Caesar: Cross-Camera Complex Activity Recognition

Xiaochen Liu
Pradipta Ghosh
Oytun Ulutan
B.S. Manjunath
Kevin Chan
Ramesh Govindan

University of Southern California
University of California, Santa Barbara
ARL
University of Southern California
Cameras are the Most Powerful and Ubiquitous Sensors Today
Cameras Can be Used to Detect Various Activities

- Public safety
- Traffic monitoring
- Smart shopping
Today’s Activity Analytics is Highly Manual
Ideal: Automated Activity Analytics

Camera Cluster

Monitoring System

User-defined activity
“A person gets into a car then walk with bag”

Detected frames and objects
No prior work on complex activity detection

But there is work on atomic activity detection

Caesar fills the gap

- Gets into a car
- Walk
- With bag

User-defined activity
“A person gets into a car then walks with bag”
Atomic Activity & Complex Activity

Atomic Activities
- Use Phone (Camera 1)
- Take Bag (Camera 1)
- Give Bag (Camera 2)

Complex Activity
“A person uses phone then takes bag and gives bag”
Prior Work on Atomic Activity Detection

**Object Detection:**
Detect people/objects as bounding boxes

**Prior Work:**
YOLO (CVPR’17), SSD (ECCV’16), ResNet (CVPR’16)
Prior Work on Atomic Activity Detection

Object Detector

Bounding Boxes

Tracker

Tubes

Tracking:
Track each bounding box’s movement as a tube

Prior Work:
DeepSORT (WACV’18), RPP (ECCV’18), DG-Net (CVPR’19)
Prior Work on Atomic Activity Detection

Object Detector → Tracker → Tubes → Action Detector

Action Detection
Detect atomic actions in a tube using DNN or rules

Prior Work:
ACAM (WACV’20), RPNN (ICCV’19), ODAS (ECCV’18)

Bounding Boxes
DNN-Based Action Detection

Input

Result:
Use Phone

Tube clips with labels

Use Phone
Talking
Reading

Images from AVA dataset

Training
Pros and Cons of DNN for Action Detection

Can detect various atomic activities with high accuracy

No training data for complex activities, which have multi-tubes and long duration

Even with data, training a separate DNN for each complex activity is an overkill and not generalizable
Rule-Based Activity Detection

Person 1 Tube

Person 2 Tube

Simple Rule:
The pixel distance between two tubes keeps decreasing

Results:
Two person getting close
Pros and Cons of Simple Rule-Based Action Detection

Intuitive and simple input (just bounding box)

Generalizable to describe a wide range of interactions between tubes

No clue about the content in the box so can only detect very limited types of tube actions
Caesar: An Framework with DNNs + Rules

Person 2 Tube

Define many more complex activities than before

Caesar’s Rule: “A person skateboarding moves with a person riding a bike”

From DNNs
From simple rules
Flexibility: Detects user-defined complex activities correctly

How to specify and detect complex activities?
Design Goals & Challenges

**Flexibility:** Detects user-defined complex activities correctly

How to specify and detect complex activities?

**Scalability:** Supports concurrent video streams in realtime

How to optimize the hardware resource usage?
Design Goals & Challenges

Flexibility: Detects user-defined complex activities correctly

- How to specify and detect complex activities?

Scalability: Supports concurrent video streams in realtime

- How to optimize the hardware resource usage?

Efficiency: Saves energy and wireless data on mobile platforms

- How to collaborate the server and the mobile efficiently?
Caesar’s Contributions

**Flexibility:** Specify complex activities as spatio-temporal relationships over an extensible vocabulary

**Scalability:** Use lazy graph matching to reduce DNN usage on GPUs

**Efficiency:** Lazily retrieve frames from mobile devices
Represent Complex Activities as Graphs

Atomic Actions:
- From Camera X
- Move/Stop
- Talking
- Use Phone

Caesar uses different approaches’ atomic actions as graph elements

Single-camera rule-based action detection

Cross-camera person Re-identification DNN

Single-camera action detection DNN
Represent Complex Activities as Graphs

Atomic Actions:
- From Camera X
- Move/Stop
- Talking
- Use Phone

Spatial Relation:
- Near
- Approach
- Leave
- Close

Logic Relation:
- And
- Not
- Then

Caesar also uses pre-defined spatial and logic relations to express rules
Represent Complex Activities as Graphs

Atomic Actions:
- From Camera X
- Move/Stop
- Talking
- Use Phone
- ...

Spatial Relation:
- Near
- Approach
- Leave
- Close
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And

Atomic Actions:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Spatial Relation:
- Close
- Move/Stop
- Near
- Use Phone
- ...

Logic Relation:
- And
- Not
- Then
- ...

An Example of Activity Graph:
- Person:
  - Use Phone
  - Move

  - With Bag
  - Then
  - With Bag
  - And
Workflow: Rule Definition and Parsing

**Action Name:**
use_phone_then_move_with_bag

**Subjects:**
Person p

**Action Definition:**
(p use_phone) and (not p with_bag) then (p move) and (p with_bag)

(1) User inputs action definition

(2) Parse definition into a graph
(3) Matching the graph to the tubes and actions
Caesar’s Contributions

**Flexibility:** Specify complex activities as spatio-temporal relationships over an extensible vocabulary

**Scalability:** Use lazy graph matching to reduce DNN usage on GPUs

**Efficiency:** Lazily retrieve frames from mobile devices
The Action DNN is the Major Scalability Bottleneck

Due to the nature of per-tube detection, the action DNN becomes the bottleneck in the processing pipeline

<table>
<thead>
<tr>
<th>DNN</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Detection</td>
<td>20~30</td>
</tr>
<tr>
<td>Tracking/ReID</td>
<td>30~45</td>
</tr>
<tr>
<td>Action Detection</td>
<td>50~60 (per tube)</td>
</tr>
</tbody>
</table>

Too slow when
- 10 people: 5~6 FPS
- 20 people: 2~3 FPS
Lazy Graph Matching

Match spatiotemporal actions in a graph before any DNN action matching

Simple-rule actions:
- stop, move, close, disappear...

DNN actions:
- talk, use phone, sit, hug...

Skip heavy nodes and try to match the rest of the graph

Complete matching
An Example of Lazy Graph Matching

Use Phone → Move

Then

And

Use Phone → Move

Then

And

Action DNN Usage (with lazy matching)

Go back and run DNN

Action DNN Usage (without lazy matching)
Caesar’s Contributions

**Flexibility:** Specify complex activities as spatio-temporal relationships over an extensible vocabulary

**Scalability:** Use lazy graph matching to reduce DNN usage on GPUs

**Efficiency:** Lazily retrieve frames from mobile devices
Different modules have different data requirements:

- The spatial-temporal action detection
  - Needs bounding box positions in every frame

- The DNN action detection and tracking/ReID
  - Needs bounding box images of some tubes

- Visualization and video storage
  - Needs full frames
Caesar Mobile uploads bounding boxes by default, and uploads requested bounding box images or full frames.
System Architecture

Caesar Mobile

Edge Server

Caesar Server

Fixed Camera

Caesar Mobile

http://
Evaluation Setup

Caesar Mobile: Eight Nvidia TX1/TX2 boards, connected with WiFi

Caesar Server: Equipped with three Nvidia 2080 GPUs

Dataset: 30-min video clips from eight cameras in DukeMTMC dataset

- We manually labeled these in the videos:
  - 1289 atomic activities
  - 149 complex activities (73.8% are cross-camera)

Accuracy Metric: The ground truth is matched only when both the person ID and all atomic actions are matched
Caesar achieves 61.0% recall and 59.5% precision.

All errors come from imperfect DNNs:

- Miss Detection: “Missed the bicycle”
- Wrong Tracking/ReID: “Same person different IDs”
- Wrong Action: “Missed talking and use phone”

Better action and ReID DNNs could improve the accuracy to 100%.
- **Strawman**: without lazy graph matching and on-demand data fetching
- For **spatial-only activities** (*only non-DNN nodes*), Caesar has FPS > 17
- For mixed activities (*contain some DNN nodes*), Caesar keeps FPS > 15
On-demand data fetching reduces
- 3x wireless data usage for mixed activities; 15x for spatial-only activities
- 6x energy consumption for mixed activities; 10x for spatial-only activities
Cross-camera complex activity recognition is essential in many fields

Caesar combines DNNs and rule-based metrics to detect customized complex activities across cameras

Caesar leverages lazy matching and on-demand data retrieval to improve efficiency and scalability

Caesar reduces data usage and detection latency by >3X, up to 10X
Caesar: Cross-Camera Complex Activity Recognition

Xiaochen Liu
Pradipta Ghosh
Oytun Ulutan
B.S. Manjunath
Kevin Chan
Ramesh Govindan

University of Southern California
University of California, Santa Barbara
ARL

https://github.com/USC-NSL/Caesar