визначенні положення КТ на результат оцінки масштабу та кута повороту.

Масштаб та кут повороту, обчислені за формулами (6, 7), є близькими до дійсних за малих змін фону або частого оновлення та адаптації КТ фону до більших змін, наприклад, внаслідок розмиття. Якщо ж ці умови не виконуються, то для надійнішого визначення вказаних параметрів доцільніше використовувати такий евристичний алгоритм.

Результатом роботи алгоритму є оцінені значення s_t та θ_t або ж нульове значення s_t для свідчення того, що вхідні дані про відповідні КТ є недостатньо надійними і містять значну кількість помилок. На початковому етапі алгоритму формується зображення, яке є результатом об'єднання по вертикалі двох кадрів: I_{t1} і поточного I_t у відео (Рис. 2, 3), номера яких показано у лівому верхньому куті на цих кадрах. Далі формується сукупність L відрізків ліній, кожна з яких з'єднує КТ у моделі об'єкта на I_{t1} з відповідною КТ на I_t, що показані зліва червоним кольором на Рис. 2(a)–Рис. 3(a) Після цього обчислюються медіанні значення довжини та кута повороту відрізків ліній L і обчислюється кількість відрізків, що мають близькі довжину та кут повороту до обчислених медіанних значень і показані справа на Рис. 2 (b)–Рис. 3 (b).

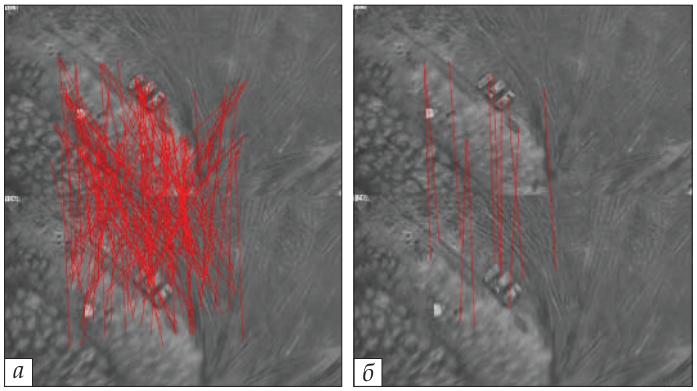

Рис. 2. Приклад визначених помилковими даних (*a*) відповідності КТ фону, яким відповідає мала кількість відібраних більш надійних пар відповідних КТ (*b*)

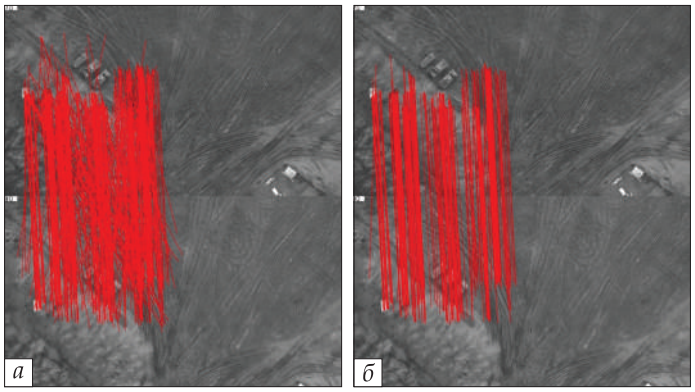

Рис. 3. Приклад вхідних даних (*a*) про відповідність КТ фону на порівнюваних кадрах 103 і 109, на основі яких визначається достатня кількість відібраних пар КТ (*b*) для подальшого обчислення масштабу та кута повороту на 109-му кадрі за медіанними значеннями

Якщо кількість цих порівняно надійних відрізків ліній є малою частиною від $|L|$, то алгоритм видає нульове значення s_t, що свідчить про порівняно малу надійність даних відповідності КТ на вході алгоритму (Рис. 2). Це означає, що поточний кадр у відео не може бути використано для оцінки масштабу внаслідок вчасно не виконаного оновлення моделі об'єкта або ж недоліків кадра, пов'язаних, наприклад, зі зміною умов освітлення або розмиттям через втрату фокусу чи швидкого пересування.

В іншому разі (Рис. 3) сукупність відібраних і порівняно надійних відрізків ліній використовується для обчислення медіанних значень масштабу та кута повороту за формулами, подібними до (6, 7). У наведеному прикладі на Рис. 3 кількість відрізків, відібраних і показаних на Рис. 3(*b*) менше половини числа $|L|$ вхідних відрізків на Рис. 2(*a*). Це означає, що обчислений масштаб за медіанним значенням відповідних вхідних даних на Рис. 3(*a*) відрізнятиметься від

дійсного значення на 109-му кадрі у відео з причини значної кількості помилок під час пошуку відповідності КТ на порівнюваних кадрах 103 і 109.

Оновлення моделі об'єкта у відео

Хай N_t^A — кількість КТ на I_t, відповідних до КТ об'єкта у M; N_T, N_R — загальні кількості КТ об'єкта у M і в межах R на I_t; $N_{KT} = N_T{*}k_T + N_R{*}k_R$ — зважена сума, де k_T, k_R можуть дорівнювати 0,6 і 0,4; $Thr_D = N_{KT}{*}k_D$, $Thr_U = N_{KT}{*}k_U$ — порогові значення для застосування в умовах детектування об'єкта і оновлення моделі M, де k_D, k_U за замовченням дорівнюють 0,15 і 0,3. Якщо $N_t^A > Thr_D$, то об'єкт вважається присутнім на поточному зображенні і обмежувальний прямокутник R може бути визначено шляхом застосування обчислених значень центру, масштабу та кута повороту об'єкта.

Якщо у відео є значні зміни у поданні об'єкта та фону на зображеннях, необхідно адаптувати до них M шляхом оновлення як кількості, так і дескрипторів КТ об'єкта та фону у цій моделі. Але ці зміни супроводжує ризик надлишкового пристосування до фону (дріфту), внаслідок чого може відбутись втрата об'єкта. Тому оновлення M відбувається за виконання порівняно надійних умов визначення положення об'єкта або його відсутності на зображеннях. Будемо відрізняти два способи оновлення M, після задіяння кожного з них на t-му кадрі $t_1 = t$.

Перший спосіб застосовується за виконання таких умов детектування об'єкта на кожному I_t: 1) $N_t^A > Thr_U \;\&\; s_t > 0$ і 2) не менше 75-и % R знаходиться на I_t. В процесі оновлення M формується із відповідних КТ об'єкта та фону, і якщо їх мало, то додатково з КТ, близьких за значеннями відстаней між дескрипторами до КТ об'єкта або фону в M у двох околах визначеного положення об'єкта.

Другий спосіб використовується, якщо об'єкт протягом певного часу відсутній на кадрах у відео. Впродовж цього часу на кожному I_t виконується пошук КТ фону, які є відповідними до КТ фону у M, і визначаються змінені значення масштабу і кута повороту відносно I_{t1}. На кожному I_t виконується також оновлення КТ фону в M (для надійнішого пошуку відповідних КТ фону на наступних зображеннях) за виконання таких умов: $N_t^A \le Thr_D \,||\, s_t = 0$. Для цього формується перелік з КТ на I_t, відповідних до КТ фону в M, і (якщо їх замало) додатково з КТ в межах ділянки на I_t, центр якої збігається з центром попереднього положення об'єкта, а розміри є тим більшими, чим довшим є час відсутності об'єкта. Далі після обчислення нових дескрипторів ці КТ використовуються як КТ фону в M. Після оновлення КТ фону додатково оновлюються також КТ об'єкта в M, якщо виконуються умови: $S_t > 0 \;\&\; (|s_t - s_{t1}| > thr_{sc} \,||\, |\theta_t - \theta_{t1}| > thr_\theta)$, де thr_{sc} і thr_θ — порогові значення. Це оновлення відбувається шляхом таких дій:

Рис. 4. Приклади детектування об'єкта під час відстеження у відео на кадрах: початковому 71-му (*а*), 283-му (*б*) перед подальшим тривалим зникненням (на кадрах з 284-го по 505-й) з поля зору і на 506-му (*в*) з появою об'єкта у відео із зменшеним у 6 разів масштабом

1) масштабування та поворот I_{t1} за масштабом та кутом повороту, визначених на I_t, і 2) детектування КТ в межах обмежувального об'єкт прямокутника на трансформованому зображенні I_{t1} з подальшим застосуванням їх як оновлених КТ об'єкта в моделі M.

На Рис. 4 показано приклади, коли визначають положення об'єкта в відео обмежувальним прямокутником, розробленими засобами, на 3х кадрах в одному з відео: оператор на 71-му (рис. 4 (*а*)) і програма на 283-му (рис. 4 (*б*)) перед тривалим (з 284-го по 505-й кадри) зникненням об'єкта з поля зору, і з його появою у відео на 506-му кадрі (рис. 4 (*в*)) зі зменшеним в 6 разів масштабом. На кожному з 71-го по 506-й кадрів обчислювались поточні значення масштабу та кута повороту об'єкта, у тому числі за відсутності об'єкта в полі зору камери, а також виконувалось оновлення моделі M на деяких з них за описаних умов, що забезпечило детектування об'єкта з його появою у відео на 506-му і наступних кадрах зі значно зміненими параметрами подання на зображеннях порівняно із 283-м кадром.

Висновки

У роботі використовується модель M об'єкта відстеження, що на відміну від [6] складається з КТ не тільки об'єкта, а і фону для пошуку відповідності КТ на кожному зображенні до M, і визначення змін масштабу та кута повороту об'єкта як за наявності на зображеннях, так і після його зникнення та подальшої появи у відео зі зміненими параметрами.

Розроблено алгоритм визначення змін масштабу та кута повороту об'єкта у відео на основі відповідних КТ об'єкта і фону за наявності значної кількості помилок у знайденій відповідності КТ на порівнюваних зображеннях. Запропоновано спосіб визначення масштабу та кута повороту зображень у відео за відсутності об'єкта на цих зображеннях з метою оновлення моделі об'єкта і його детектування після появи на зображеннях із суттєво зміненими значеннями цих параметрів.

ЛІТЕРАТУРА / REFERENCES

1. Bolme D.S., Beveridge J.R., Draper B.A., Lui Y.M.. Visual object tracking using adaptive correlation filters. *The IEEE conference on Computer Vision and Pattern,* 2010, 1–10. https://doi.org/10.1109/CVPR.2010.5539960
2. Henriques J.F., Caseiro R., Martins P., Batista J. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 2015, Vol. 37 (3), 583–596. https://doi.org/10.1109/TPAMI.2014.2345390
3. Danelljan M., Häger G., Khan F.S., Felsberg M. Discriminative scale space tracking. *IEEE transactions on pattern analysis and machine intelligence*, 2016, Vol. *39* (8), 1561–1575. https://doi.org/10.1109/TPAMI.2016.2609928
4. Tao R., Gavves E., Smeulders A.W. Siamese Instance Search for Tracking. *The IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 1420–1429. https://doi.org/10.1109/CVPR.2016.158
5. Bertinetto L., Valmadre J., Henriques J.F., Vedaldi A., Torr P.H. Fully-convolutional siamese networks for object tracking. *Computer vision–ECCV 2016 workshops,* Amsterdam, the Netherlands, 2016, 850–865. https://doi.org/10.1109/CVPR.2016.158
6. Nebehay G., Pflugfelder R.P. Clustering of static-adaptive correspondences for deformable object tracking. *The IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, 2784–2791. https://doi.org/10.1109/CVPR.2015.7298895
7. Wu B., Xie Y., Luo W. Robust and adaptive object tracking via correspondence clustering. *IEICE Trans. Information & Systems*, 2016, Vol. E99-D (10), 2664–2667. https://doi.org/10.1587/transinf. 2016EDL8065
8. Hong Z., Chen Z., Wang C., Mei X., Prokhorov D., Tao D. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. *The IEEE conference on computer vision and pattern recognition,* 2015, 749–758. https://doi.org/10.1109/CVPR.2015.7298675
9. Kalal Z., Mikolajczyk K., Matas J. Tracking-learning detection. *TPAMI*, 2012, Vol. 34 (7), 1409–1422. https://doi.org/10.1109/TPAMI.2011.239
10. Ma C., Yang X., Zhang C., Yang M.H. Long-term correlation tracking. *The IEEE conference on computer vision and pattern recognition,* 2015, 5388–5396. https://doi.org/10.1109/CVPR.2015.7299177
11. Ma C., Huang J.B., Yang X., Yang M.H. Adaptive correlation filters with long-term and short-term memory for object tracking. *International Journal of Computer Vision*, 2018, Vol. 126, 771–796. https://doi.org/10.1007/s11263-018-1076-4
12. Lukežič A., Zajc L.Č., Vojíř T., Matas J., Kristan M. FuCoLot–a fully-correlational long-term tracker. *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia,* 2018, Revised Selected Papers, Part II 14, 2019, 595–611. https://doi.org/10.1007/978-3-030-20890-5_38
13. Lukežič A., Vojíř, T., C̆ehovin Zajc, L., Matas, J., Kristan, M. Discriminative correlation filter with channel and spatial reliability. *Comp. Vis. Patt. Recognition,* 2017, 6309–6318. https://doi.org/10.1109/CVPR.2017.515
14. Yan B., Zhao H., Wang D., Lu H., Yang X. 'Skimming-Perusal' Tracking: a framework for Real-Time and robust Long-Term tracking. *IEEE/CVF International Conference on Computer Vision (ICCV),* 2019. https://doi.org/10.1109/iccv.2019.00247
15. Huang L., Zhao X., Huang K. GlobalTrack: a simple and strong baseline for Long-Term tracking. *AAAI Conference on Artificial Intelligence*, 2020, Vol. *34* (07), 11037–11044. https://doi.org/10.1609/aaai.v34i07.6758
16. Dai K., Zhang Y., Wang D., Li J., Lu H., Yang X. High-Performance Long-Term tracking with Meta-Updater. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 6297–6306. https://doi.org/10.1109/cvpr42600.2020.00633
17. Dunnhofer M., Micheloni C. CoCoLoT: Combining Complementary Trackers in Long-Term Visual Tracking. *26th International Conference on Pattern Recognition (ICPR)*, 2022, 5132–5139. https://doi.org/10.1109/ICPR56361.2022.9956082

18. Chen X., Yan B., Zhu J., Wang D., Yang X., Lu H. Transformer Tracking. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 8122–8131. https://doi.org/10.1109/CVPR46437.2021.00803
19. Swati, Kumar V.N., Dinesh Kawa S., Engineer P.J. An Efficient Object Tracking on Edge Devices with Quantized Siamese Networks. *Devices for Integrated Circuit (DevIC)*, 2025, 604–609. https://doi.org/10.1109/DevIC63749.2025.11012629
20. Kristan M., Matas J., Leonardis A., Vojír T., Pflugfelder R.P., Fernandez G.J., Nebeha, G., Porikli F.M., Cehovin L. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2015, Vol. 38, 2137–2155. https://doi.org/10.1109/TPAMI.2016.2516982
21. Dalal N., Triggs B. Histograms of oriented gradients for human detection. *Comp. Vis. Patt. Recognition,* 2005, Vol. 1, 886–893. https://doi.org/10.1109/CVPR.2005.177
22. Lowe D.G. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004, Vol. 60 (2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
23. Danelljan M., Hager G., Khan F.S., Felsberg M. Accurate Scale Estimation for Robust Visual Tracking. *BMVC,* 2014, 1–11. http://doi.org/10.5244/C.28.65
24. Li Y., Zhu J. A scale adaptive kernel correlation filter tracker with feature integration. *Proc. European Conf. Computer Vision*. 2014, 254–265. https://doi.org/10.1007/978-3-319-16181-5_18
25. Dolla´r P., Appel R., Belongie S., Perona P. Fast feature pyramids for object detection. *TPAMI*, 2014. https://doi.org/10.1109/TPAMI.2014.2300479
26. Krizhevsky A., Sutskever I., Hinton G.. ImageNet classification with deep convolutional neural networks. *NIPS*, 2012, 84–90. https://doi.org/10.1145/3065386
27. Shrivastava A., Gupta A., Girshick R. Training region- based object detectors with online hard example mining. *IEEE conference on Computer Vision and Pattern Recognition*, 2016, 761–769. https://doi.org/10.1109/CVPR.2016.89
28. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015, 91–99.
29. Lin T.-Y. *et al.* Feature pyramid networks for object detection. *IEEE conference on Computer Vision and Pattern Recognition*, 2017, 2117–2125. https://doi.org/10.1109/CVPR.2017.106
30. Zhou S., Zhou H., Qian L. A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images. *Sci Rep,* 2025, Vol. 15, Article 9255. https://doi.org/10.1038/s41598-025-92344-7
31. Leutenegger S., Chli M., Siegwart R.Y. BRISK: Binary robust invariant scalable keypoints. In *ICCV*, 2011. https://doi.org/10.1109/ICCV.2011.6126542
32. Mikolajczyk K., Schmid C. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2005, Vol. 27, 1115–1125 1615–1630. https://doi.org/10.1109/TPAMI.2005.188
33. Lucas B.D., Kanade T. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981, 674–679. URL: https://www.ijcai.org/Proceedings/81-2/Papers/017.pdf [Accessed 03 Oct. 2025]

Отримано / Received 30.10.2025

M. KYYKO, PhD (Engineering), Senior Researcher, Leading Researcher,
Institute of Information Technologies and Systems of the NAS of Ukraine,
40, Hlushkova Akad. ave., Kyiv, 03187, Ukraine
https://orcid.org/0009-0005-6605-0339
vkiiko@gmail.com

DETERMINING THE SCALE AND ROTATION ANGLE
FOR LONG-TERM OBJECT TRACKING IN VIDEO

Introduction. The task of tracking is to determine the position of an arbitrary object (target) in a video after detection in the initial frame. Algorithms for performing tracking are divided into those that provide short-term or more complex long-term

tracking in a video. A key problem in long-term tracking is the recovery of a target after a period of absence or tracking failures, because during this time not only the coordinates but also the scale and angle of rotation of the target can change significantly, knowledge of which increases the accuracy and reliability of detection. As a result, after a long disappearance, the search for the target must be performed not locally, but within the entire image and significant intervals of possible changes in scale and rotation angle. The reliability of tracking in video largely depends on the efficiency (accuracy and low computational complexity) of the algorithms used to determine the scale and rotation angle of the tracked target in the images. There are known algorithms that determine the scale and rotation angle based on the correspondence of key points (KPs) of the target without sufficient consideration of the background KP, and can provide tracking in conditions of only a short-term absence of the target, during which the scale and rotation angle change little.

Purpose of the research is to develop an algorithm for determining the scale and angle of rotation of an object, which overcomes these shortcomings to obtain more reliable object tracking results in difficult conditions.

Methods for searching key points and determining their correspondence in images were used.

Results. An algorithm is proposed for estimation the scale and rotation angle of the tracked object in the images based on finding corresponding KPs in each frame to the KPs in the object model *M*. The algorithm can be mainly used in conditions where changes in scale and angle of rotation are mainly a consequence of changes in camera movement or operator actions. These changes are largely correlated with changes in the background, which usually corresponds to video surveillance from an aircraft, in particular a UAV. The advantages of the algorithm are that it is relatively more resistant to errors in determining the corresponding KP pairs, and can also be used during the prolonged absence of an object in the video to estimate the scale and angle of rotation.

Conclusions. The paper uses a tracking object model consisting of KPs in the object and the background to search for KPs correspondence. The algorithm is proposed for determining the scale and rotation angle of the object both when present and absent in images to update *M* and detect this object after it appears in images with significantly changed parameters. Examples of using the algorithm for long-term tracking with the proposed criterion for the presence of the object, as well as two methods of updating *M*, when it is present or absent in images, are given.

***Keywords**: object tracking in video, BRISK key points, KCF tracking algorithm, HOG features, scale and rotation angle of an object in an image, object detection and recognition.*