6.808 Mobile and Sensor Computing

aka IoT Systems

Spring 22 Lecture #11

Split Computing / Continuous Object Recognition

Glimpse

Continuous, Real-Time Object Recognition on Mobile Devices

Tiffany Chen

Lenin Ravindranath

Shuo Deng

Victor Bahl

Hari Balakrishnan

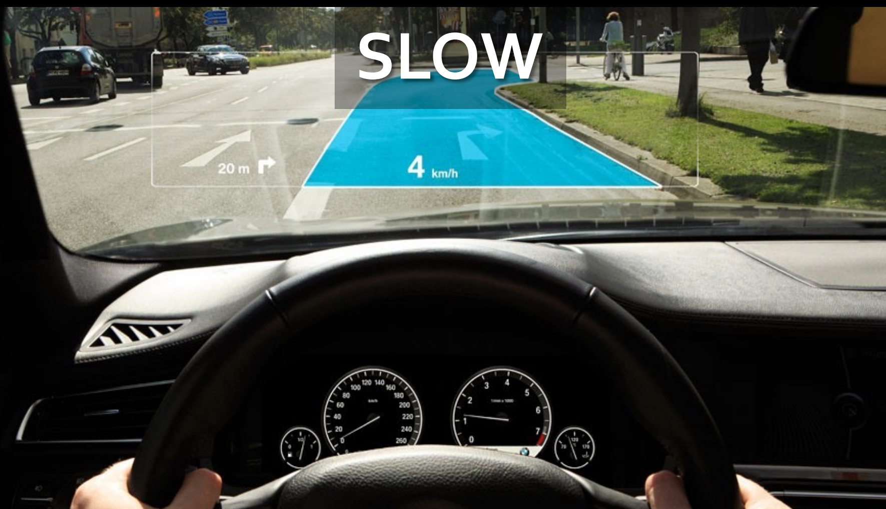
CSAIL

Microsoft Research

Continuous, Real-Time Recognition Apps

- Apps that continuously **locate** and **label** objects in a video stream

Continuous, Real-Time Recognition Apps



Driver Assistance

Augmented Reality Shopping

Face Recognition

Augmented Reality Tourist App

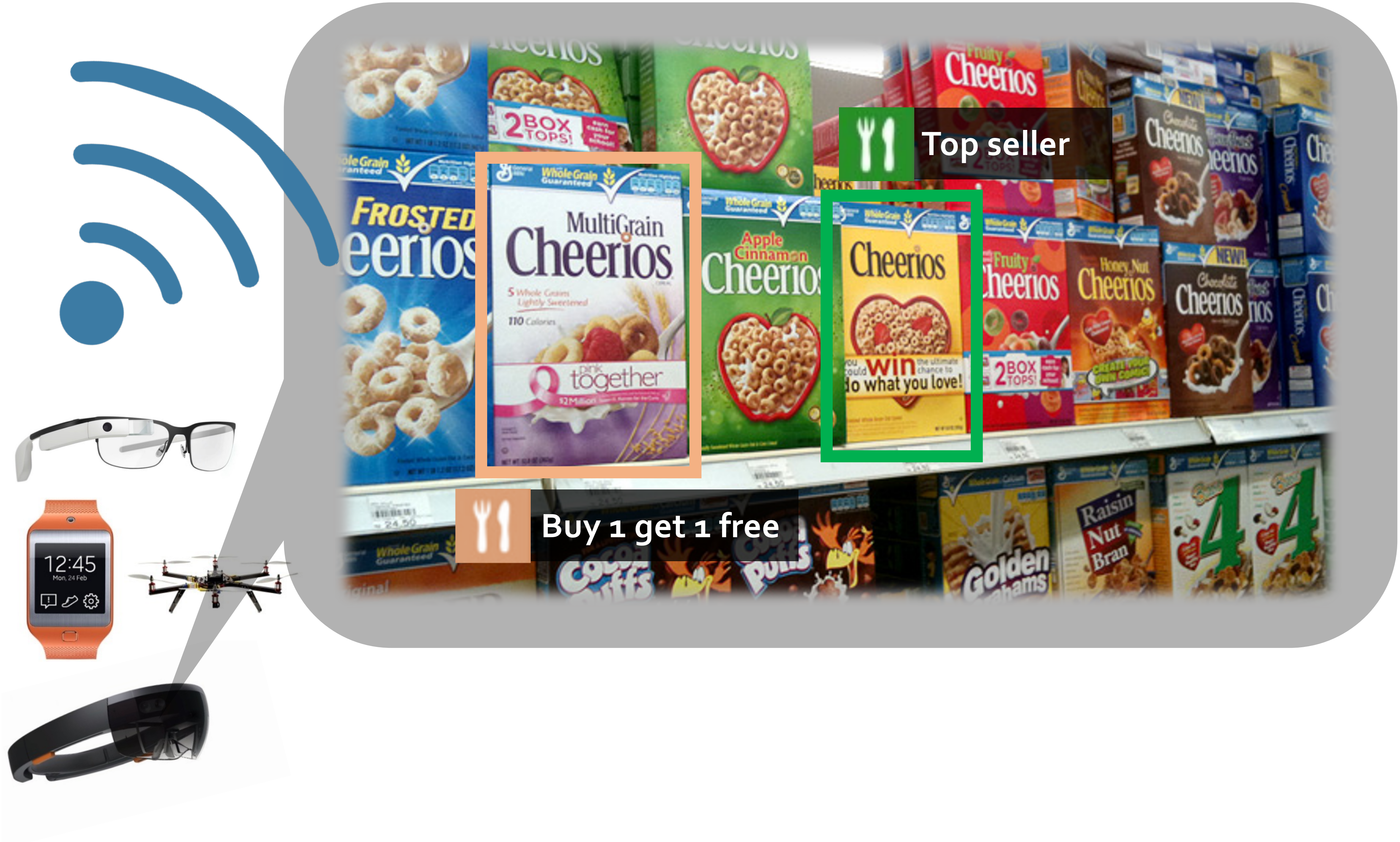
Earlier Designs: Picture-Based Object Recognition

Earlier Designs: Picture-Based Object Recognition

Earlier Designs: Picture-Based Object Recognition

Video-Based Object Recognition

Video-Based Object Recognition



Last Week in IoT (Mobile World Congress)

MWC: Meta asks for better networks to support the metaverse

Zuckerberg says Meta can't do it alone

March 01, 2022 By: Peter Judge Comment

Meta is calling for better networks at Mobile World Congress (MWC) in Barcelona today, with CEO Mark Zuckerberg saying that the metaverse needs more than just smart headsets.

Help us build our virtual world

"Creating a true sense of presence in virtual worlds delivered to smart glasses and VR headsets will require massive advances in connectivity, bigger than any of the step changes we've seen before," Facebook founder and Meta CEO Zuckerberg said in a statement before the opening of MWC.

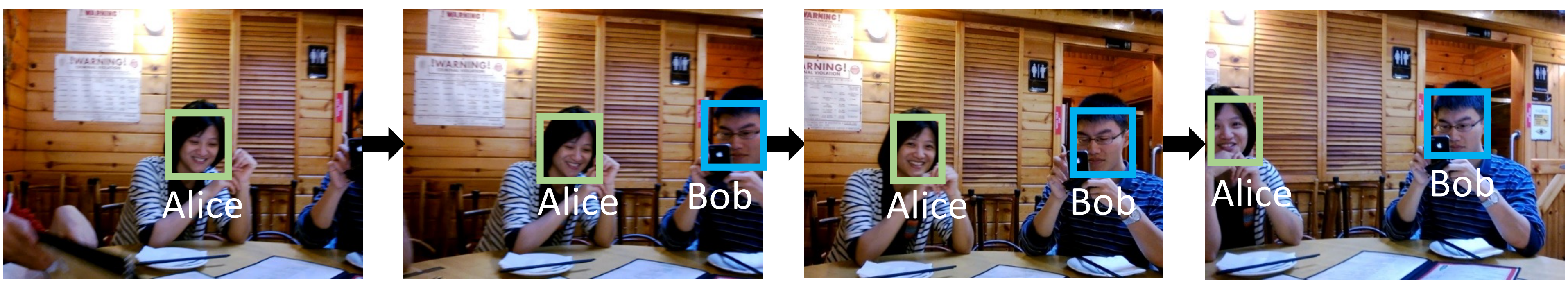
To build an immersive virtual environment and share it with people in real time will require the ability to process data quickly in ways which are not widely supported right now. Zuckerberg's statement says the metaverse needs infrastructure that can evolve quickly, and it can't deliver this without partners.

Glimpse

- Continuous, real-time object recognition on mobile devices in a video stream

Glimpse

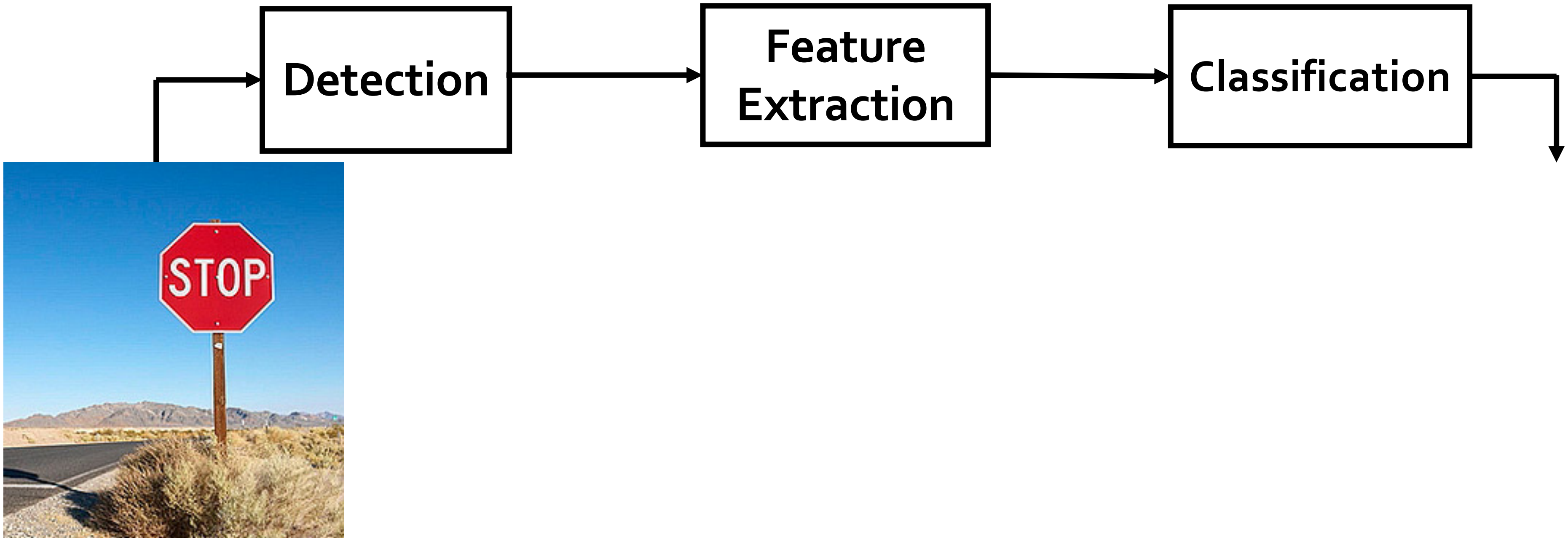
- Continuous, real-time object recognition on mobile devices in a video stream
- Continuously ***identify*** and ***locate*** objects in each frame



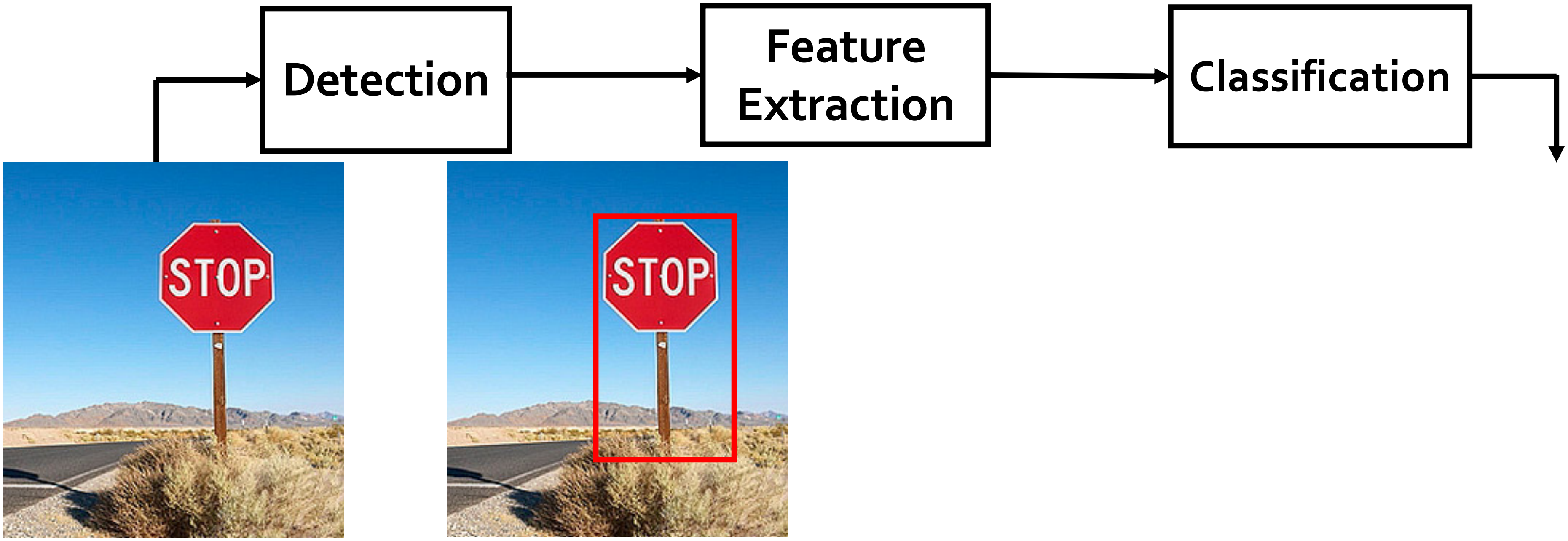
Object Recognition Pipeline

Object Recognition Pipeline

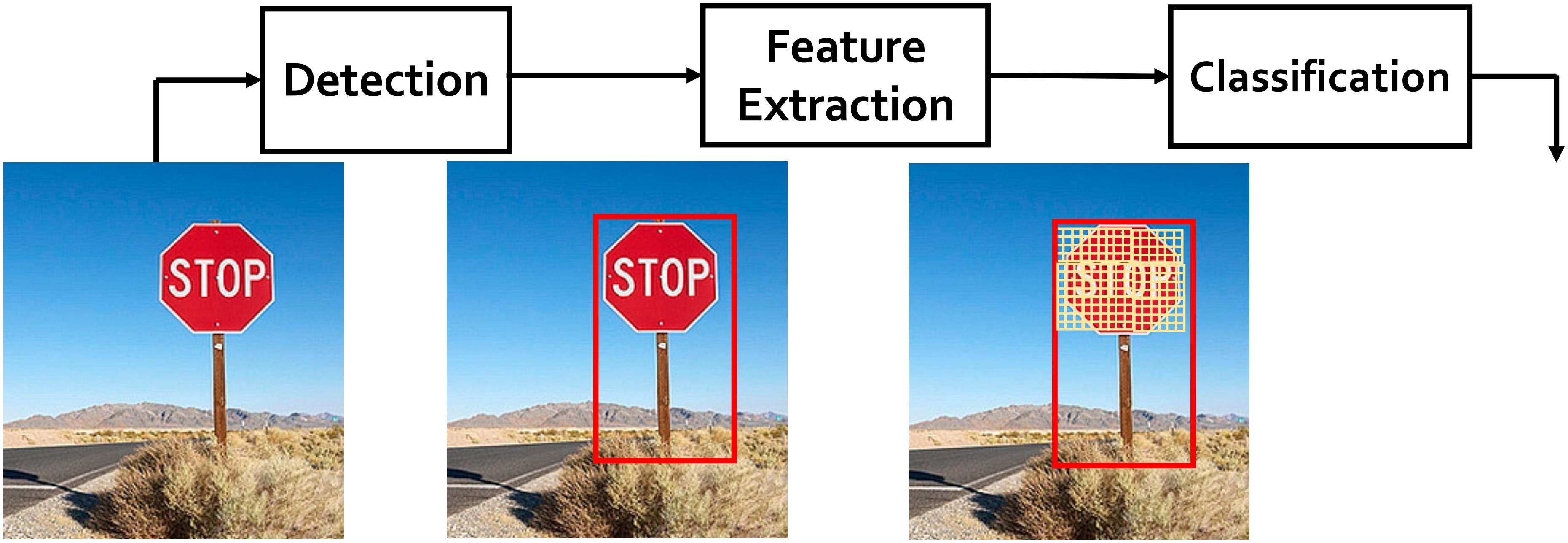
Object Recognition Pipeline



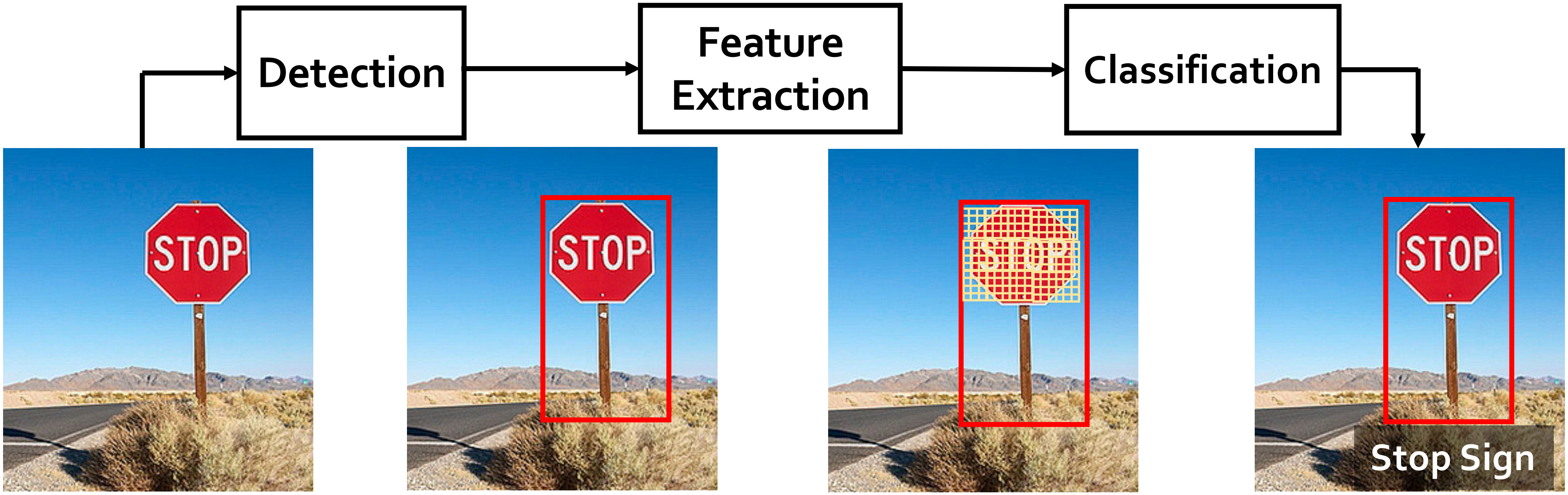
Object Recognition Pipeline



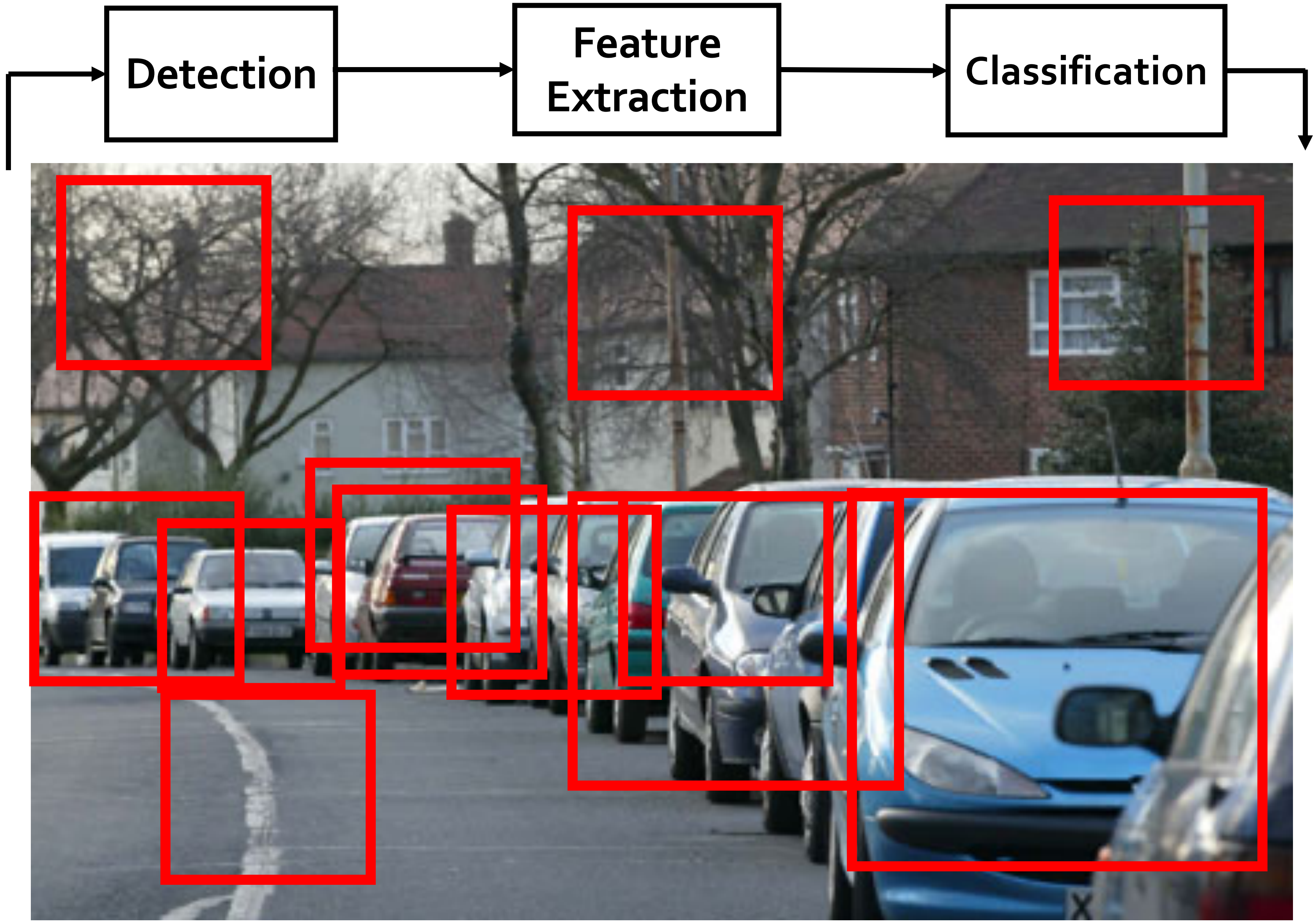
Object Recognition Pipeline



Object Recognition Pipeline



Before Convolutional Neural Network



Before Convolutional Neural Network

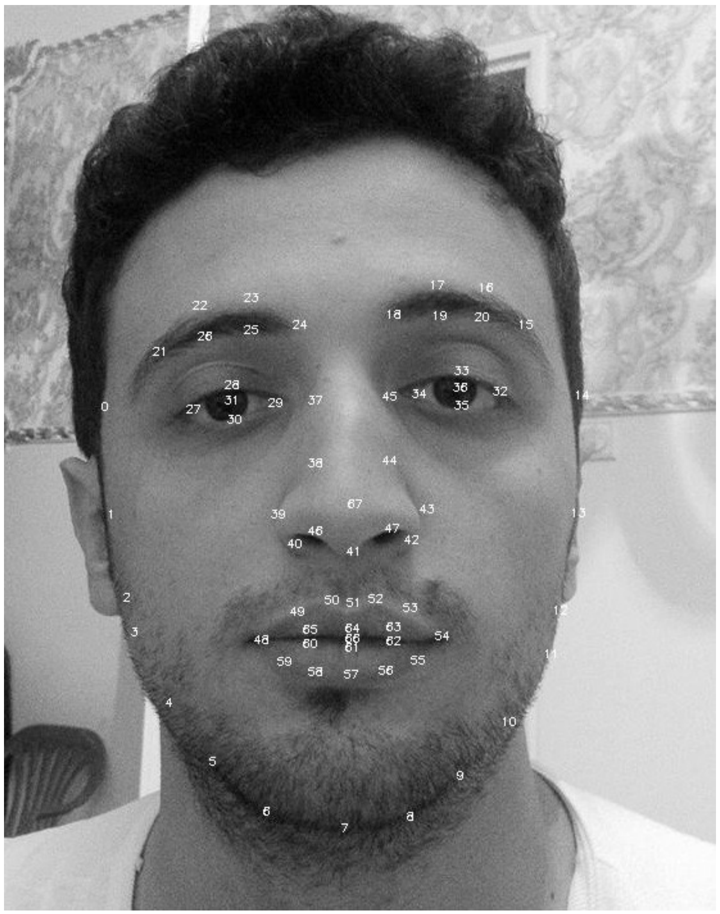
Feature engineering

Feature Extraction

Before Convolutional Neural Network

Feature engineering

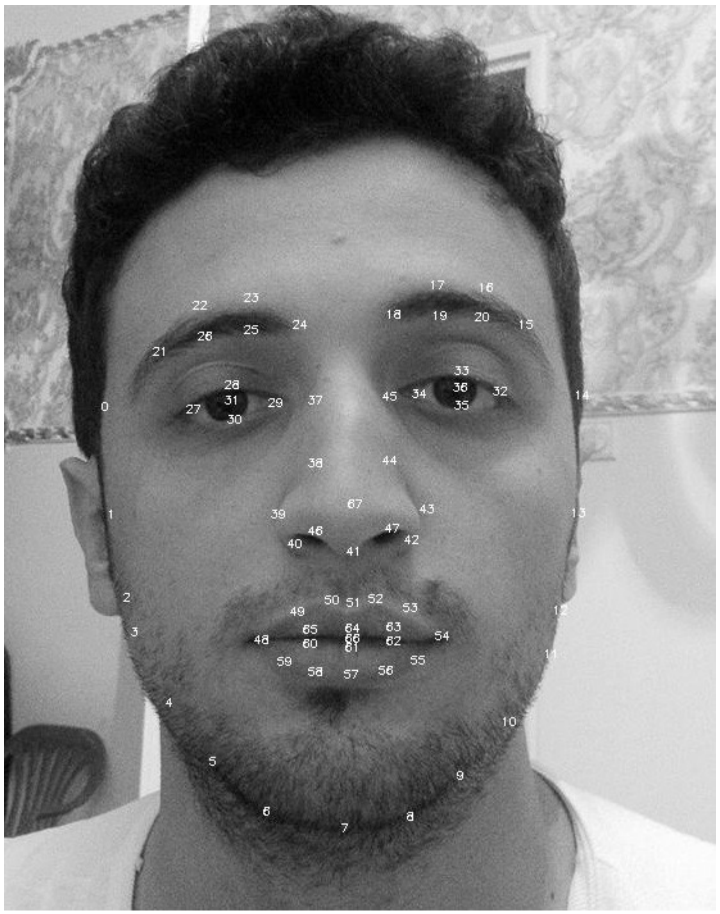
Feature Extraction

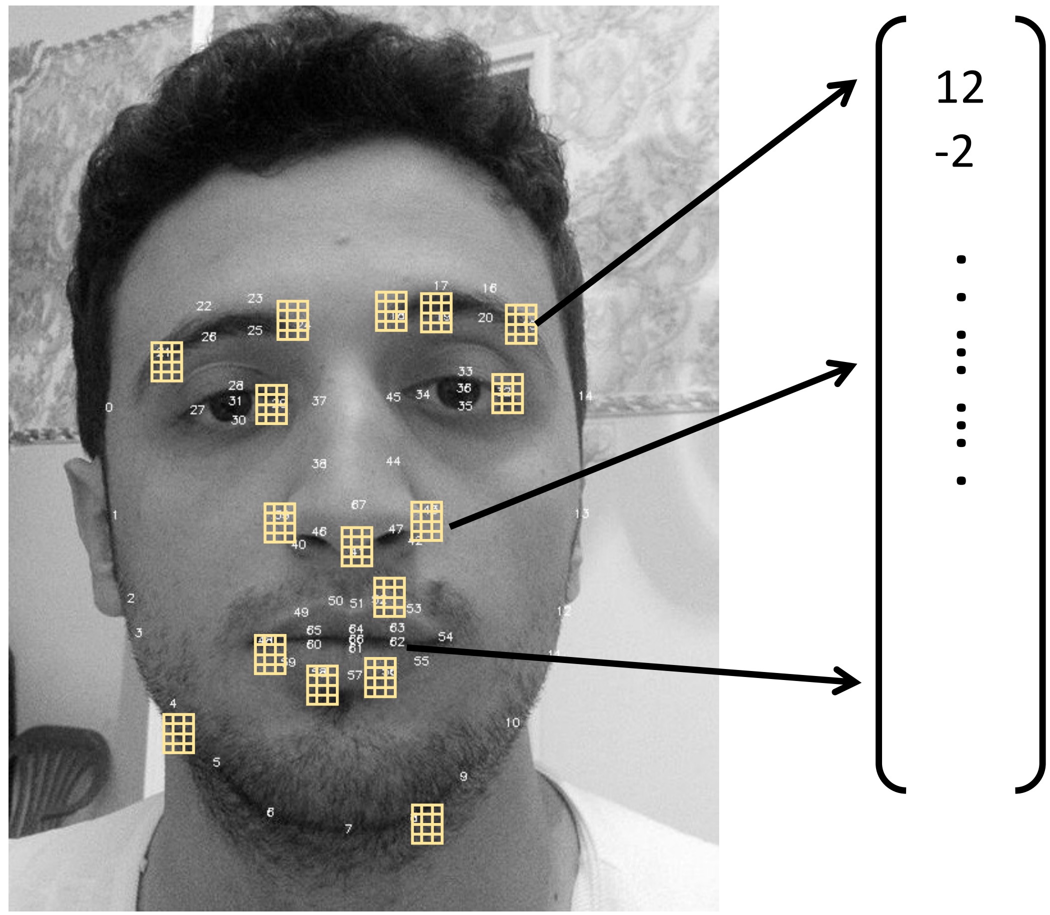


Before Convolutional Neural Network

Feature engineering

Feature Extraction

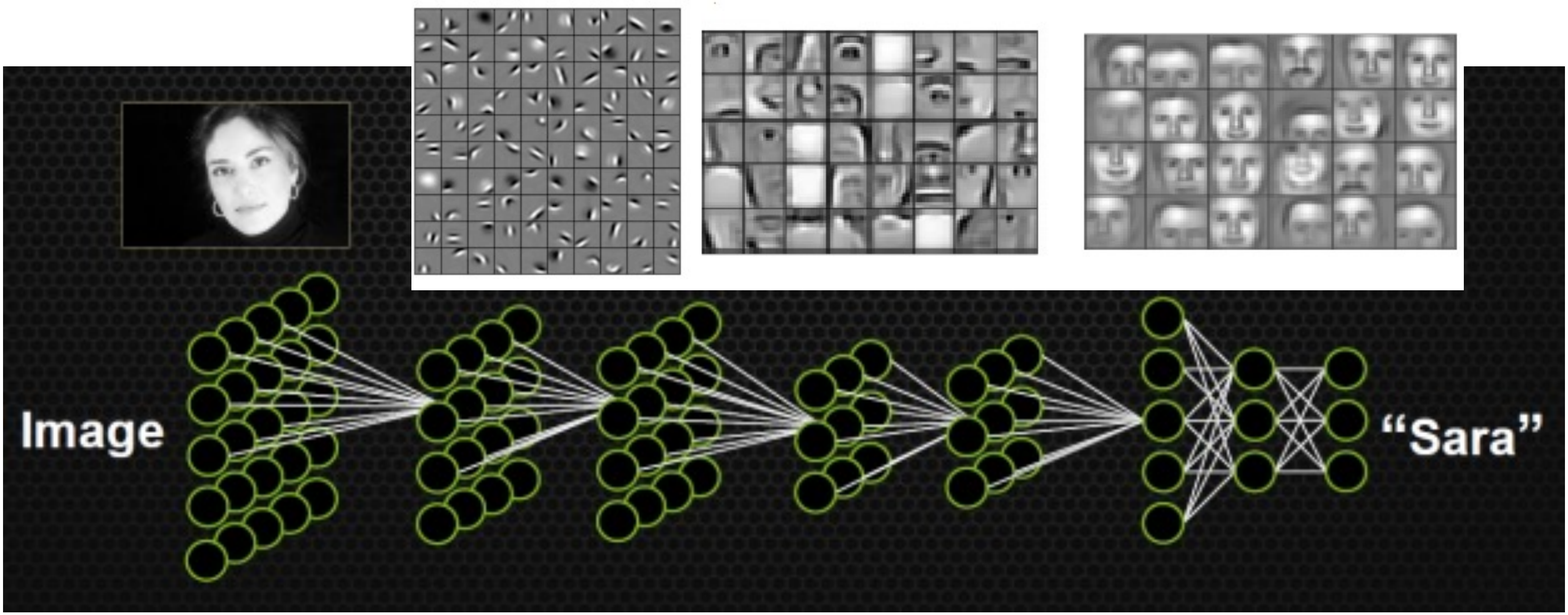




Convolutional Neural Network

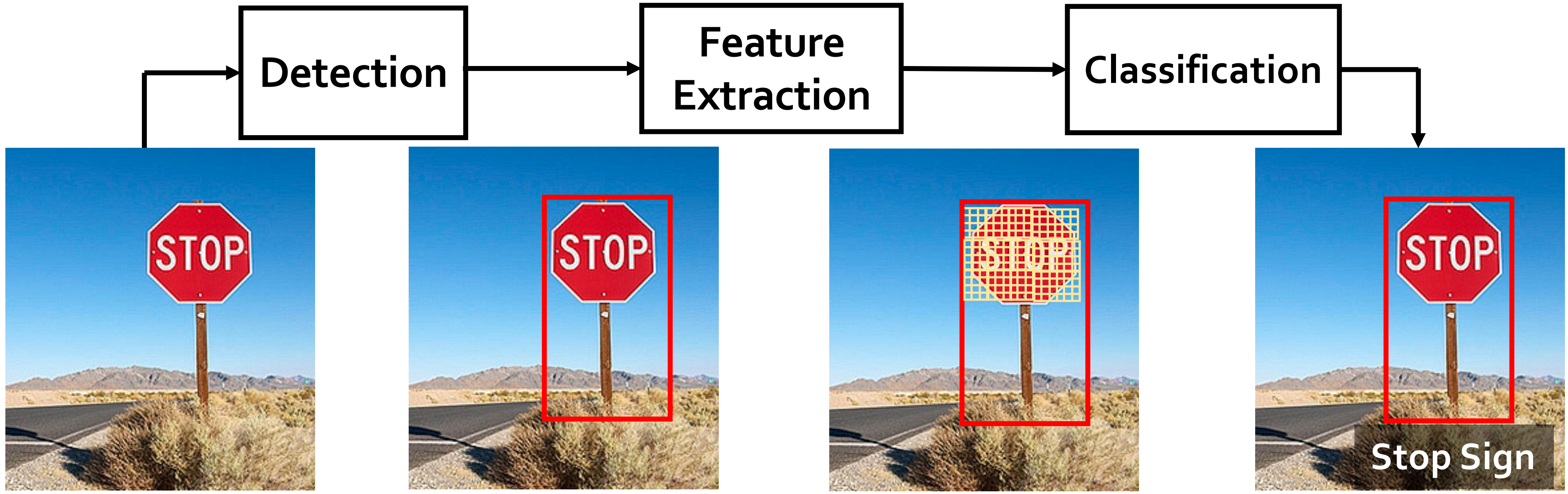
Feature learning

Feature Extraction

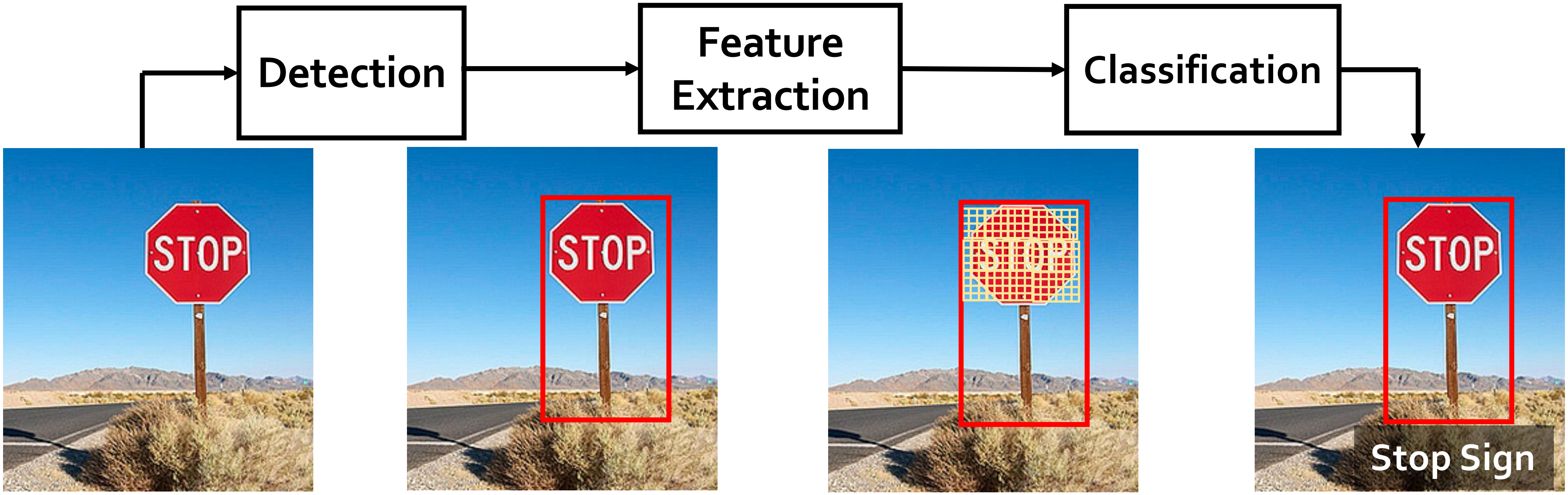


Berkeley caffe http://caffe.berkeleyvision.org/

Object Recognition Pipeline

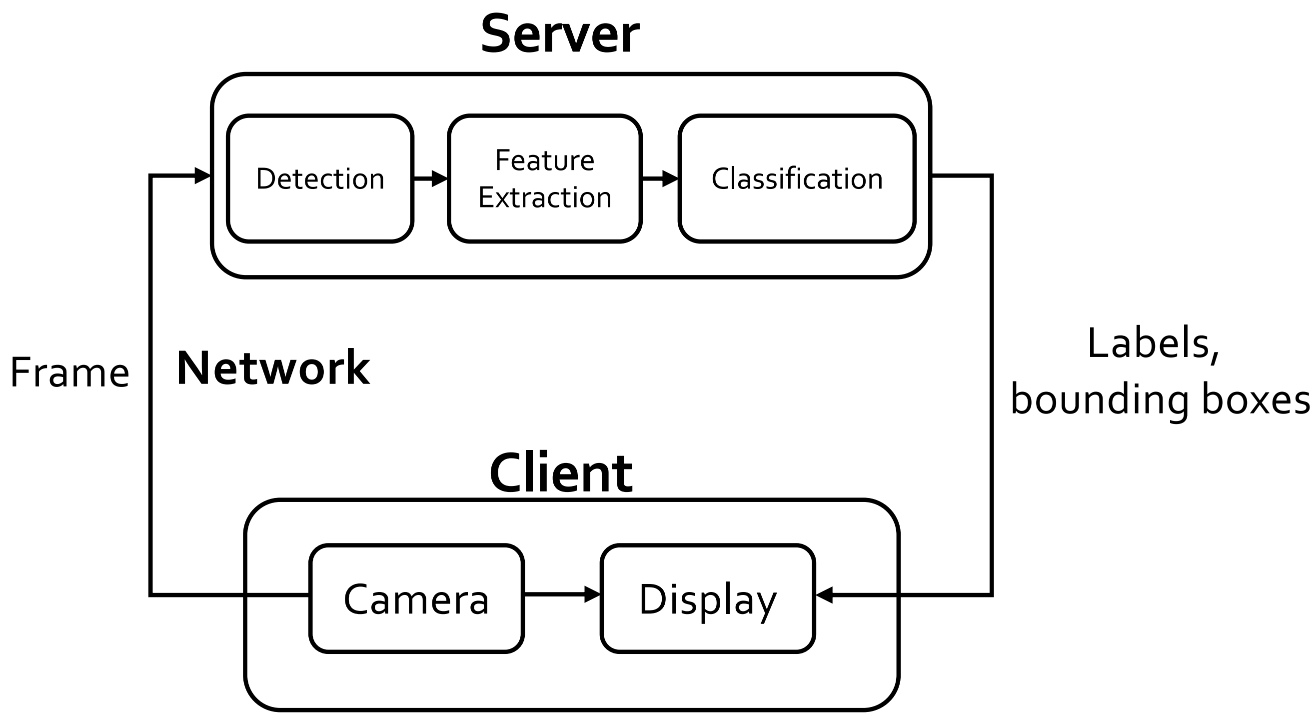


Object Recognition Pipeline



- Computationally expensive and memory-intensive
  - Server is 700x faster than Google Glass
  - Scalability
- We need to offload the recognition pipeline to servers

Client-Server Architecture

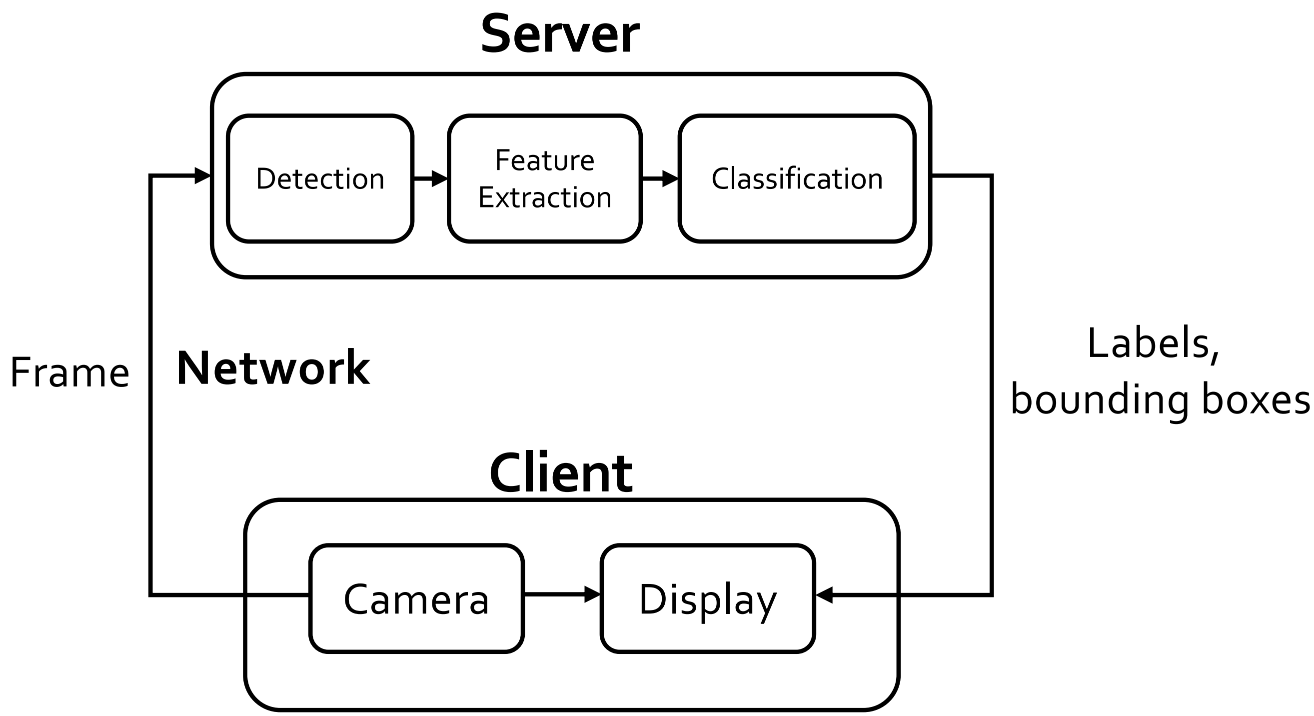


End-to-End Latency Lowers Accuracy

Expected

In reality…

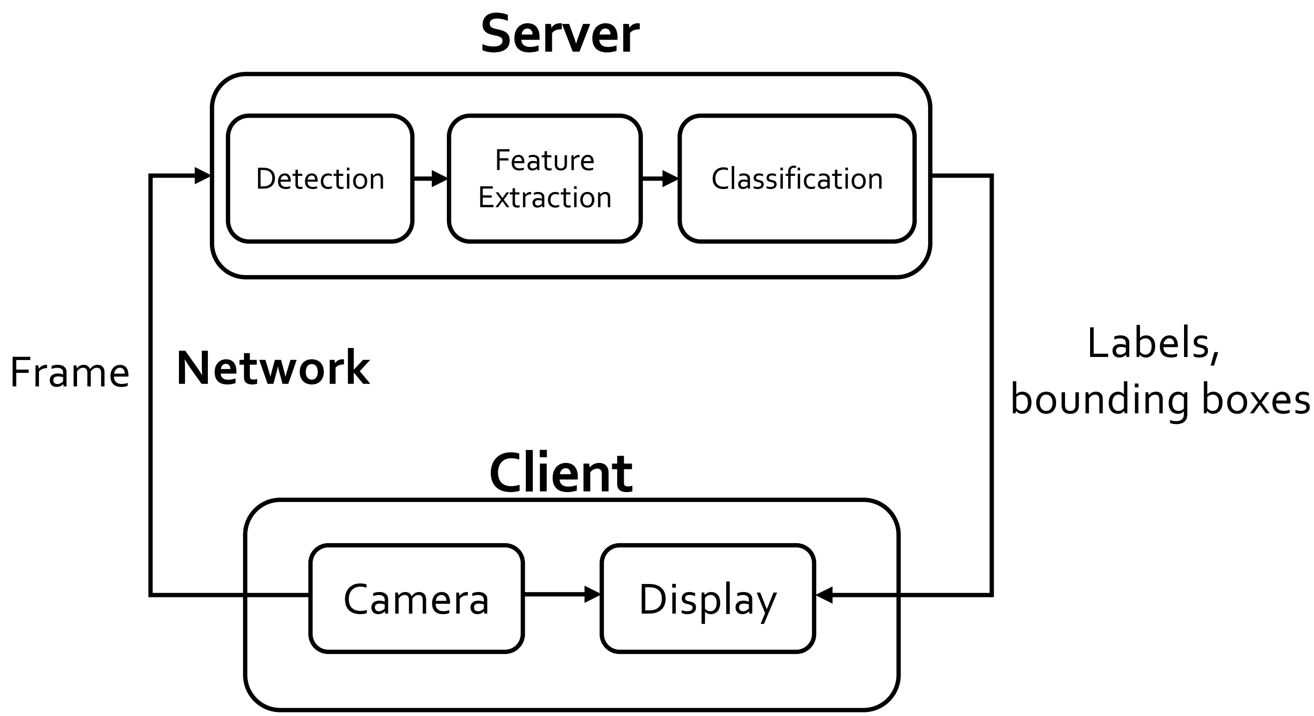
Client-Server Architecture



Challenges

1. **End-to-end latency** lowers object recognition accuracy

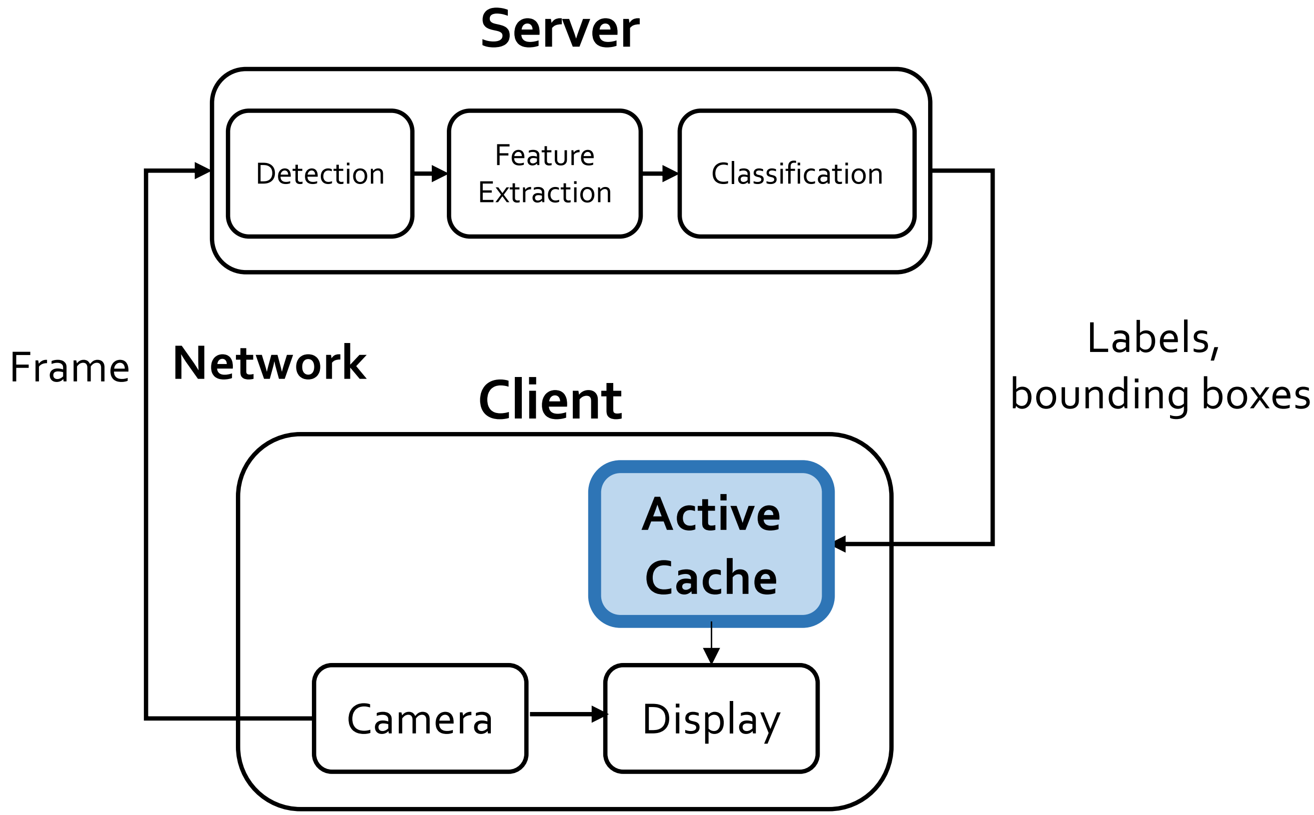
Client-Server Architecture



Challenges

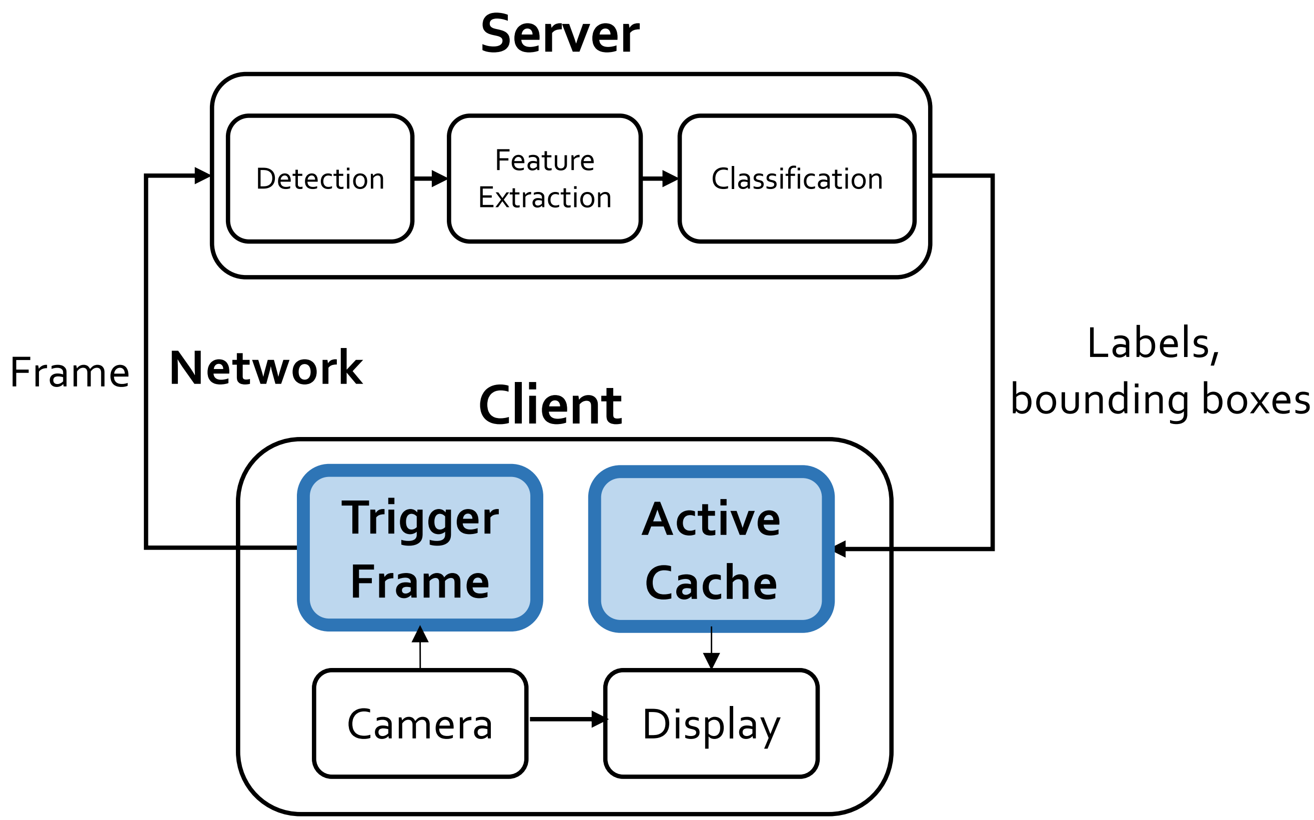
1. **End-to-end latency** lowers object recognition accuracy
2. **Bandwidth** and **energy-efficiency**

Glimpse Architecture



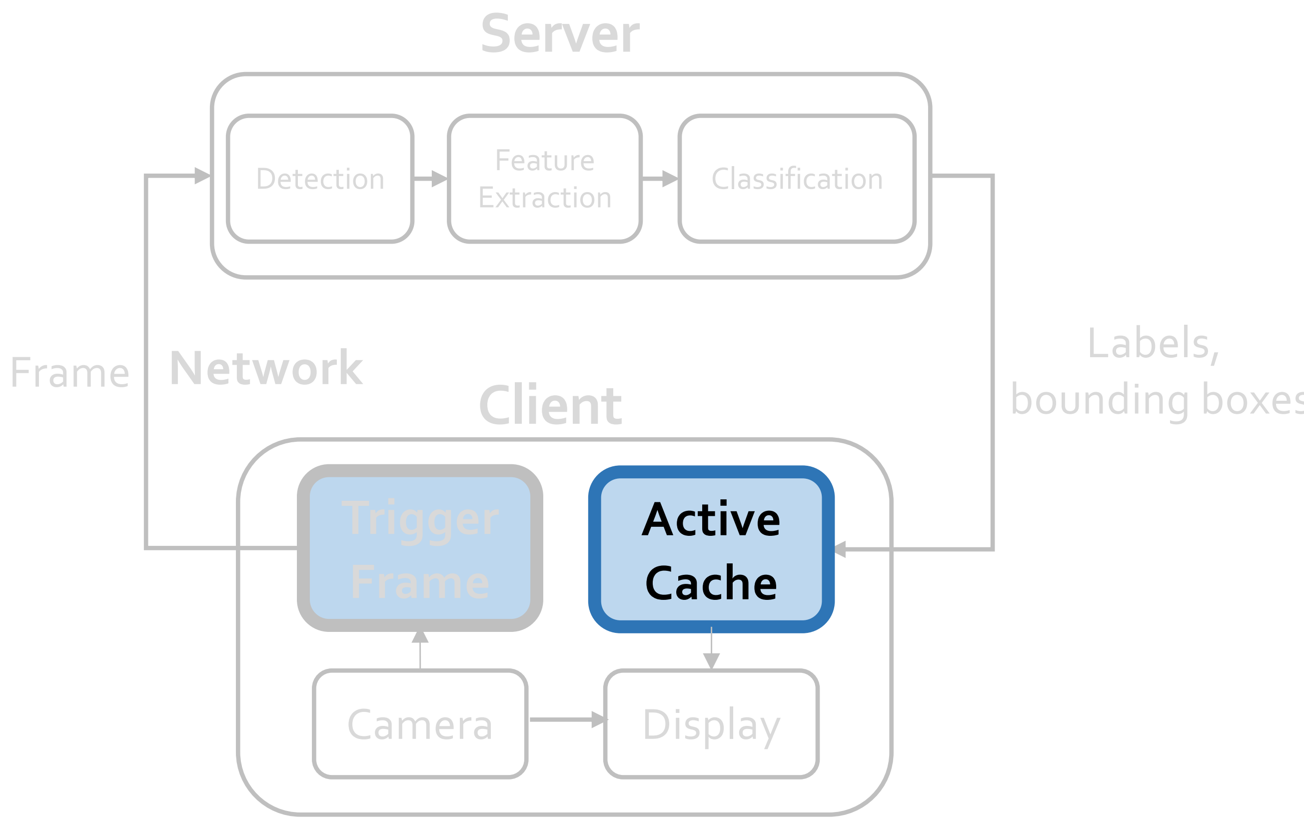
1. **Active Cache** combats e2e latency and regains accuracy

Glimpse Architecture



1. **Active Cache** combats e2e latency and regains accuracy
2. **Trigger Frames** reduce bandwidth usage

Glimpse Architecture



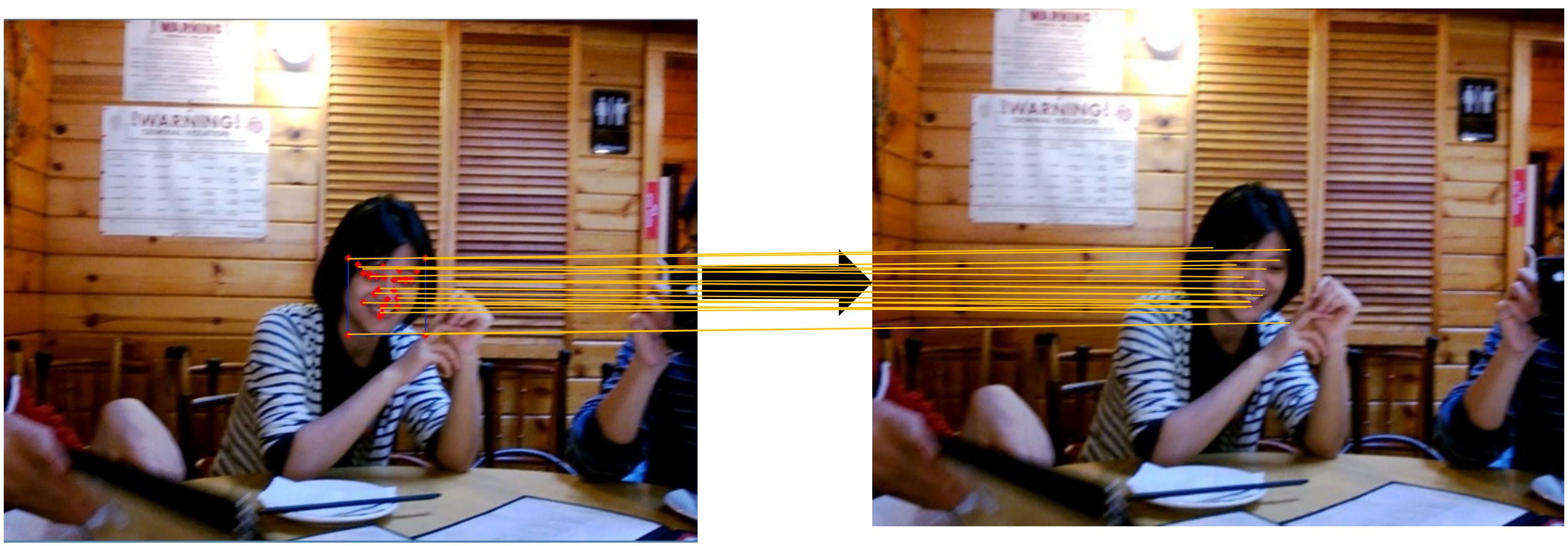
1. **Active Cache** combats e2e latency and regains accuracy

End-to-End Latency Lowers Accuracy

Is it possible to combat latency and regain accuracy?

Relocate Moving Object with Tracking

- Object tracking on the client to re-locate the object

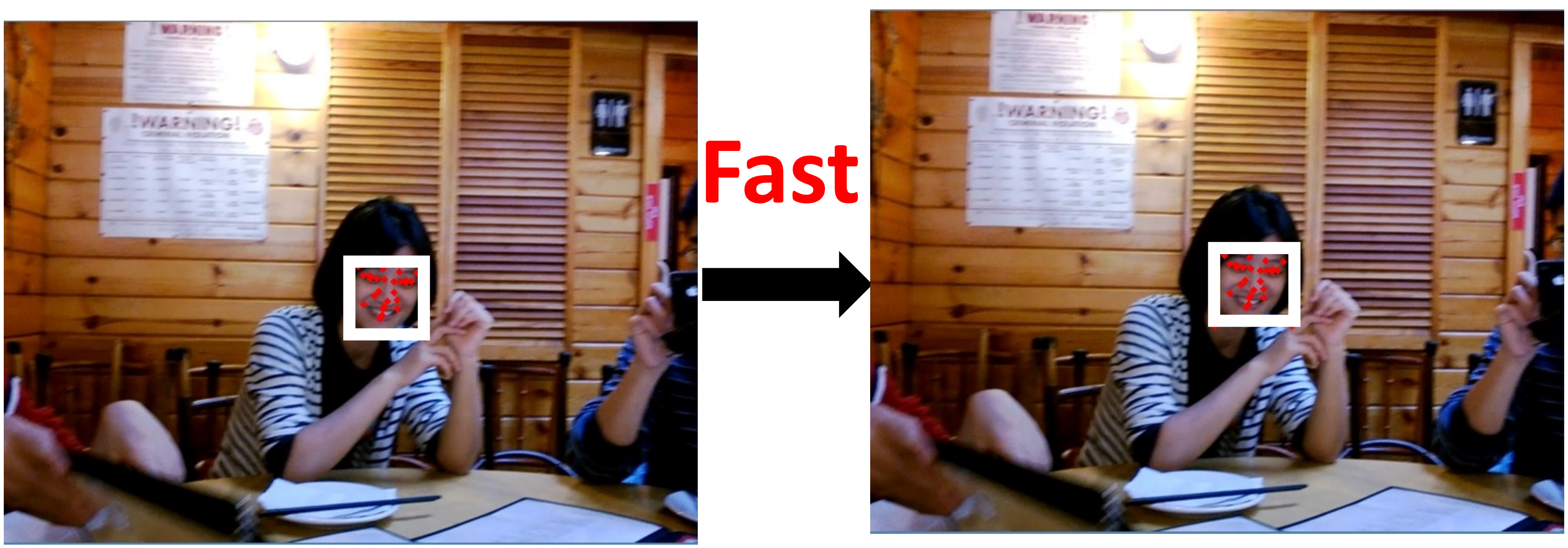


Frame 0

Frame 12 (delay = 360 ms)

Relocate Moving Object with Tracking

- Object tracking on the client to re-locate the object



Frame 0

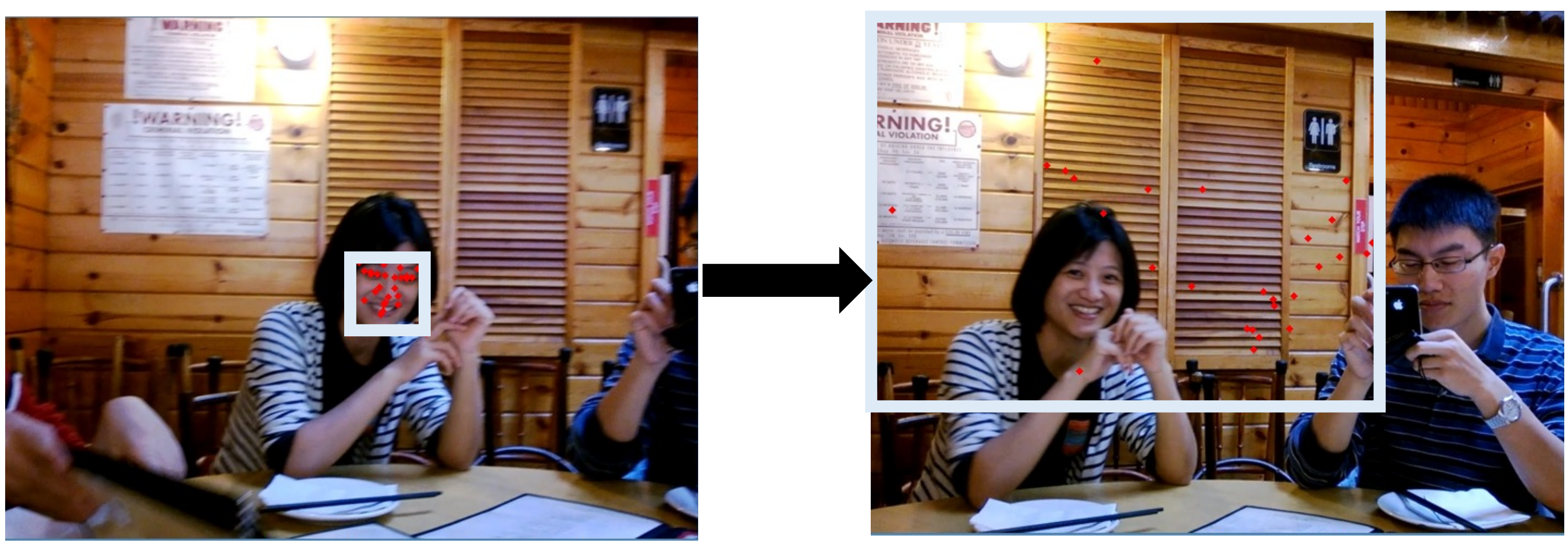
Frame 12 (delay = 360 ms)

Relocate Moving Object with Tracking

- Object tracking on the client to re-locate the object
- Fails to work when object displacement is large

Relocate Moving Object with Tracking

- Object tracking on the client to re-locate the object
- Fails to work when object displacement is large



Frame 0

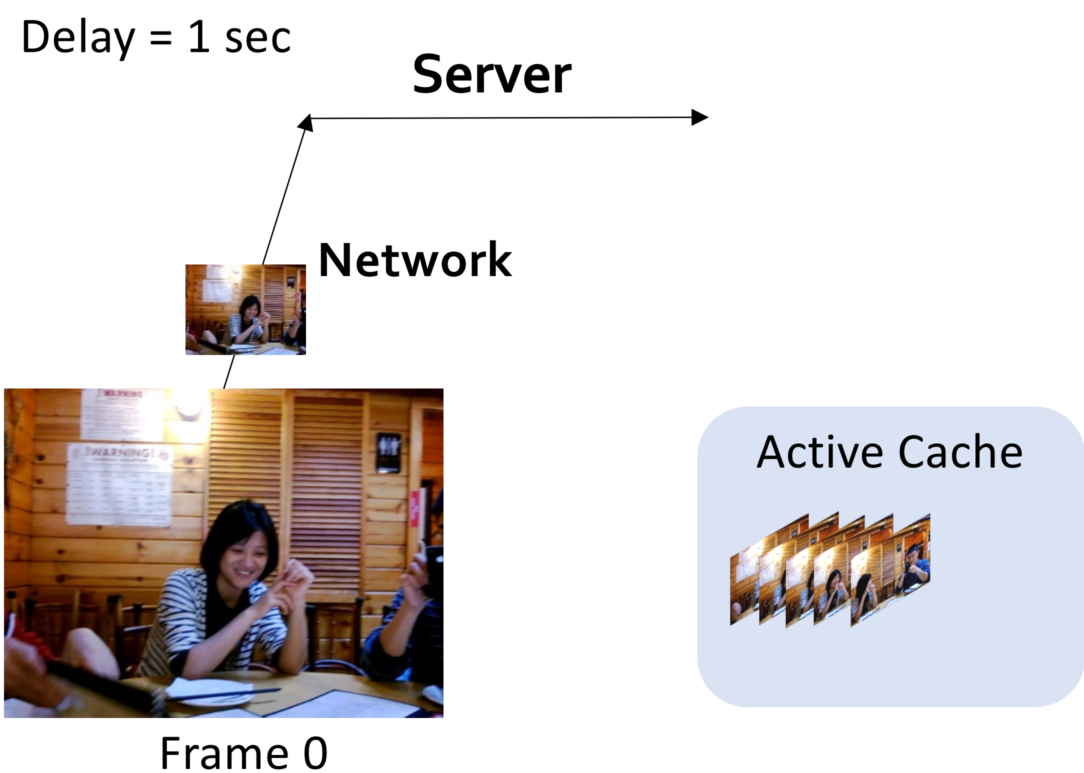
Frame 30 (delay= 1 sec)

Regain Accuracy with *Active Cache*

- Cache and run tracking through the cached frames

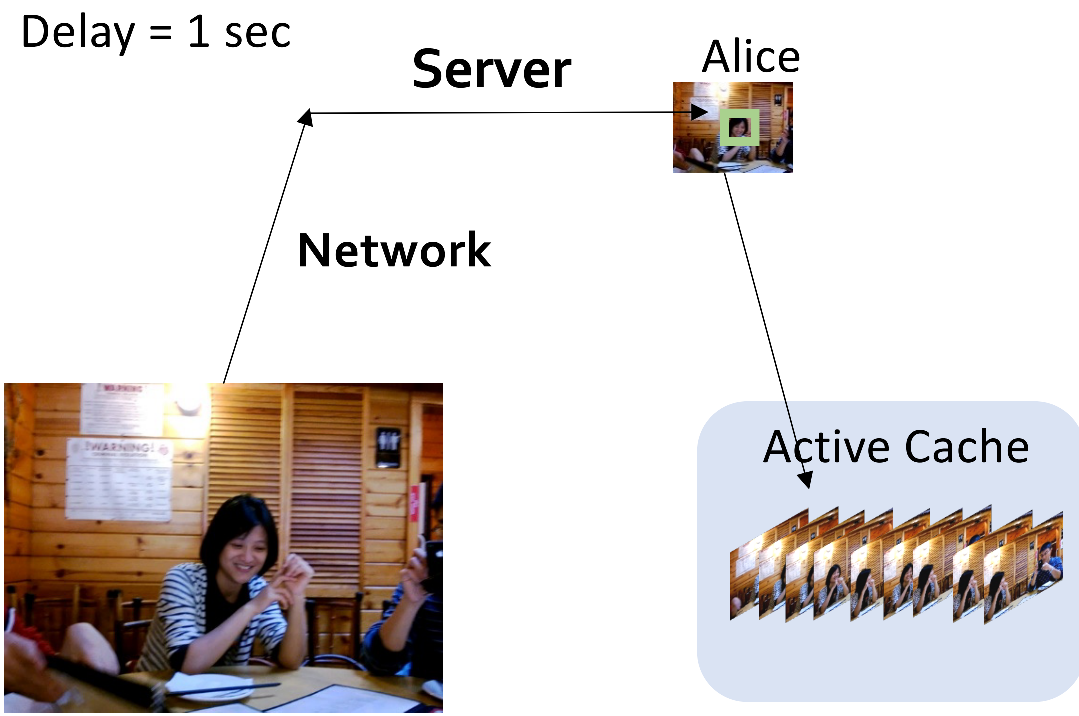
Regain Accuracy with *Active Cache*

- Cache and run tracking through the cached frames



Regain Accuracy with *Active Cache*

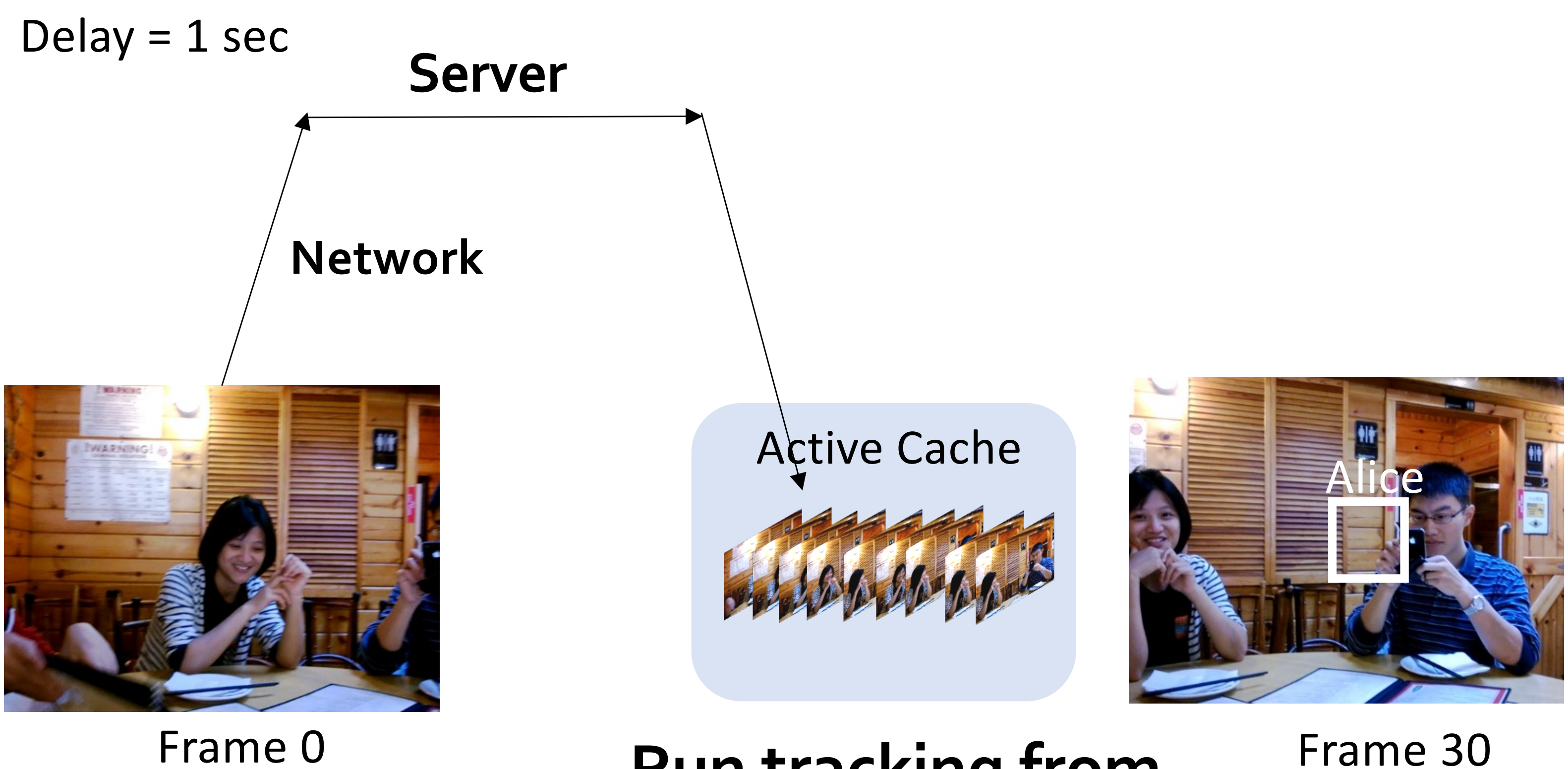
- Cache and run tracking through the cached frames



Frame 0

Regain Accuracy with *Active Cache*

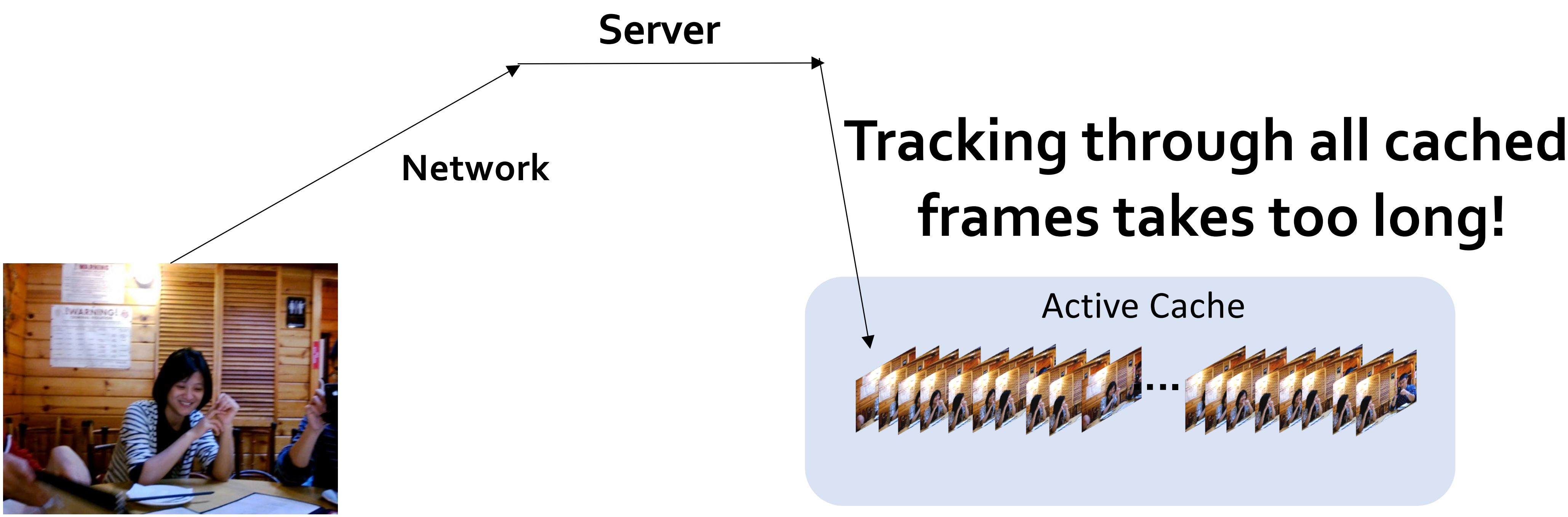
- Cache and run tracking through the cached frames



Run tracking from Frame 0 to Frame 30

Regain Accuracy with *Active Cache*

- Cache and run tracking through the cached frames



Regain Accuracy with *Active Cache*

- Cache and run tracking through the cached frames

Server

Network

Tracking through all cached frames takes too long!

Active Cache

Adaptive Frame Selection

Given ***n_cached*** frames, select ***s_selected*** frames so that we can catch up without sacrificing tracking performance

Adaptive Frame Selection

Given ***n_cached*** frames, select ***s_selected*** frames so that we can catch up without sacrificing tracking performance

1. How many frames to select?
2. Which frames to select?

Adaptive Frame Selection

Given ***n_cached*** frames, select ***s_selected*** frames so that we can catch up without sacrificing tracking performance

1. **How many frames to select?**

- ***s_selected:*** active cache processing time vs. tracking accuracy

Adaptive Frame Selection

Given ***n_cached*** frames, select ***s_selected*** frames so that we can catch up without sacrificing tracking performance

1. **How many frames to select?**

- ***s_selected:*** active cache processing time vs. tracking accuracy

What is the maximum number of frames that can be tracked?

e = execution time for processing any frame in the active cache

N frames per second

=> have 1/N seconds before next frame

=> Can process s_selected = (1/N)/e frames

Adaptive Frame Selection

Given ***n_cached*** frames, select ***s_selected*** frames so that we can catch up without sacrificing tracking performance

1. **How many frames to select?**

- ***s_selected:*** active cache processing time vs. tracking accuracy

What is the maximum number of frames that can be tracked?

What if I'm okay with increasing the latency a bit?

e = execution time for processing any frame in the active cache

N frames per second => have 1/N seconds before next frame

If I'm fine with a lag of t frames

=> Can process s_selected = (t/N)/e frames

Adaptive Frame Selection

Given ***n_cached*** frames, select ***s_selected*** frames so that we can catch up without sacrificing tracking performance

2. **Given *s_selected*, which frames to select?**

- Temporal redundancy between frames

Adaptive Frame Selection

Given ***n_cached*** frames, select ***s_selected*** frames so that we can catch up without sacrificing tracking performance

2. Given *s_selected*, which frames to select?

- Temporal redundancy between frames
- Use *frame differencing* to quantify movement and select frames to capture as much movement as possible

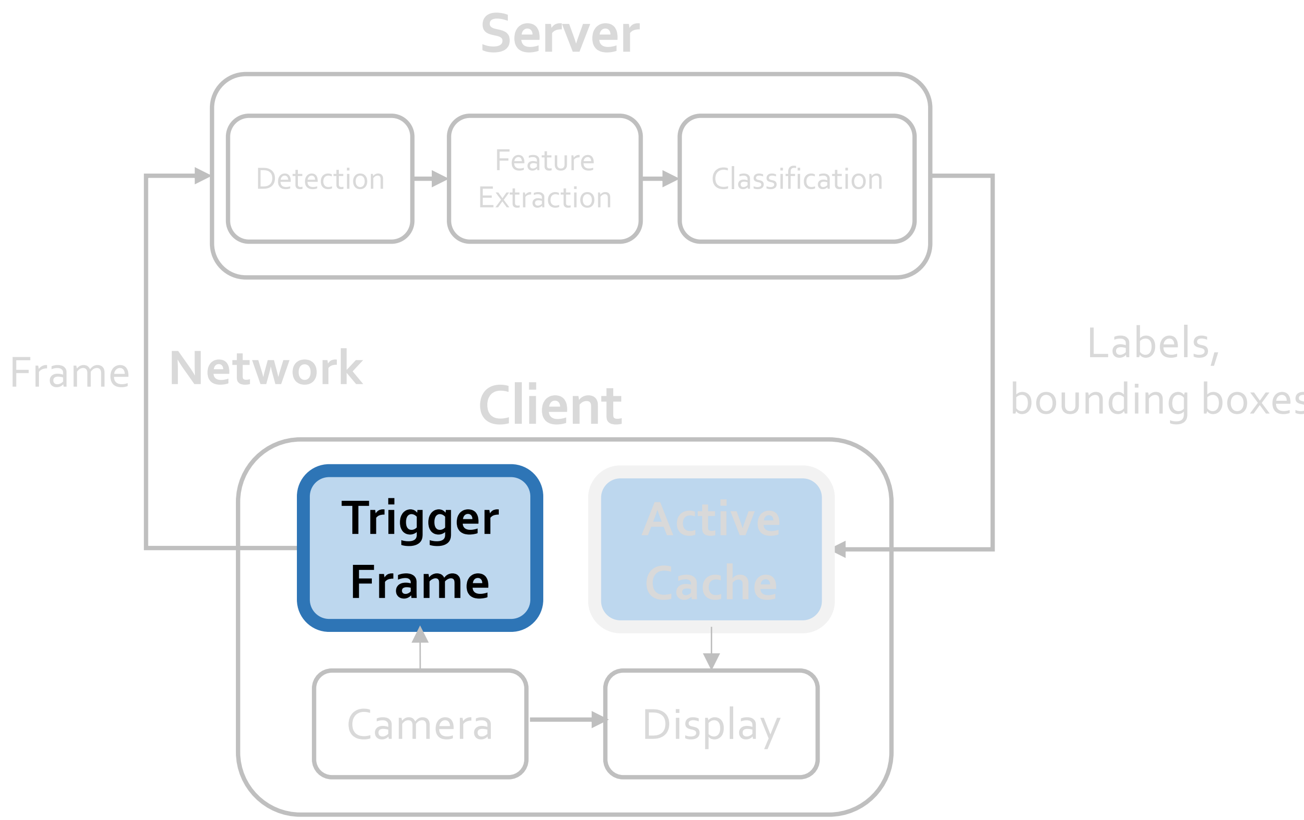
Active Cache Short Question

- Does Glimpse reduce the end-to-end latency of object recognition?

Active Cache Short Question

- Does active cache reduce the end-to-end latency of object recognition?
  - No. It's a trick to fool the user into thinking that the recognition is in real time.

Glimpse Architecture



1. **Active Cache** combats e2e latency and regains accuracy
2. **Trigger Frames** reduce bandwidth usage

Reduce Bandwidth Usage with Trigger Frames

- Strategically send certain trigger frames to the server

Reduce Bandwidth Usage with Trigger Frames

- Strategically send certain trigger frames to the server

1. Measuring scene changes from the previously processed frame

Reduce Bandwidth Usage with Trigger Frames

- Strategically send certain trigger frames to the server

1. Measuring scene changes from the previously processed frame
2. Detecting tracking failure

- Feature points deviate when the size, angle, or appearance of the object changes.
- The standard deviation of distance of all tracked points between two frames

Reduce Bandwidth Usage with Trigger Frames

- Strategically send certain trigger frames to the server

1. Measuring scene changes from the previously processed frame
2. Detecting tracking failure

- Limiting the number of frames in-flight
  - 1 frame in-flight strikes the best balance between bandwidth and accuracy

Evaluation

- **Object recognition pipelines**
  1. Face recognition
  2. Road sign recognition

Evaluation

- **Object recognition pipelines**
  1. Face recognition
  2. Road sign recognition
- **Datasets**
  1. **Face Dataset:**
     - 26 videos recorded with a smartphone
     - 30 minutes, 54K frames, and 36K faces
     - Scenarios: shopping with friends and waiting at a subway station
  2. **Road Sign Dataset:**
     - 4 walking videos recorded using Google Glass from YouTube
     - 35 minutes, 63K frames, and 5K road signs

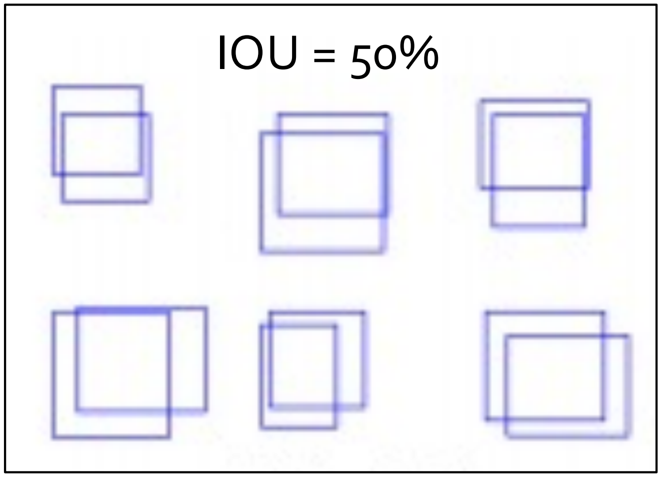
Evaluation

- **Evaluation Metrics**
  - Intersection over union (IOU) to measure recognition accuracy

$$IOU_i = \frac{area\,|O_i \cap G_i|}{area\,|O_i \cup G_i|}$$

O*i:* bounding box of the detected object *i*

G*i:* bounding box of object *i*'s ground truth



  - **Correct if IOU > 50% and the label matches ground truth**

Evaluation

- **Evaluation Metrics**
  - Precision

$$\frac{\text{\# of objects correctly labeled and located}}{\text{total \# of objects detected}}$$

  - Recall

$$\frac{\text{\# of objects correctly labeled and located}}{\text{total \# of objects in the ground truth}}$$

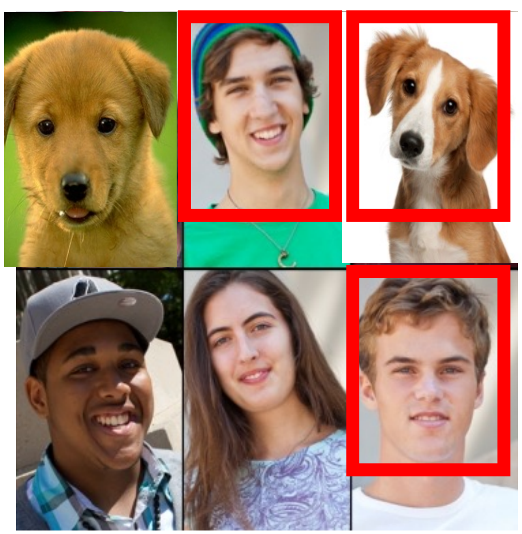
Evaluation

- **Evaluation Metrics**
  - Precision

$$\frac{\text{\# of objects correctly labeled and located}}{\text{total \# of objects detected}}$$

  - Recall

$$\frac{\text{\# of objects correctly labeled and located}}{\text{total \# of objects in the ground truth}}$$



faces in the ground truth:4

faces detected: 3

faces correctly labeled and detected: 2

Precision:

Recall:

Evaluation

- **Evaluation Metrics**
  - Precision

$$\frac{\text{\# of objects correctly labeled and located}}{\text{total \# of objects detected}}$$

  - Recall

$$\frac{\text{\# of objects correctly labeled and located}}{\text{total \# of objects in the ground truth}}$$

faces in the ground truth:4
faces detected: 3
faces correctly labeled and detected: 2

Precision: 2/3
Recall: 2/4

Evaluation

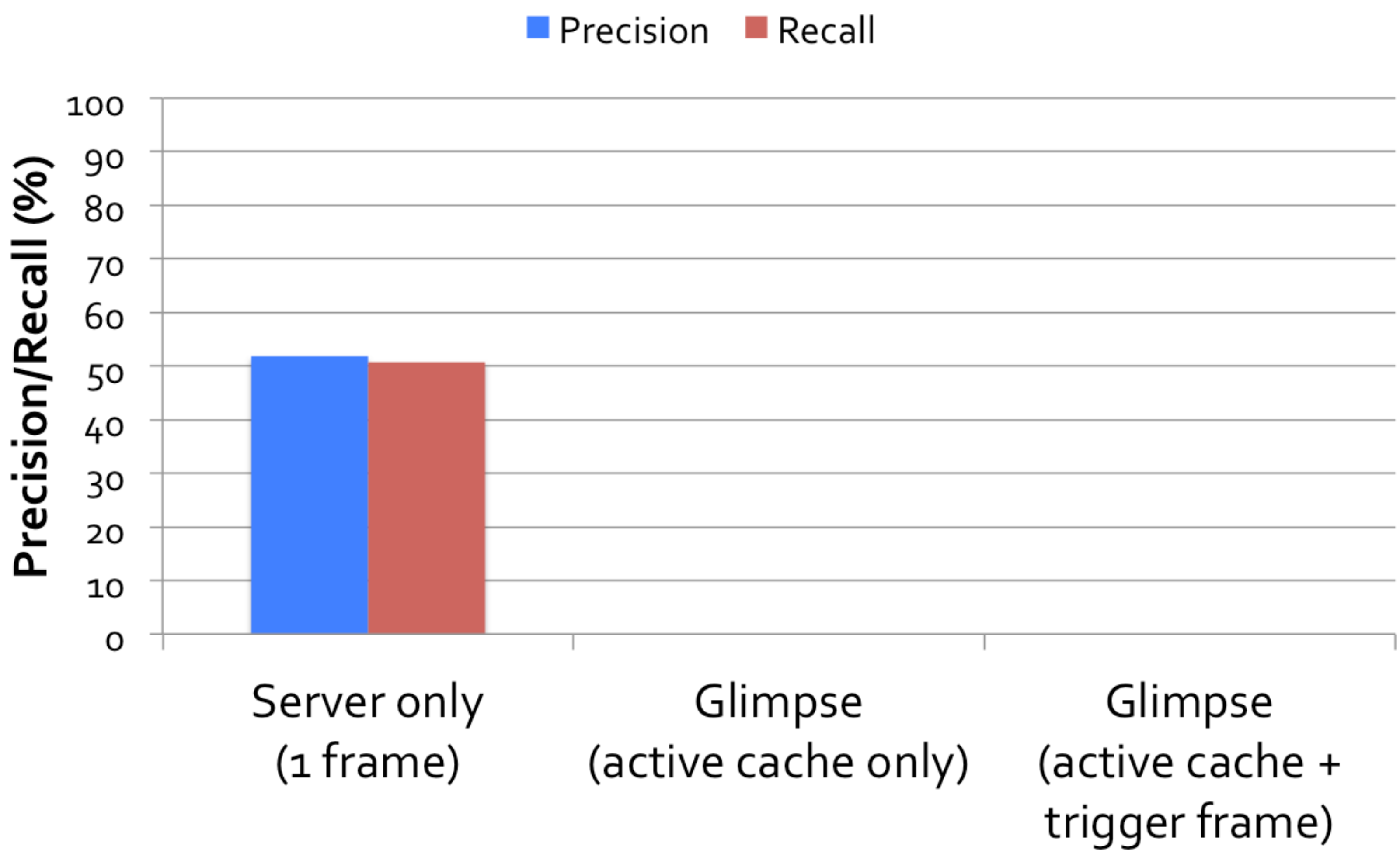
- **Network conditions**
  - Wi-Fi, Verizon's LTE, and AT&T's LTE network

Results Outline

1. Face recognition
2. Road sign recognition
3. Face recognition with hardware-assisted face detection

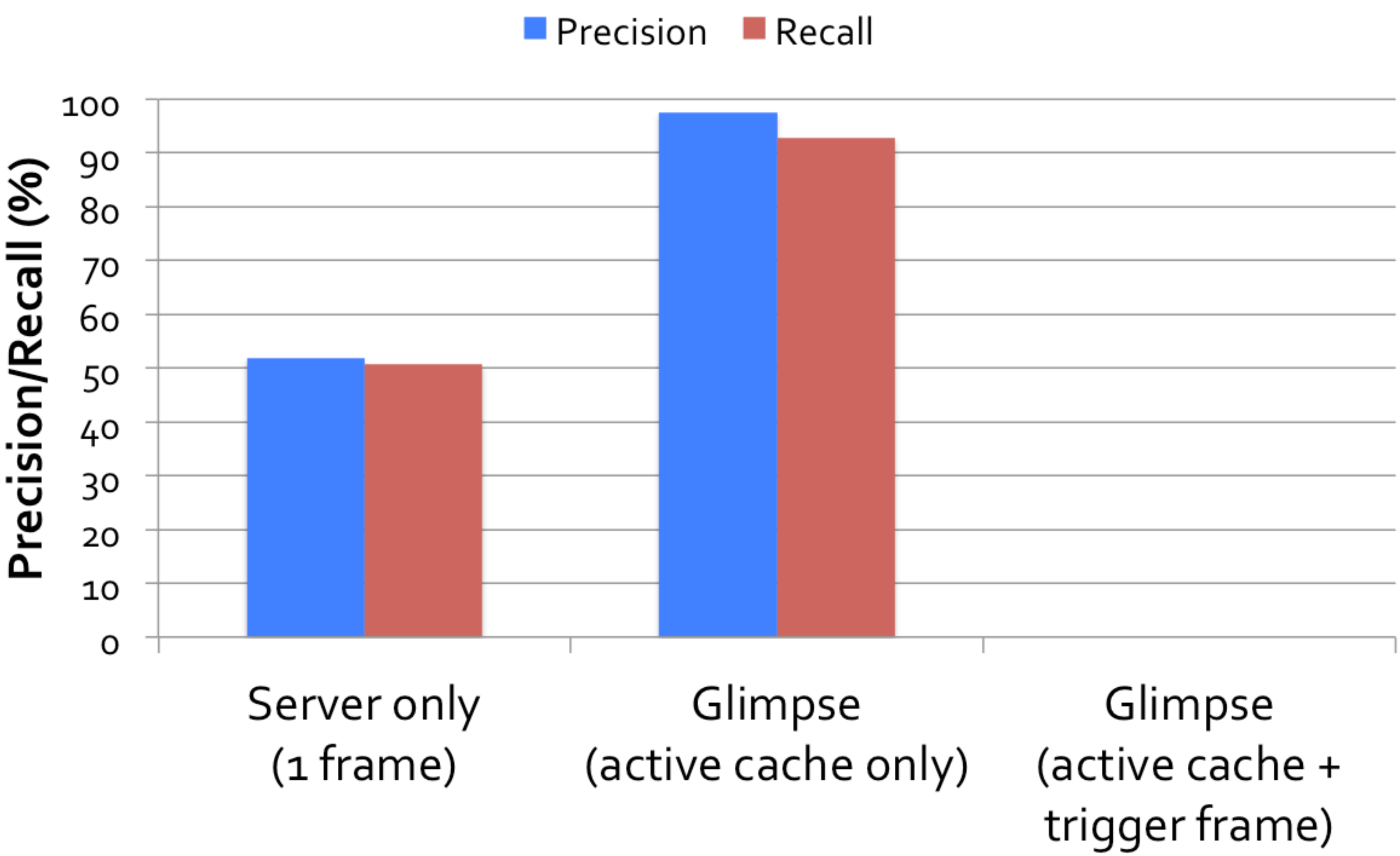
Active Cache Achieves High Accuracy

- Face dataset
- Wi-Fi (End-to-end delay: 430 ms)



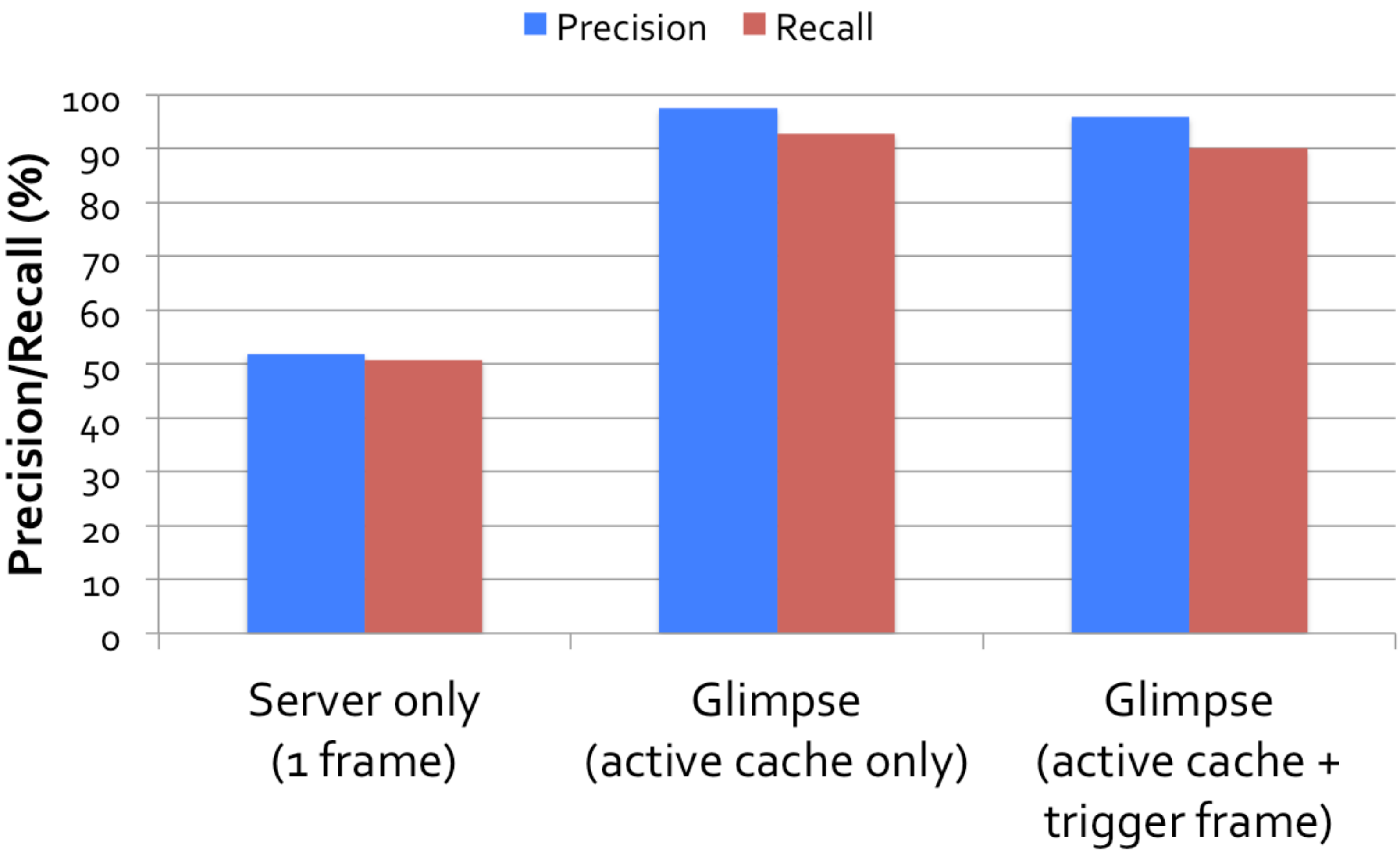
Active Cache Achieves High Accuracy

- Face dataset
- Wi-Fi (End-to-end delay: 430 ms)



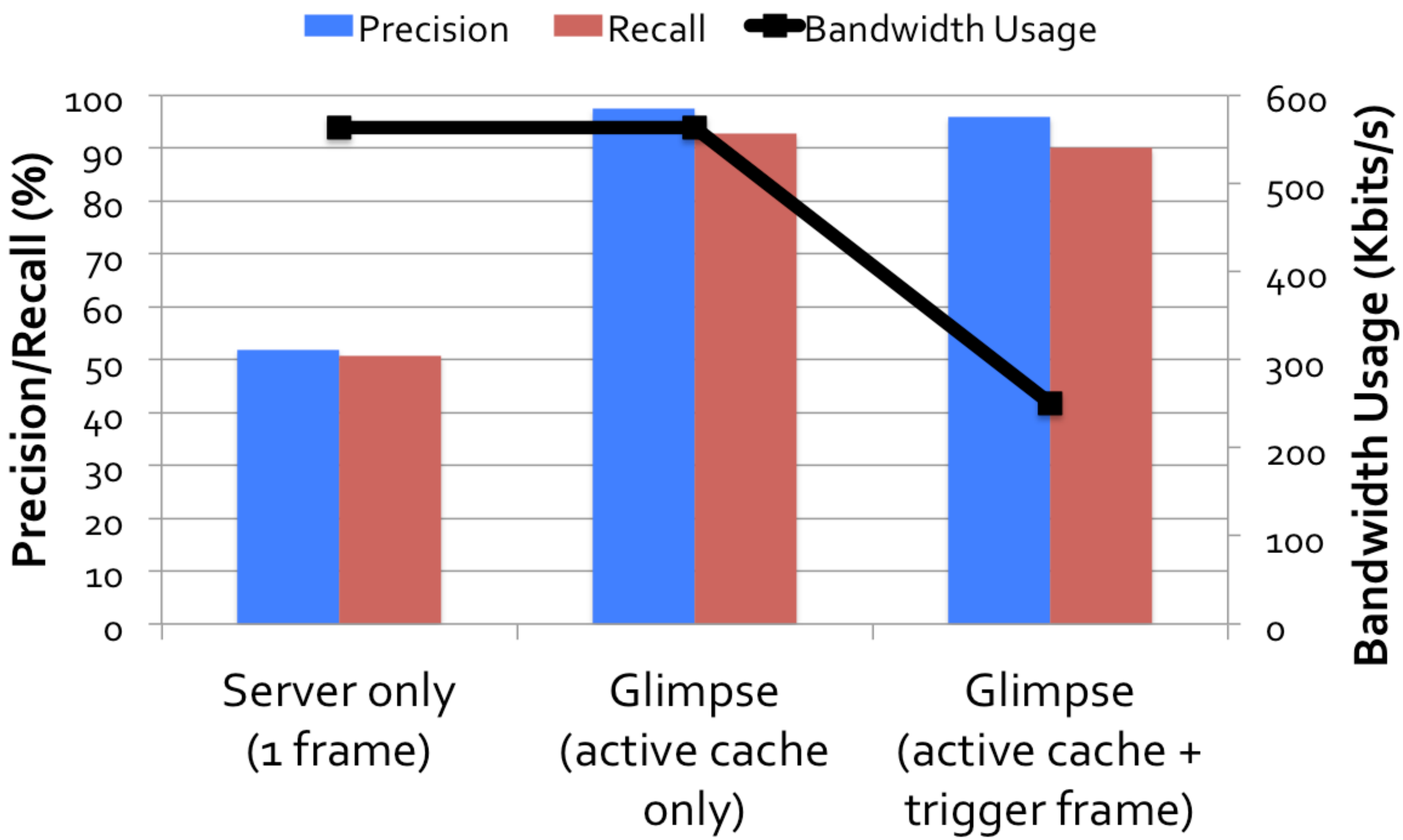
Trigger Frames

- Face dataset
- Wi-Fi (End-to-end delay: 430 ms)



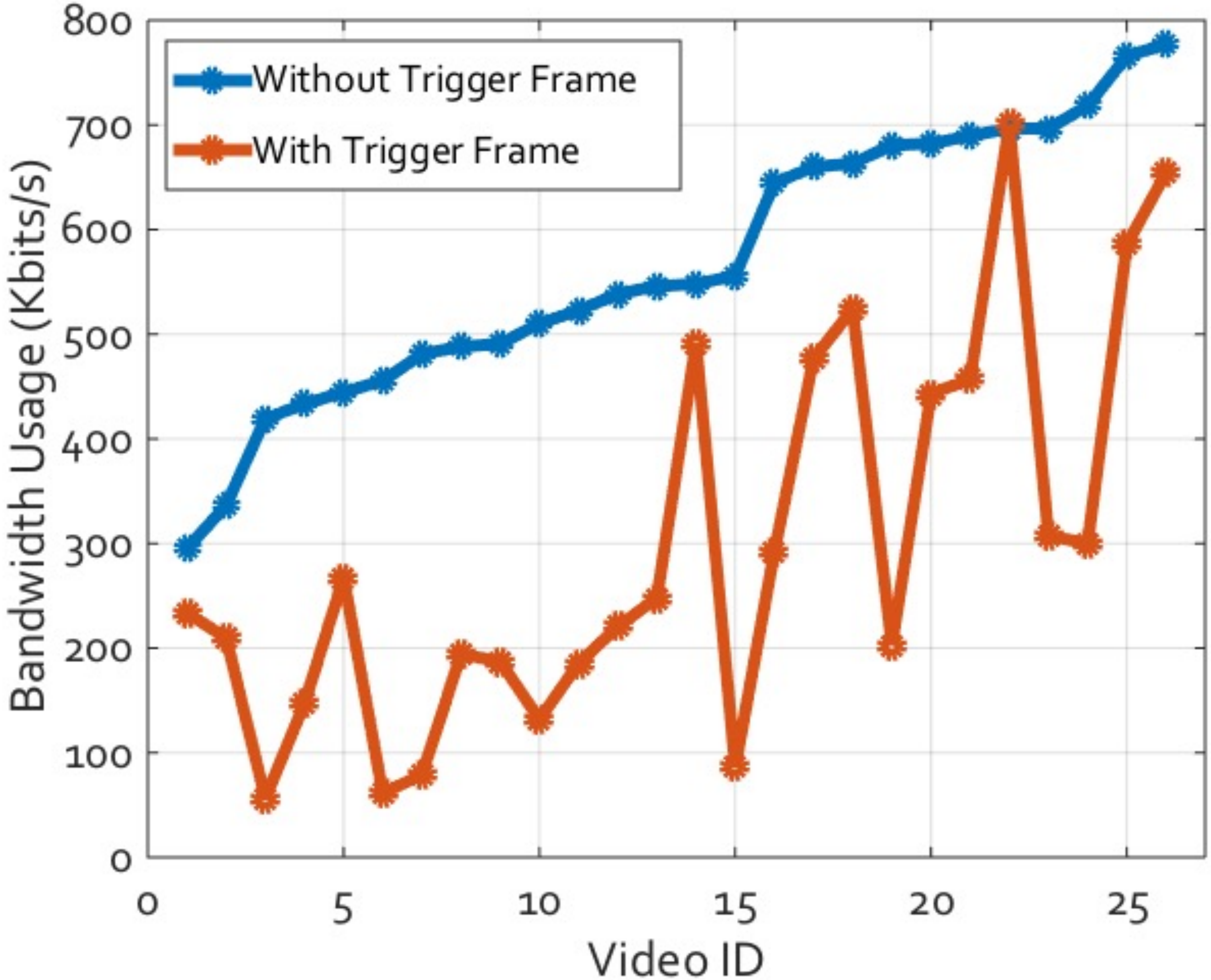
Trigger Frames Reduce Bandwidth Used without Sacrificing Accuracy

- Face dataset
- Wi-Fi (End-to-end delay: 430 ms)



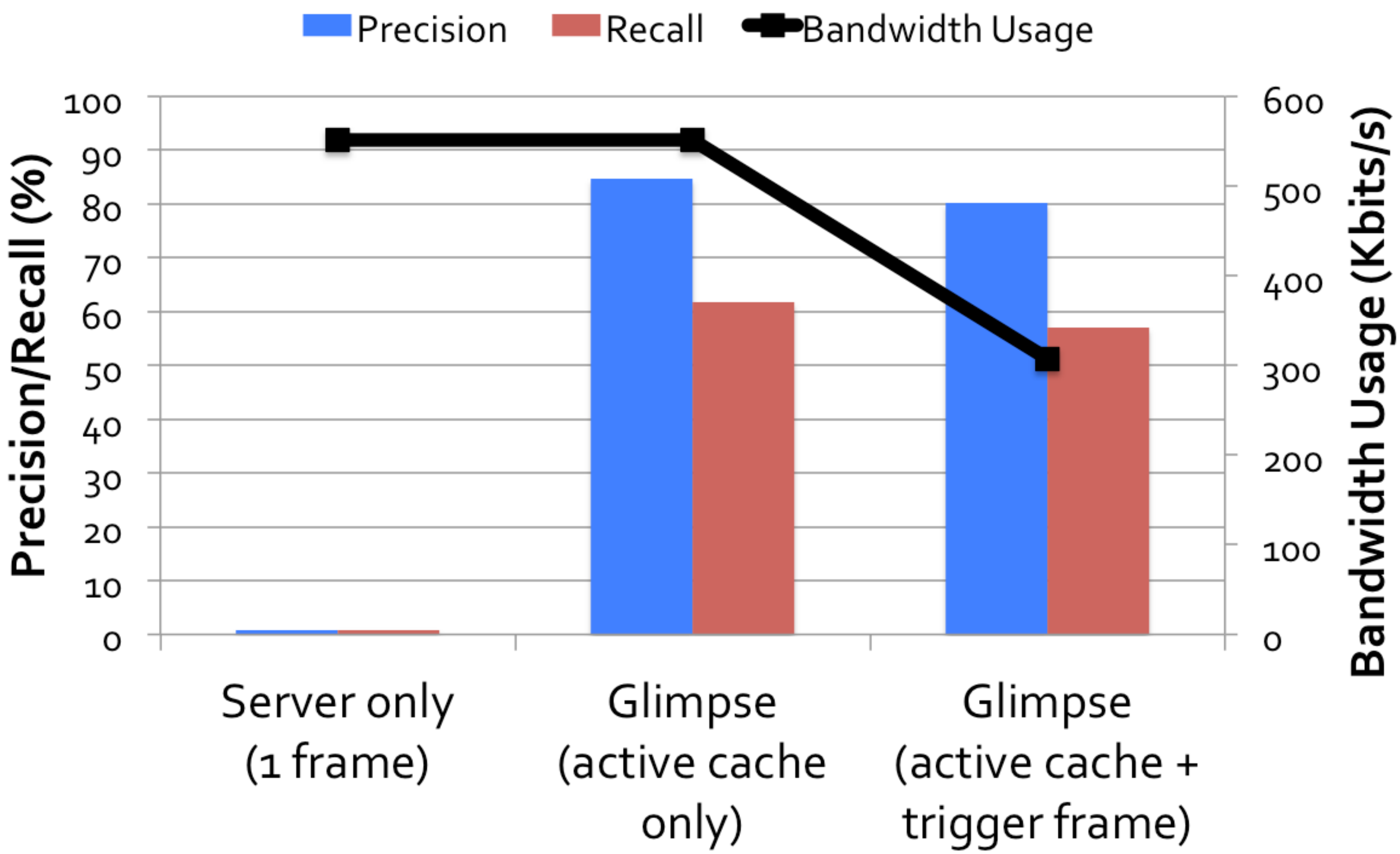
Trigger Frames Reduce Bandwidth Used

- Face Dataset (Wi-Fi)



Glimpse Achieves Higher Accuracy and Lower Bandwidth Usage

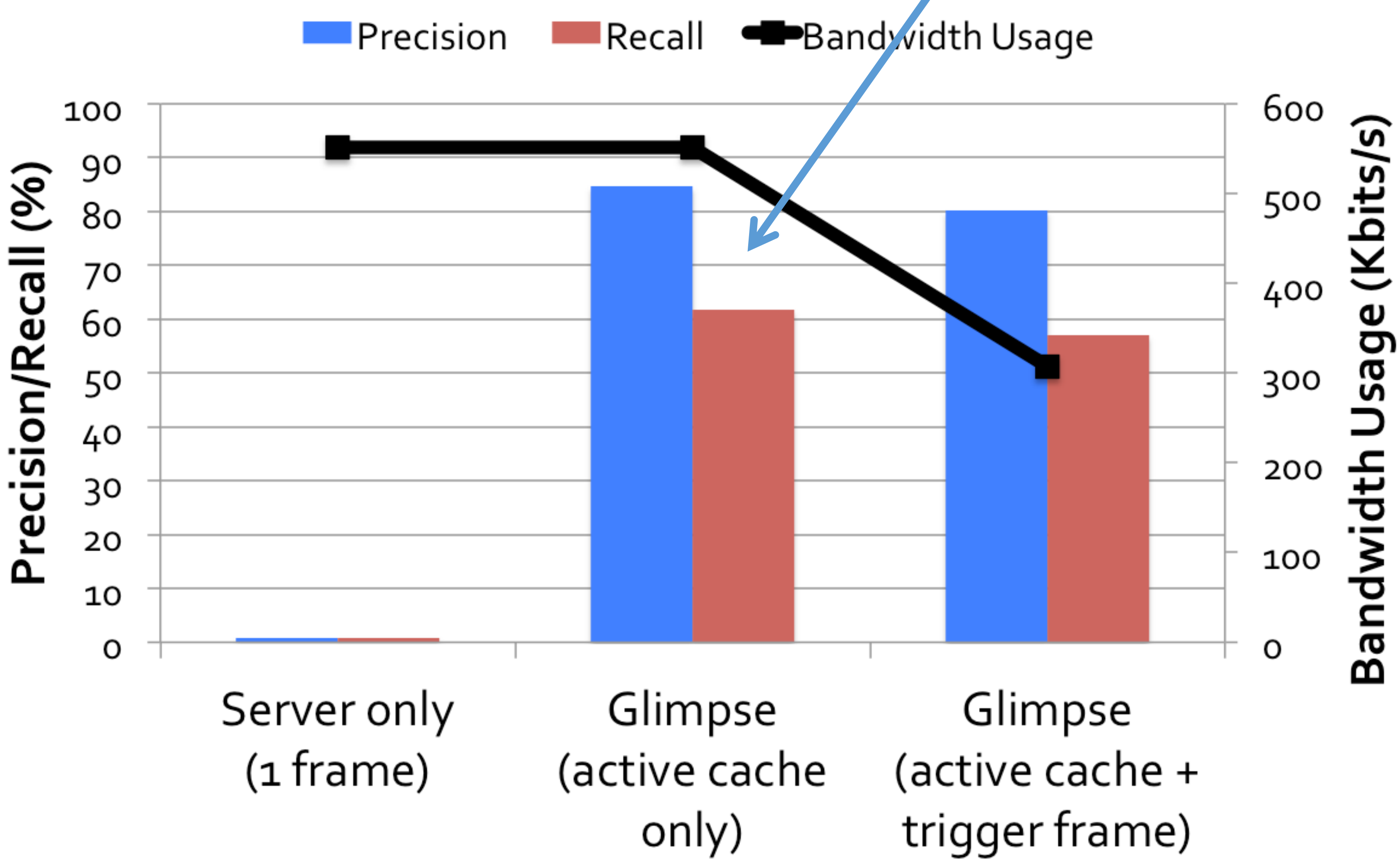
- Road sign dataset
- Wi-Fi (End-to-end delay: 520 ms)



Glimpse Achieves Higher Accuracy and Lower Bandwidth Usage

- Road sign dataset
- Wi-Fi (End-to-end delay: 520 ms)

Recall lower than precision
Why?



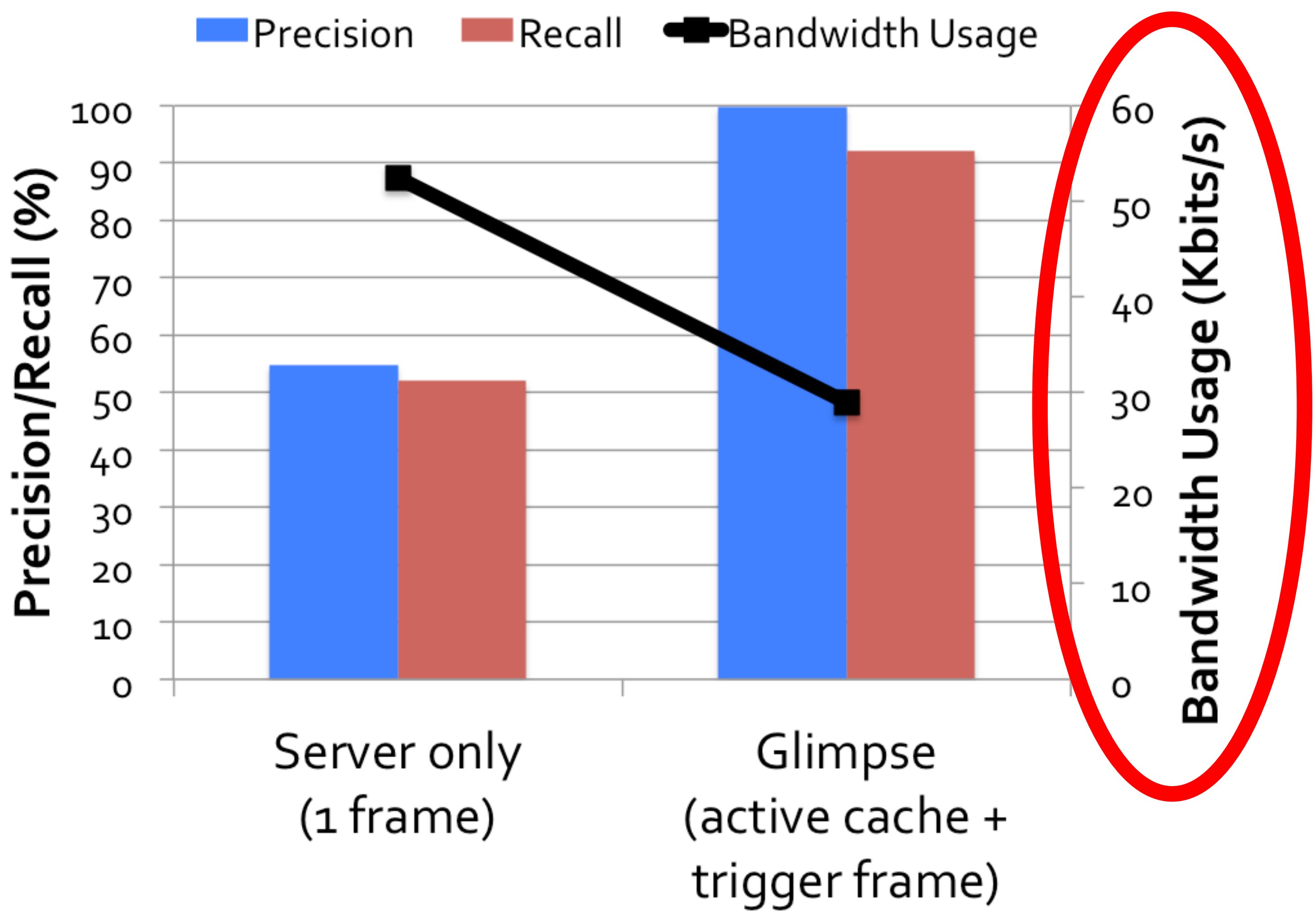
Hardware-Assisted Object Detection

- Mobile devices are now equipped with object detection hardware
- Is Glimpse still helpful?

Glimpse Improves Accuracy even with Detection Hardware on Devices

- Face dataset (Wi-Fi)
- Face detection in hardware

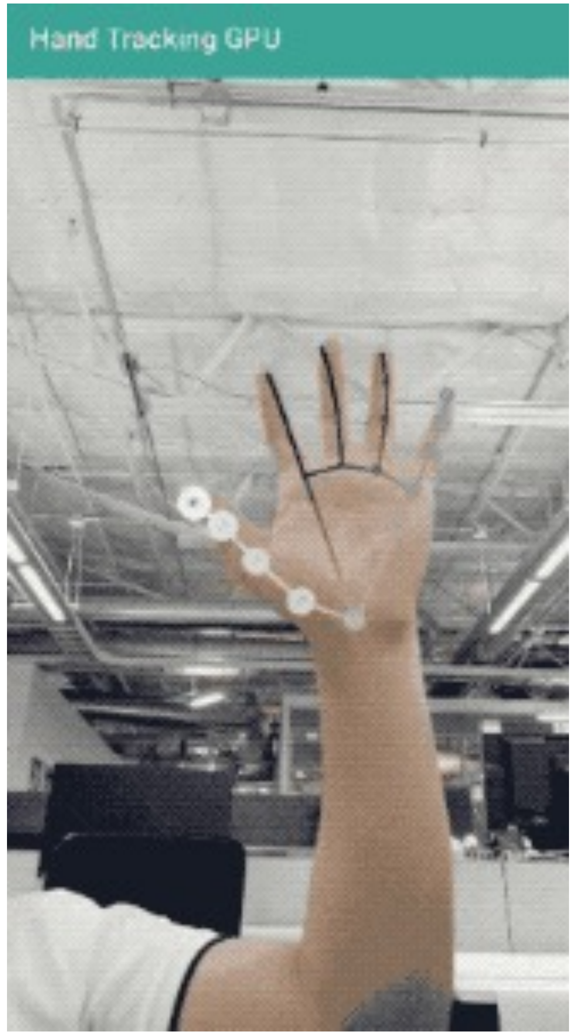
Why lower than before?



Glimpse

- Glimpse enables continuous, real time object recognition on mobile devices
- Glimpse achieves high recognition accuracy by maintaining an *active cache* of frames on the client
- Glimpse reduces bandwidth consumption by strategically sending only certain *trigger frames*

- CoreML (Apple)
- MediaPipe (cross-platform)

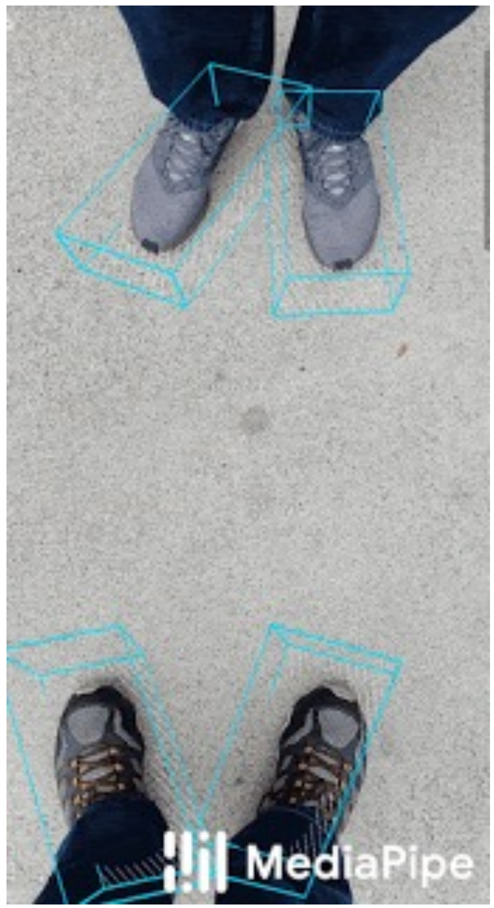


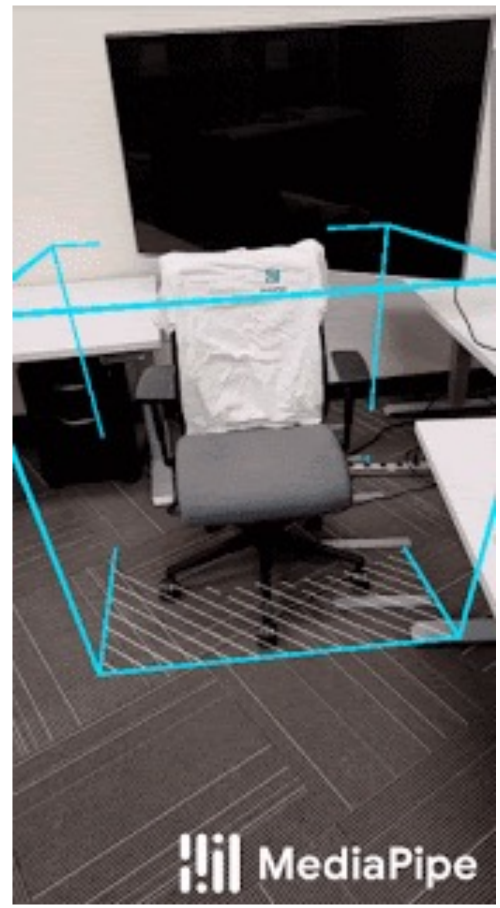
Google AI Blog

The latest news from Google AI

Real-Time 3D Object Detection on Mobile Devices with MediaPipe

Wednesday, March 11, 2020





Live Streaming is Gaining Popularity

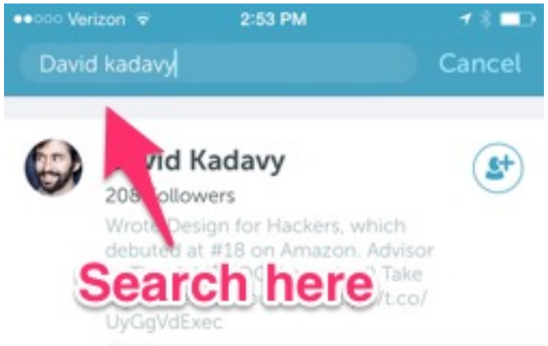
Meerkat
Tweet Live Video

Periscope

Search Capability is Limited

Location

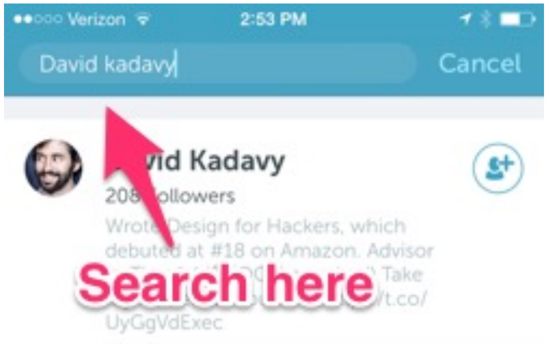
Tags/Username



Search Capability is Limited

Location

Tags/Username



Example Tags:

it ended ☺☺☺

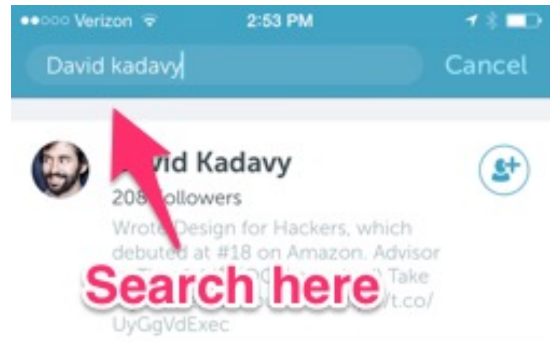
Hola buenas noches

Back at it again ☹☹☹

Search Capability is Limited

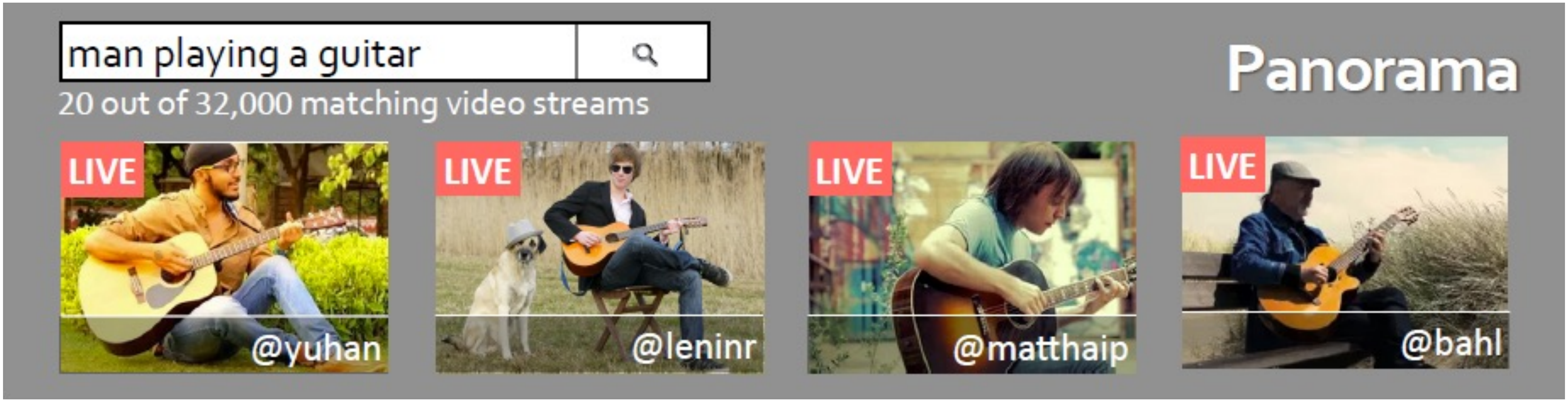
Location

Tags/Username



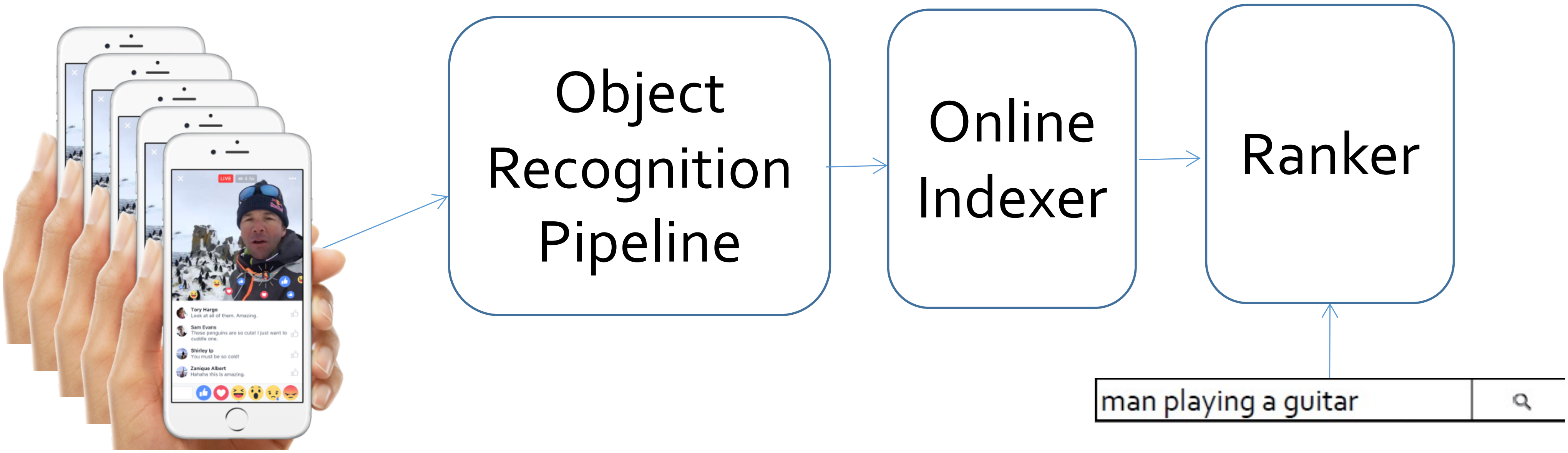
Example Tags:
it ended ☺☺☺
Hola buenas noches
Back at it again ☹☹☹

I want to search based on the contents!

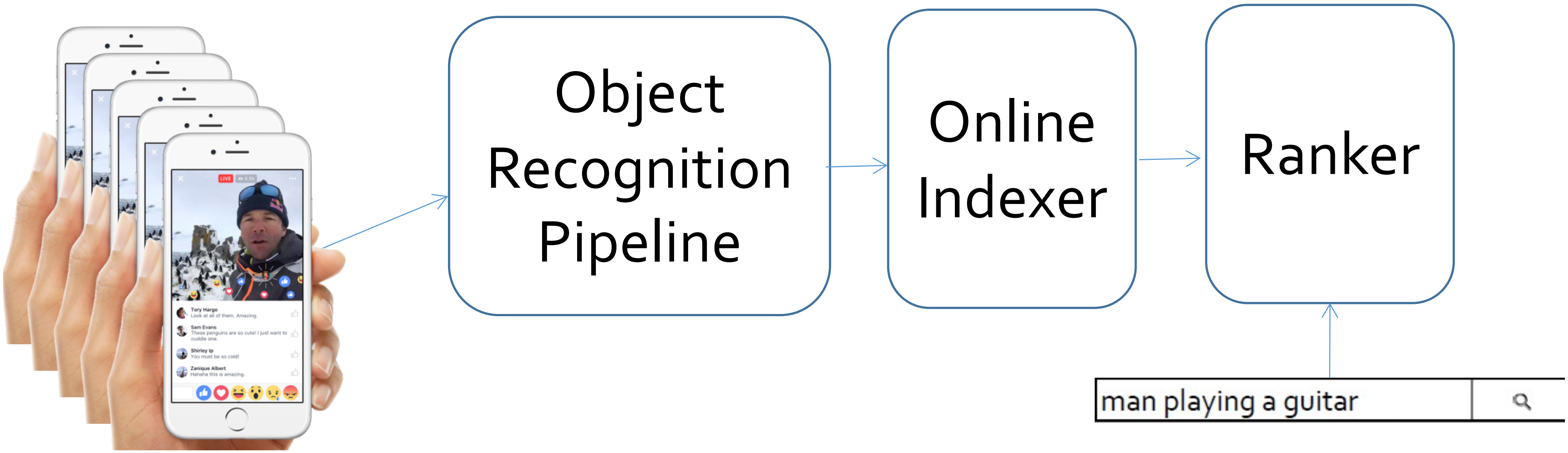


- Why now?
  1. Cameras everywhere
     - dashcams, GoPro, phones
  2. Advances in computer vision
     - CNN
  3. Faster compute
     - GPUs

System Design



Challenges



1. **Scalability**

   - more than 1M incoming streams (30 fps)

2. **Liveliness and Relevance**

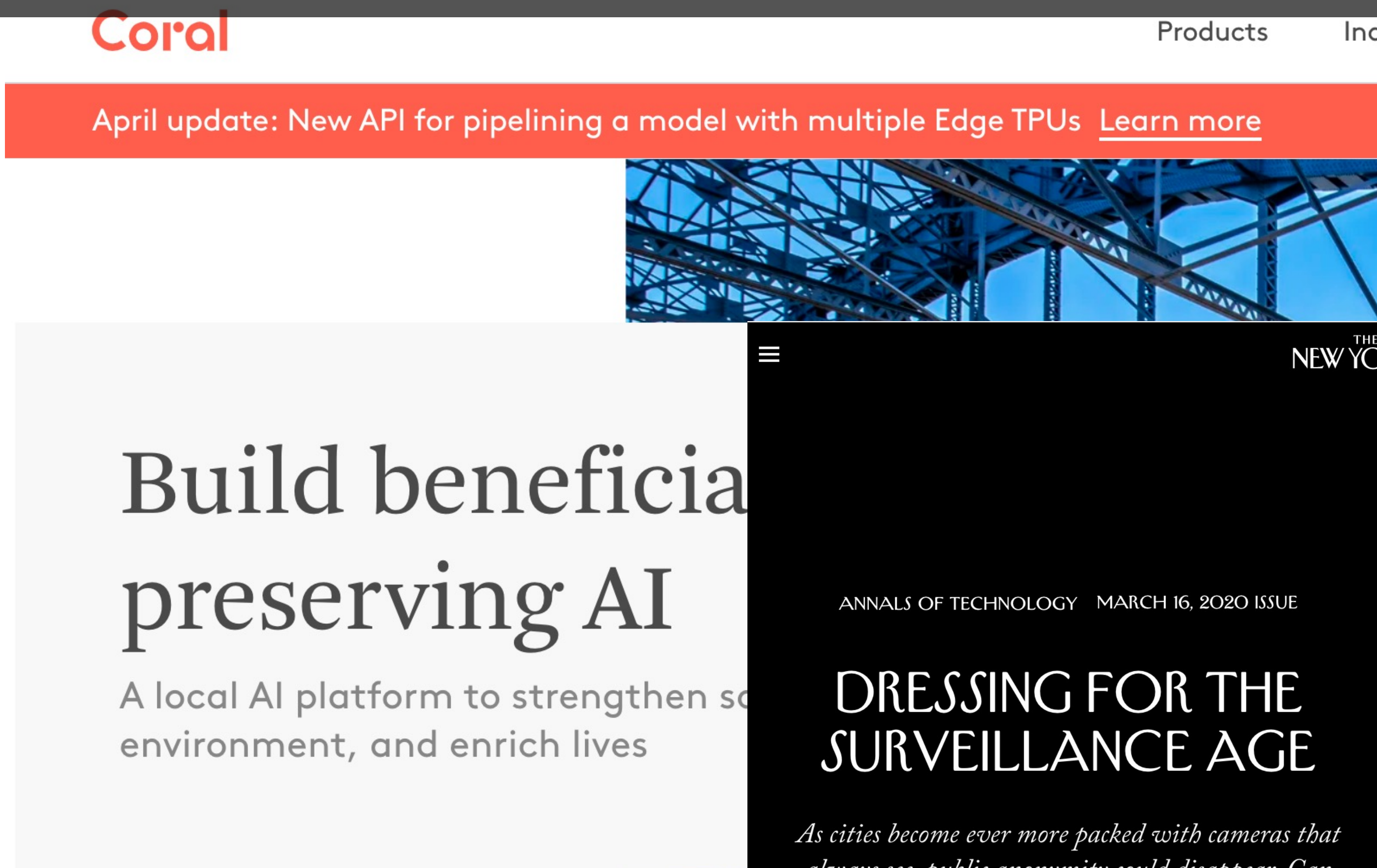
   - contents keep changing – perhaps an incremental online indexer
   - the provided videos match what users want

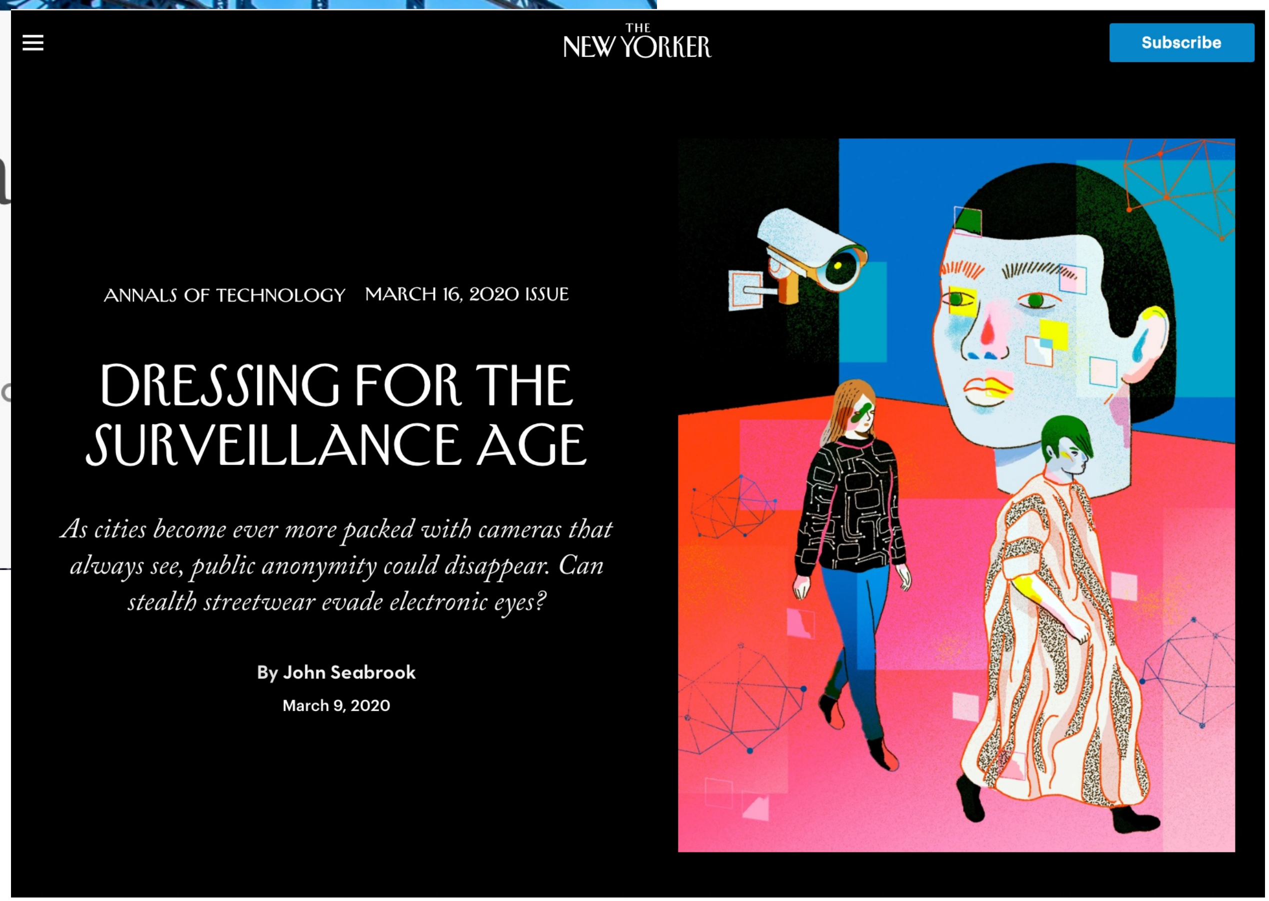
Unleash the Power of Cameras –
What are the possible apps?

- Indoor localization
- Pothole detection
- Activity recognition
- Habitat monitoring
- Agriculture IoT
- Smart city camera networks

- Beyond cameras?
    - RF reconstruction

Concerns


Coral
Products
Ind
April update: New API for pipelining a model with multiple Edge TPUs Learn more
Build beneficia
preserving AI
A local AI platform to strengthen sc
environment, and enrich lives


THE NEW YORKER
Subscribe
ANNALS OF TECHNOLOGY MARCH 16, 2020 ISSUE
DRESSING FOR THE SURVEILLANCE AGE
As cities become ever more packed with cameras that always see, public anonymity could disappear. Can stealth streetwear evade electronic eyes?
By John Seabrook
March 9, 2020