Lecture 13: Learning on Videos, 3D Deep Learning and Scene Graphs

Areej Alasiry

CIS 6217 – Computer Vision for Data Representation

College of Computer Science, King Khalid University

Outline

1. Explain visual relationship modeling (object–object & object–scene relations).
2. Understand how Graph Neural Networks encode structured visual information.
3. Describe motion detection and multi-object tracking pipelines.
4. Infer human actions from image sequences using temporal deep learning models.

Visual Relationships

- Visual relationships describe **how objects interact** in a scene. Examples:
- *person riding horse*
- *dog next to table*
- *car under bridge*
- They form **triplets**:
- (subject, *predicate, object)* $= (s, p, o)$

Justification

- Move beyond object detection → **scene understanding**
- Enable:
  - Image captioning
  - Visual question answering (VQA)
  - Robotics reasoning
  - Surveillance & activity recognition
- Supports **commonsense reasoning** (e.g., “cup on table” implies support).

Modelling Visual Relationships

- **CNN-based object detection** → extract object features
- **Predicate classification** (interaction between pairs)
- **Graph-based reasoning** (scene graph generation)
- Key idea: Combine **appearance + spatial layout + context**.

GNNs for Visual Reasoning

- **Graph Neural Networks** process **nodes** (objects) and **edges** (relationships).
  Useful for scenes represented as **Scene Graphs**.
- Example:
  Nodes: person, bike
  Edges: riding, next-to

Why GNNs in CV?

- Encode **structured relationships**
- Model **contextual reasoning** across objects
- Improve tasks like:
  - Scene graph generation
  - Visual question answering
  - Relationship detection
  - Human interaction understanding

Motion Detection

- **Goal:** Identify pixels or regions with motion across frames.
- Common methods:
- **Frame differencing**
- **Background subtraction**
- **Optical flow (Horn–Schunck, Lucas–Kanade)**
- **CNN-based motion segmentation**

Optical Flow

- Estimates pixel-level motion vectors between frames. Applications:
- Video stabilization
- Action recognition
- Object tracking
- Autonomous driving

Object Tracking

- **Definition:** Maintain object identity across frames.
- Two major types:
- **Single-object tracking (SOT):** Track one target.
- **Multiple-object tracking (MOT):** Track many objects with IDs.
- Pipeline:
- Detection (CNN-based: YOLO, Faster R-CNN)
- Tracking (Kalman filters, SORT, DeepSORT)
- Data association (matching detections to past tracks)

Activity Recognition

- Recognizing **actions** or **activities** from a sequence of images or video.
  Examples:
  - Running
  - Jumping
  - Cooking
  - Fighting
  - Playing sports

Challanges

- Temporal dependencies
- Variations in viewpoint, scale, occlusion
- Multi-person interactions
- Complex, long-duration activities

Techniques for Activity Inference

- **1. CNN + LSTM Models**
  - CNN extracts frame-level features
  - LSTM/RNN models capture temporal sequences
- **2. 3D CNNs (e.g., C3D, I3D)**
  - Convolution in space + time
  - Strong temporal modeling
- **3. Transformers for Video**
  - Spatiotemporal attention
  - State-of-the-art (Video Swin Transformer, TimeSformer)
- **4. Pose-based Activity Recognition**
  - Keypoints (skeleton) tracked across time
  - Useful for sports, gestures, safety systems

Architecture - Example

- Input video frames
- CNN feature extraction (ResNet, EfficientNet)
- Sequence modeling (LSTM / GRU / Transformer)
- Activity classification layer

References

- Guide to CNNs for CV – Khan et al. (2018)
- Deep Learning with Python – Chollet (2018)
- Deep Learning in Computer Vision – Awad & Hassaballah (2020)

- Deep Learning for Vision Systems by Mohamed Elgendy (2020)