МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ "ЛЬВІВСЬКА ПОЛІТЕХНІКА"

ЛЕКЦІЯ 5. Застосування машинного навчання в комп'ютерному зорі

Львів -- 2025

Лекція зі штучного інтелекту 2025-05

Вступ

Комп'ютерний зір (Computer Vision) — це галузь штучного інтелекту, яка займається розробкою методів і алгоритмів для автоматичного аналізу та інтерпретації візуальної інформації. За останні десятиліття ця галузь пройшла значний шлях розвитку: від простих алгоритмів обробки зображень до складних нейромережевих архітектур, здатних розв'язувати широкий спектр задач.

У цій лекції ми розглянемо основні задачі комп'ютерного зору, класичні та сучасні підходи до їх вирішення, з особливим фокусом на задачах зіставлення зображень (image matching) та відновлення тривимірної структури з руху (structure-from-motion). Ми також простежимо еволюцію архітектур нейронних мереж для задач комп'ютерного зору та розглянемо причини їх ефективності.

Основні задачі комп'ютерного зору

Комп'ютерний зір охоплює широкий спектр задач, серед яких:

1. **Класифікація зображень** — визначення категорії об'єкта на зображенні
2. **Виявлення об'єктів** (Object Detection) — знаходження та класифікація об'єктів на зображенні
3. **Сегментація зображень** — розділення зображення на смислові області
4. **Розпізнавання облич** — ідентифікація та верифікація осіб
5. **Оцінка пози** (Pose Estimation) — визначення положення людського тіла
6. **Зіставлення зображень** (Image Matching) — знаходження відповідностей між точками на різних зображеннях
7. **Відновлення 3D-структури з 2D-зображень** (Structure-from-Motion, SfM) — реконструкція тривимірної сцени з набору зображень
8. **Локалізація камери** (Camera Localization) — визначення положення камери у просторі
9. **Оптичний потік** (Optical Flow) — визначення руху об'єктів між кадрами
10. **Генерація та синтез зображень** — створення нових зображень

Класичні підходи до задач комп'ютерного зору

Виділення ознак (Feature Extraction)

Класичні методи комп'ютерного зору базуються на виділенні інформативних ознак із зображень. Ці ознаки повинні бути:

- **Інваріантними** до змін масштабу, повороту, освітлення
- **Відмінними** для різних об'єктів
- **Стабільними** при невеликих змінах зображення
- **Локальними** для стійкості до оклюзій

Популярні дескриптори ознак:

1. **SIFT** (Scale-Invariant Feature Transform):
    - Виявляє ключові точки, інваріантні до масштабу та обертання
    - Створює 128-вимірний дескриптор для кожної точки
    - Використовує різницю гаусіанів (DoG) для виявлення ключових точок
2. **SURF** (Speeded-Up Robust Features):
    - Швидша альтернатива SIFT
    - Використовує інтегральні зображення та вейвлети Хаара
    - Створює 64-вимірний дескриптор
3. **ORB** (Oriented FAST and Rotated BRIEF):
    - Поєднує детектор FAST та дескриптор BRIEF
    - Обчислювально ефективний, підходить для мобільних пристроїв
    - Бінарний дескриптор, що дозволяє швидке порівняння
4. **HOG** (Histogram of Oriented Gradients):
    - Обчислює гістограми напрямків градієнтів у локальних областях
    - Ефективний для розпізнавання форм об'єктів
    - Широко використовується для виявлення пішоходів

Зіставлення зображень (Image Matching)

Зіставлення зображень — це процес знаходження відповідностей між точками на різних зображеннях однієї сцени. Цей процес є фундаментальним для багатьох задач, включаючи панорамне зшивання, відстеження об'єктів та відновлення 3D-структури.

Класичний конвеєр зіставлення зображень:

1. **Виявлення ключових точок** на обох зображеннях (використовуючи SIFT, SURF, ORB тощо)
2. **Обчислення дескрипторів** для кожної ключової точки
3. **Зіставлення дескрипторів** для знаходження потенційних відповідностей:
    - Метод найближчого сусіда (Nearest Neighbor)
    - Метод співвідношення відстаней (Ratio Test) для фільтрації ненадійних відповідностей
4. **Геометрична верифікація** для видалення хибних відповідностей:
    - RANSAC (Random Sample Consensus) для оцінки фундаментальної або гомографічної матриці

Structure-from-Motion (SfM)

Structure-from-Motion — це процес відновлення тривимірної структури сцени та положень камер з набору двовимірних зображень. SfM широко застосовується в фотограмметрії, картографії, доповненій реальності та робототехніці.

Основні етапи класичного SfM:

1. **Виявлення та зіставлення ознак** між усіма парами зображень
2. **Оцінка відносного положення камер** для пар зображень:
    - Обчислення фундаментальної матриці F або істотної матриці E
    - Декомпозиція E для отримання обертання R та переносу t
3. **Триангуляція** для відновлення 3D-координат точок
4. **Зв'язування пар зображень** у глобальну модель:
    - Послідовний SfM: поступове додавання зображень
    - Глобальний SfM: одночасна оптимізація всіх положень камер
5. **Bundle Adjustment** — нелінійна оптимізація для мінімізації помилки перепроєкції:
    - Мінімізація суми квадратів відстаней між спостережуваними та проєктованими точками
    - Використання алгоритму Левенберга-Марквардта для оптимізації

Математична основа SfM:

Для пари зображень з каліброваними камерами, зв'язок між відповідними точками x і x' описується істотною матрицею E:

$$x'^T E x = 0$$

де E = [t]_× R, [t]_× — кососиметрична матриця, що відповідає вектору переносу t, а R — матриця обертання.

Для некаліброваних камер використовується фундаментальна матриця F:

$$x'^T F x = 0$$

де F = K'^{-T} E K^{-1}, K і K' — матриці внутрішніх параметрів камер.

Локалізація камери

Локалізація камери — це задача визначення положення та орієнтації камери відносно відомої сцени або карти. Ця задача критично важлива для систем доповненої реальності, автономної навігації та робототехніки.

Класичні підходи до локалізації камери:

1. **Perspective-n-Point (PnP)** — визначення положення камери за n відповідностями між 3D-точками та їх 2D-проєкціями:
    - P3P (n=3): мінімальний випадок, дає до 4 можливих рішень
    - EPnP: ефективний алгоритм для довільної кількості точок
    - RANSAC-PnP: робастна версія для даних з викидами
2. **Візуальна одометрія** — послідовне відстеження положення камери:
    - Відстеження ознак між послідовними кадрами
    - Оцінка відносного руху між кадрами
    - Інтеграція руху для отримання траєкторії

3. **Relocalization** — повторна локалізація після втрати відстеження:
    - Пошук зображення в базі даних (image retrieval)
    - Зіставлення ознак з 3D-моделлю
    - Оцінка положення за допомогою PnP

Глибоке навчання в комп'ютерному зорі

Революція глибокого навчання кардинально змінила підходи до вирішення задач комп'ютерного зору. Замість ручного проєктування ознак, глибокі нейронні мережі автоматично вивчають ієрархічні представлення безпосередньо з даних.

Еволюція архітектур CNN для комп'ютерного зору

LeNet (1998)

- Перша успішна архітектура CNN, розроблена Яном ЛеКуном
- Використовувалася для розпізнавання рукописних цифр
- Складалася з двох згорткових шарів та трьох повнозв'язних шарів

AlexNet (2012)

- Переможець змагання ImageNet 2012, що започаткував революцію глибокого навчання
- Глибша архітектура з 5 згортковими та 3 повнозв'язними шарами
- Використовувала ReLU, dropout та нормалізацію локальної відповіді
- Навчалася на двох GPU

VGGNet (2014)

- Проста, але глибока архітектура з 16-19 шарами
- Використовувала малі фільтри 3×3 замість більших
- Продемонструвала важливість глибини для якості розпізнавання

GoogLeNet/Inception (2014)

- Введення модуля Inception з паралельними згортками різних розмірів
- Ефективне використання обчислювальних ресурсів
- Глибока архітектура з 22 шарами

ResNet (2015)

- Революційна архітектура з залишковими зв'язками (skip connections)
- Дозволила навчати надглибокі мережі (до 152 шарів і більше)
- Вирішила проблему зникаючого градієнта
- Залишкові блоки: F(x) + x замість F(x)

DenseNet (2017)

- Кожен шар отримує входи від всіх попередніх шарів
- Краще поширення градієнтів та повторне використання ознак
- Менша кількість параметрів порівняно з ResNet

EfficientNet (2019)

- Систематичний підхід до масштабування CNN
- Балансування глибини, ширини та роздільної здатності
- Досягнення найкращої точності при меншій кількості параметрів

Сучасні підходи до зіставлення зображень

Глибоке навчання трансформувало підходи до зіставлення зображень, замінивши ручно спроєктовані дескриптори навченими представленнями.

Навчені дескриптори ознак:

1. **LIFT** (Learned Invariant Feature Transform):
    - Повністю навчений конвеєр виявлення, орієнтації та опису ключових точок
    - Навчається на відповідностях, отриманих з 3D-реконструкції
2. **SuperPoint**:
    - Самонавчання на синтетичних даних
    - Одночасне виявлення ключових точок та обчислення дескрипторів
    - Використовує повністю згорткову архітектуру
3. **D2-Net**:
    - Виявлення та опис з одного CNN представлення
    - Адаптивне виявлення ознак на різних рівнях деталізації
    - Ефективний для складних сцен з великими змінами точки зору
4. **R2D2**:
    - Спільне навчання надійності та відмінності ознак
    - Щільне передбачення ключових точок та дескрипторів
    - Оптимізований для точної локалізації

Навчені методи зіставлення:

1. **SuperGlue**:
    - Використовує графові нейронні мережі та механізм уваги (attention)
    - Розглядає зіставлення як задачу оптимального транспортування
    - Враховує як схожість дескрипторів, так і просторову узгодженість
2. **LoFTR** (Local Feature TRansformer):

- Використовує трансформери для встановлення відповідностей
- Працює з щільними ознаками замість розріджених ключових точок
- Ефективний для сцен з малою текстурою

3. **COTR** (Correspondence Transformer):
   - Трансформерна архітектура для точного зіставлення
   - Запитує відповідності для конкретних точок
   - Підходить для щільного зіставлення

Глибоке навчання для Structure-from-Motion

Глибоке навчання вносить значні покращення в різні компоненти конвеєру SfM.

Навчені підходи до SfM:

1. **BA-Net** (Bundle Adjustment Network):
   - Диференційований шар Bundle Adjustment
   - Навчається передбачати глибину та положення камери
   - Поєднує класичну оптимізацію з глибоким навчанням
2. **DeepSfM**:
   - Ітеративне уточнення глибини та положення камери
   - Використовує згорткові рекурентні мережі
   - Робастний до шуму та оклюзій
3. **DROID-SLAM**:
   - Щільна рекурентна одометрія та SLAM з глибоким навчанням
   - Відстежує всі пікселі замість розріджених ключових точок
   - Досягає точності, порівнянної з класичними методами

Навчені підходи до локалізації камери:

1. **PoseNet**:
   - Пряме регресійне передбачення положення та орієнтації камери
   - Не вимагає явного зіставлення ознак
   - Швидкий, але менш точний порівняно з геометричними методами
2. **MapNet**:
   - Використовує геометричні обмеження для покращення точності
   - Навчається з додатковими втратами для часової узгодженості
   - Поєднує навчання з візуальною одометрією
3. **Hierarchical Localization**:

- Поєднує пошук зображень з локальним зіставленням ознак
- Використовує навчені глобальні та локальні дескриптори
- Масштабується до великих середовищ

Трансформери в комп'ютерному зорі

Архітектура трансформерів, яка революціонізувала обробку природної мови, знаходить все більше застосувань у комп'ютерному зорі.

Vision Transformer (ViT):

- Розділяє зображення на патчі та обробляє їх як токени
- Використовує механізм самоуваги для моделювання глобальних залежностей
- Досягає найкращих результатів у класифікації зображень при навчанні на великих наборах даних

DETR (DEtection TRansformer):

- Формулює виявлення об'єктів як задачу прямого набору передбачень
- Усуває потребу в ручному проєктуванні компонентів, таких як NMS
- Використовує двонаправлений трансформер-енкодер-декодер

Swin Transformer:

- Ієрархічна архітектура з вікнами зсувної уваги
- Ефективно обробляє зображення високої роздільної здатності
- Універсальна основа для різних задач комп'ютерного зору

Нейромережеві архітектури для специфічних задач комп'ютерного зору

Сегментація зображень:

- **U-Net**: Архітектура енкодер-декодер зі skip-з'єднаннями для збереження просторової інформації
- **DeepLab**: Використовує атрофовані (dilated) згортки для збільшення рецептивного поля
- **Mask R-CNN**: Розширення Faster R-CNN для сегментації екземплярів

Оцінка глибини:

- **MonoDepth**: Навчання без нагляду з використанням стереопар
- **DenseDepth**: Щільна оцінка глибини з використанням енкодера-декодера
- **BTS (Big to Small)**: Локальне планарне керівництво для точної оцінки глибини

Оцінка пози людини:

- **OpenPose**: Виявлення ключових точок з використанням полів частин
- **HRNet**: Підтримка високої роздільної здатності протягом всієї мережі
- **VIBE**: Оцінка 3D-пози та форми людини з відео

Сучасні тенденції та майбутні напрямки

Самонавчання (Self-supervised Learning)

Самонавчання дозволяє моделям вчитися з нерозмічених даних, що особливо важливо для комп'ютерного зору, де розмічені дані часто обмежені.

Популярні підходи до самонавчання:

- **Контрастивне навчання** (SimCLR, MoCo): Максимізація схожості між різними представленнями одного зображення
- **Маскування та відновлення** (MAE, BEiT): Маскування частин зображення та їх відновлення
- **DINO**: Самодистиляція з відсутністю негативних пар

Нейромережеві архітектури для 3D-даних

Обробка 3D-даних стає все важливішою для комп'ютерного зору, особливо для автономних транспортних засобів та доповненої реальності.

Основні архітектури для 3D-даних:

- **PointNet/PointNet++**: Обробка невпорядкованих наборів 3D-точок
- **3D CNN**: Розширення CNN для воксельних представлень
- **Graph Neural Networks**: Обробка 3D-даних як графів
- **Neural Radiance Fields (NeRF)**: Представлення 3D-сцен як неявних нейронних полів

Нейронні поля для представлення 3D-сцен

Нейронні поля — це новий підхід до представлення 3D-сцен, який використовує нейронні мережі для моделювання неявних функцій.

Neural Radiance Fields (NeRF):

- Представляє сцену як неперервну функцію, що відображає 3D-координати та напрямки погляду в колір та щільність
- Дозволяє синтезувати фотореалістичні види з нових ракурсів
- Навчається лише з набору 2D-зображень

Розширення NeRF:

- **Instant NGP**: Прискорення навчання та рендерингу NeRF
- **NeRF in the Wild**: Робастність до змін освітлення та інших варіацій
- **SLAM-NeRF**: Поєднання SLAM з NeRF для одночасної локалізації та картографування

Мультимодальне навчання

Поєднання різних модальностей (зображення, текст, аудіо) відкриває нові можливості для комп'ютерного зору.

Приклади мультимодальних моделей:

- **CLIP** (Contrastive Language-Image Pretraining): Навчання на парах зображення-текст для розуміння візуальних концепцій
- **DALL-E/Stable Diffusion**: Генерація зображень з текстових описів
- **ImageBind**: Єдине представлення для різних модальностей

Практичні аспекти та інструменти

Бібліотеки та фреймворки для комп'ютерного зору:

1. **OpenCV**:
    - Класична бібліотека з широким набором функцій
    - Підтримка як традиційних алгоритмів, так і глибокого навчання
    - Оптимізована для продуктивності
2. **PyTorch/TensorFlow**:
    - Основні фреймворки для глибокого навчання
    - Підтримка GPU-прискорення
    - Багатий екосистем інструментів
3. **COLMAP**:
    - Сучасна реалізація Structure-from-Motion
    - Повний конвеєр від зіставлення зображень до щільної реконструкції
    - Відкритий код та активна спільнота
4. **Kornia**:
    - Диференційована комп'ютерна графіка та комп'ютерний зір
    - Інтеграція з PyTorch
    - Реалізація класичних алгоритмів з підтримкою автоматичного диференціювання

Набори даних для комп'ютерного зору:

1. **ImageNet**:
    - Великий набір даних для класифікації зображень
    - Понад 14 мільйонів розмічених зображень
    - Стандарт для оцінки моделей класифікації
2. **COCO** (Common Objects in Context):
    - Набір даних для виявлення об'єктів, сегментації та підписів
    - Складні сцени з кількома об'єктами
    - Детальні анотації

3. **MegaDepth**:
    - Великомасштабний набір даних для оцінки глибини
    - Отриманий з реконструкцій інтернет-фотографій
    - Різноманітні сцени та умови
4. **ScanNet**:
    - Набір даних RGB-D сканувань внутрішніх приміщень
    - Анотації для сегментації та 3D-реконструкції
    - Реалістичні сцени для навчання та оцінки

Висновки

Комп'ютерний зір пройшов значний шлях розвитку від класичних алгоритмів обробки зображень до сучасних нейромережевих підходів. Особливо вражаючим є прогрес у задачах зіставлення зображень та відновлення 3D-структури з руху, які є фундаментальними для багатьох практичних застосувань.

Класичні підходи, засновані на ручно спроєктованих ознаках та геометричних алгоритмах, заклали міцний теоретичний фундамент для галузі. Вони продовжують відігравати важливу роль, особливо в задачах, де потрібна інтерпретованість та математична строгість.

Глибоке навчання революціонізувало комп'ютерний зір, дозволивши автоматично вивчати ієрархічні представлення безпосередньо з даних. Еволюція архітектур CNN від LeNet до ResNet та трансформерів демонструє постійне вдосконалення здатності моделей захоплювати складні візуальні патерни.

Сучасні підходи до зіставлення зображень та SfM поєднують сильні сторони класичних методів та глибокого навчання, досягаючи безпрецедентної точності та робастності. Нейронні поля, такі як NeRF, відкривають нові можливості для представлення та рендерингу 3D-сцен.

Майбутнє комп'ютерного зору лежить у подальшій інтеграції з іншими модальностями, розвитку самонавчання та створенні все більш універсальних моделей, здатних розуміти візуальний світ на рівні, порівнянному з людським.

Розвиток комп'ютерного зору має величезний потенціал для трансформації багатьох галузей, від автономних транспортних засобів та робототехніки до медичної діагностики та доповненої реальності, відкриваючи нові можливості для взаємодії людей з технологіями та навколишнім світом.