GabriellaV2: UCF DIVA System
Center for Research in Computer Vision (CRCV) at the University of Central Florida (UCF)

Zacchaeus Scheffer, Ishan Dave, Mubarak Shah, Yogesh Rawat

December 8, 2021
Outline
Outline

- Problem Statement
Outline

▶ Problem Statement
▶ New in GabriellaV2
Outline

▶ Problem Statement
▶ New in GabriellaV2
  ▶ Localization
Outline

- Problem Statement
- New in GabriellaV2
  - Localization
  - Action Classification
Outline

▶ Problem Statement
▶ New in GabriellaV2
  ▶ Localization
  ▶ Action Classification
  ▶ Post Processing
Outline

- Problem Statement
- New in GabriellaV2
  - Localization
  - Action Classification
  - Post Processing
- Scores
Outline

► Problem Statement
► New in GabriellaV2
  ▶ Localization
  ▶ Action Classification
  ▶ Post Processing
► Scores
► Qualitative Results
Outline

- Problem Statement
- New in GabriellaV2
  - Localization
  - Action Classification
  - Post Processing
- Scores
- Qualitative Results
- Questions
Problem Statement

Untrimmed Video:
- No special processing
- Should be comparable to "real-world"

Detect activities:
- Human and vehicle
- Indoor and outdoor

Types:
- Single-actor
- Multi-actor
- Actor-object

Output:
- Start/end times
- Spacial location
Problem Statement

Untrimmed Video
Problem Statement

Untrimmed Video

- No special processing
Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"
Problem Statement

Untrimmed Video
  ▶ No special processing
  ▶ Should be comparable to "real-world"

Detect activities
Problem Statement

Untrimmed Video
- No special processing
- Should be comparable to "real-world"

Detect activities
- Human and vehicle
Problem Statement

Untrimmed Video

▶ No special processing
▶ Should be comparable to "real-world"

Detect activities

▶ Human and vehicle
▶ Indoor and outdoor
Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

- Single-actor
- Multi-actor
- Actor-object
Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

- Single-actor
Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

- Single-actor
- Multi-actor
Problem Statement

Untrimmed Video
- No special processing
- Should be comparable to "real-world"

Detect activities
- Human and vehicle
- Indoor and outdoor

Types
- Single-actor
- Multi-actor
- Actor-object
Problem Statement

Untrimmed Video
- No special processing
- Should be comparable to "real-world"

Detect activities
- Human and vehicle
- Indoor and outdoor

Types
- Single-actor
- Multi-actor
- Actor-object

Output
Problem Statement

Untrimmed Video
- No special processing
- Should be comparable to "real-world"

Detect activities
- Human and vehicle
- Indoor and outdoor

Types
- Single-actor
- Multi-actor
- Actor-object

Output
- Start/end times
Problem Statement

Untrimmed Video
- No special processing
- Should be comparable to "real-world"

Detect activities
- Human and vehicle
- Indoor and outdoor

Types
- Single-actor
- Multi-actor
- Actor-object

Output
- Start/end times
- Spacial location
GabriellaV2 System

Untrimmed Video

Clip 1
Clip 2
Clip N

Object Detection

Input Clip

Background Subtractor

Tracklet Generation

Filtered Detected Objects

Object Tracker

Tracklets

id-1
id-2
id-n

Clip 1
Clip 2
Clip N

Activity Tubelets

Tracklets

id=1

id=2

id=n

Action Split

Tubelet Merge

TMAS system

Classified Clips

Classified Tracklets

Classification Network

L_{cls}

Extracted Tracklets

Classified Clips

Classification Block

Predictions

Spatio Temporal Deduplication

Activity Tubes

Action 1

Action 2

Action 3

UCF CENTER FOR RESEARCH IN COMPUTER VISION
PreProcessing

First, we split the video using a sliding window approach. Every 4th frame is used for tracklet generation. The full clip is used in tracklet extraction/classification.
First, we split the video using a sliding window approach.
**PreProcessing**

- First, we split the video using a sliding window approach
  - 0-16
- Every 4th frame used for tracklet generation
- Full clip is used in tracklet extraction/classification
PreProcessing

- First, we split the video using a sliding window approach
  - 0-16
  - 8-24
  - Every 4th frame used for tracklet generation
  - Full clip is used in tracklet extraction/classification
First, we split the video using a sliding window approach:
- 0-16
- 8-24
- 16-32
- Every 4th frame used for tracklet generation
- Full clip is used in tracklet extraction/classification
PreProcessing

- First, we split the video using a sliding window approach
  - 0-16
  - 8-24
  - 16-32
- Every 4th frame used for tracklet generation
PreProcessing

- First, we split the video using a sliding window approach
  - 0-16
  - 8-24
  - 16-32
- Every 4th frame used for tracklet generation
- Full clip is used in tracklet extraction/classification
Tracklet Generation

Tracklet Generation happens in 3 steps:

1. **Object Detector (YOLOv5)**
   - Spacially localizes objects

2. **Background Subtractor (MOG)**
   - Removes stationary objects

3. **Object Tracker (SORT)**
   - Groups detections of the same object
Tracklet Generation

Tracklet Generation happens in 3 steps
Tracklet Generation

Tracklet Generation happens in 3 steps

▶ Object Detector (YOLOv5)
Tracklet Generation

Tracklet Generation happens in 3 steps

▶ Object Detector (YOLOv5)
  ▶ spacially localizes objects
Tracklet Generation

Tracklet Generation happens in 3 steps
► Object Detector (YOLOv5)
  ► spatially localizes objects
► Background Subtractor (MOG)
Tracklet Generation

Tracklet Generation happens in 3 steps

➢ Object Detector (YOLOv5)
  ➢ spacially localizes objects

➢ Background Subtractor (MOG)
  ➢ Removes Stationary Objects
Tracklet Generation

Tracklet Generation happens in 3 steps

▶ Object Detector (YOLOv5)
  ▶ spatially localizes objects
▶ Background Subtractor (MOG)
  ▶ Removes Stationary Objects
▶ Object Tracker (SORT)
Tracklet Generation

Tracklet Generation happens in 3 steps

▶ Object Detector (YOLOv5)
  ▷ spacially localizes objects
▶ Background Subtractor (MOG)
  ▷ Removes Stationary Objects
▶ Object Tracker (SORT)
  ▷ groups detections of the same object
Object Detection

For each Clip

▶ send 4 frames to Object Detector
▶ Place bounding box around potential actor
▶ person
▶ vehicle
Object Detection

For each Clip

- send 4 frames to Object Detector
Object Detection

For each Clip

- send 4 frames to Object Detector
- Place bounding box around potential actor
Object Detection

For each Clip
- send 4 frames to Object Detector
- Place bounding box around potential actor
  - person
Object Detection

For each Clip

- send 4 frames to Object Detector
- Place bounding box around potential actor
  - person
  - vehicle
We use the MOG2 background subtractor, simple to use, to drastically reduce localization false alarm, achieving a 40% FA reduction and runtime reduced by half.
We use the MOG2 background subtractor
We use the MOG2 background subtractor

simple to use
Background Subtraction

- We use the MOG2 background subtractor
- simple to use
- drastically reduce localization false alarm
We use the MOG2 background subtractor
simple to use
drastically reduce localization false alarm
  40% FA reduction
We use the MOG2 background subtractor
- simple to use
- drastically reduce localization false alarm
  - 40% FA reduction
  - Runtime reduced by half
Background Subtractor Results
Object Tracker

- We use the SORT tracker
- IOU-based matching
- Same object gets same ID in subsequent frames
- Can "remember" 2 frames prior
- Track continues if one bounding box is missing
Object Tracker

- We use the SORT tracker
Object Tracker

- We use the SORT tracker
- IOU-based matching
Object Tracker

- We use the SORT tracker
- IOU-based matching
- Same object gets same ID in subsequent frames
Object Tracker

- We use the SORT tracker
- IOU-based matching
- Same object gets same ID in subsequent frames
- Can "remember" 2 frames prior
Object Tracker

- We use the SORT tracker
- IOU-based matching
- Same object gets same ID in subsequent frames
- Can "remember" 2 frames prior
  - Track continues if one bounding box is missing
Recall Analysis of Tracklet Generation

Legacy method used action localization instead of object detection/tracking.

New method attains higher activity recall than legacy method.

22% higher recall using 0.8 IOU threshold.
Recall Analysis of Tracklet Generation

- Legacy method used action localization instead of object detection/tracking.
- New method attains higher activity recall than legacy method.
- 22% higher recall using 0.8 IOU threshold.
Recall Analysis of Tracklet Generation

- Legacy method used action localization instead of object detection/tracking
- New method attains higher activity recall than legacy method

![Graph showing activity recall vs ground-truth spatio-temporal overlap]
Recall Analysis of Tracklet Generation

- Legacy method used action localization instead of object detection/tracking
- New method attains higher activity recall than legacy method
  - 22% higher recall using 0.8 IOU threshold
Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

▶ collect all bounding boxes with that object ID
▶ Combine them by taking the smallest-bounding box
▶ Extend the resulting bounding box in the shorter dimension to obtain a square

classifier gets consistent aspect ratio

Take spatial crop of clip. <- Extracted Tracklet
Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding box
- Extend the resulting bounding box in the shorter dimension to obtain a square
- classifier gets consistent aspect ratio

Extracted Tracklet
Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding bounding box
Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding bounding box
- Extend the resulting bounding box in the shorter dimension to obtain a square
Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding bounding box
- Extend the resulting bounding box in the shorter dimension to obtain a square
  - classifier gets consistent aspect ratio
Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding bounding box
- Extend the resulting bounding box in the shorter dimension to obtain a square
  - classifier gets consistent aspect ratio
- Take spatial crop of clip. <- Extracted Tracklet
Action Classification

Surveillance video activities are multi-label
Ground truth tracks to train 3D-CNN backbone + sigmoid
activation with standard BCE loss

$$\text{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \left[ y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right]$$

Each action class is independent of each other
Action Classification

Surveillance video activities are multi-label

Ground truth tracks to train 3D-CNN backbone + sigmoid activation with standard BCE loss

$L_{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)]$

Each action class is independent of each other
Action Classification

Surveillance video activities are multi-label
Ground truth tracks to train 3D-CNN backbone + sigmoid
activation with standard BCE loss

\[
\text{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \left[ y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right]
\]

Each action class is independent of each other
Action Classification

Surveillance video activities are multi-label
Ground truth tracks to train 3D-CNN backbone + sigmoid
activation with standard BCE loss

\[ \mathcal{L}_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)] \]
Action Classification

Surveillance video activities are multi-label
Ground truth tracks to train 3D-CNN backbone + sigmoid activation with standard BCE loss

\[ \mathcal{L}_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)] \]

Each action class is independent of each other
Adapting to Yolo Tracklets

Need to train classifier to take YOLO tracklets
Adapting to Yolo Tracklets

Need to train classifier to take YOLO tracklets
Class Balanced Training

Some classes are very frequent, some very rare. In the most extreme, a 1:1000 difference was observed.

Used Partial Label Masking (PLM) for class imbalance.
Class Balanced Training

Some classes are very frequent, some very rare
Class Balanced Training

Some classes are very frequent, some very rare
In the most extreme, 1:1000 difference
Class Balanced Training

Some classes are very frequent, some very rare
In the most extreme, 1:1000 difference
Used Partial Label Masking (PLM) for class imbalance
Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.

$L_{LSEP} = \log \left( \frac{1}{\sum_{i \in y} \sum_{j \not\in y} e^{x_j - x_i}} \right)$

Standard LSEP is based on BCE loss. Use LSEP loss with PLM to reweight samples to class balance and learn class correlations together.
Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.
Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.

$$L_{LSEP} = \log \left( 1 + \sum_{i \in y} \sum_{j \notin y} e^{x_j - x_i} \right)$$
Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.

\[ \mathcal{L}_{\text{LSEP}} = \log \left( 1 + \sum_{i \in y} \sum_{j \notin y} e^{x_j - x_i} \right) \]

Standard LSEP is based on BCE loss
Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.

\[ \mathcal{L}_{\text{LSEP}} = \log \left( 1 + \sum_{i \in y} \sum_{j \notin y} e^{x_j - x_i} \right) \]

Standard LSEP is based on BCE loss
Use LSEP loss with PLM to reweight samples to class balance and learn class correlations together
Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:
Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:
▶ standard phase
Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:
▶ standard phase
▶ reduce ensemble size
Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:
▶ standard phase
  ▶ reduce ensemble size
▶ compression phase
Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:

- standard phase
  - reduce ensemble size
- compression phase
  - reduce model size
Knowledge Distillation (Standard)

- Train student model with ensemble
- Loss based on hidden layers
- Loss on raw outputs
- Loss on label prediction
- Val set mAP >4% improvement
- Same mAP as full ensemble
Knowledge Distillation (Standard)

- Train student model with ensemble
Knowledge Distillation (Standard)

- Train student model with ensemble
- Loss based on hidden layers

Val set mAP > 4% improvement

Same mAP as full ensemble
Knowledge Distillation (Standard)

- Train student model with ensemble
- Loss based on hidden layers
- Loss on raw outputs
Knowledge Distillation (Standard)

- Train student model with ensemble
- Loss based on hidden layers
- Loss on raw outputs
- Loss on label prediction
Knowledge Distillation (Standard)

- Train student model with ensemble
- Loss based on hidden layers
- Loss on raw outputs
- Loss on label prediction
- Val set mAP >4% improvement
Knowledge Distillation (Standard)

- Train student model with ensemble
- Loss based on hidden layers
- Loss on raw outputs
- Loss on label prediction
- Val set mAP >4% improvement
- Same mAP as full ensemble
Knowledge Distillation (Model Compression)

Teacher
34-layer R2+1D

Student
irCSN-152

L2 loss for KD

BCE loss from Ground truth label

Backpropagation

val set mAP > 7%

same mAP as full ensemble

loss on raw outputs

loss on label prediction
Knowledge Distillation (Model Compression)

- Student model from before as Teacher

Diagram:
- Teacher: 34-layer R2+1D
- Student: irCSN-152
- L2 loss for KD
- BCE loss from Ground truth label
- Backpropagation
Knowledge Distillation (Model Compression)

- Student model from before as Teacher
- New student is slimmer
Knowledge Distillation (Model Compression)

- Student model from before as Teacher
- New student is slimmer
- Loss on raw outputs

Teacher
34-layer R2+1D

L2 loss for KD

Student
irCSN-152

BCE loss from Ground truth label

Backpropagation

Backpropagation
Knowledge Distillation (Model Compression)

- Student model from before as Teacher
- New student is slimmer
- Loss on raw outputs
- Loss on label prediction
Knowledge Distillation (Model Compression)

- Student model from before as Teacher
- New student is slimmer
- loss on raw outputs
- loss on label prediction
- val set mAP >7% improvement
Knowledge Distillation (Model Compression)

- Student model from before as Teacher
- New student is slimmer
- Loss on raw outputs
- Loss on label prediction
- Val set mAP >7% improvement
- Same mAP as full ensemble
Post-Processing

Post-Processing has two steps:

▶ TMAS
   - smooth detection temporally
   - convert actor tracks into action tubes

▶ NMS
   - Remove duplicate predictions
Post-Processing

Post-Processing has two steps:

- TMAS
- NMS

Remove duplicate predictions
Post-Processing

Post-Processing has two steps:

- TMAS
  - smooth detection temporally

- NMS
  - Remove duplicate predictions
Post-Processing

Post-Processing has two steps:

- TMAS
  - smooth detection temporally
  - convert actor tracks into action tubes

- NMS
  - Remove duplicate predictions
Post-Processing has two steps:

- TMAS
  - smooth detection temporally
  - convert actor tracks into action tubes
- NMS
Post-Processing

Post-Processing has two steps:

- **TMAS**
  - smooth detection temporally
  - convert actor tracks into action tubes
- **NMS**
  - Remove duplicate predictions
TMAS

Chain tracklets with same ID into one actor track

Smooth classwise detection temporally

Create action tubes for each class using (connected) regions of actor tracks where class-scores is high
TMAS

Chain tracklets with same ID into one actor track
Chain tracklets with same ID into one actor track
Smooth classwise detection temporally
Chain tracklets with same ID into one actor track
Smooth classwise detection temporally
Create action tubes for each class using (connected) regions of actor tracks where class-scores is high
NMS

We square our object detections

Very large overlaps

Actions contained in multiple bounding boxes

Multi-actor actions

Need to remove duplicate predictions
NMS

- We square our object detections

- Very large overlaps
- Actions contained in multiple bounding boxes
- Multi-actor actions
- Need to remove duplicate predictions
NMS

- We square our object detections
- Very large overlaps
NMS

- We square our object detections
- Very large overlaps
- Actions contained in multiple bounding boxes
NMS

- We square our object detections
- Very large overlaps
- Actions contained in multiple bounding boxes
- Multi-actor actions

UCF CENTER FOR RESEARCH IN COMPUTER VISION
NMS

- We square our object detections
- Very large overlaps
- Actions contained in multiple bounding boxes
- Multi-actor actions
- Need to remove duplicate predictions
NMS

Need method to work regardless of actor size

\[ d(A, B) = \sqrt[4]{\text{Area}_A \ast \text{Area}(B)} \]

Where \( d_e \) is the Euclidean distance. If actors are close, treat as same action, if VERY close, same actor. False Negatives very rare.
NMS

Need method to work regardless of actor size
use dimensionless distance metric

\[ d(A, B) = \sqrt[4]{\frac{\text{Area}_A \times \text{Area}(B)}{d_e(A, B)}} \]

Where \( d_e \) is the Euclidean distance.

If actors are close, treat as same action, if VERY close, same actor
False Negatives very rare
NMS

Need method to work regardless of actor size
use dimensionless distance metric

\[ d(A, B) = \frac{d_e(A, B)}{\sqrt[4]{\text{Area}_A \ast \text{Area}(B)}} \]
Need method to work regardless of actor size 
use dimensionless distance metric 

\[ d(A, B) = \frac{d_e(A, B)}{\sqrt[4]{AreaA \ast Area(B)}} \]

Where \( d_e \) is the euclidian distance.
NMS

Need method to work regardless of actor size
use dimensionless distance metric

\[ d(A, B) = \frac{d_e(A, B)}{\sqrt[4]{\text{Area}_A \ast \text{Area}(B)}} \]

Where \( d_e \) is the euclidian distance. If actors are close, treat as same action, if VERY close, same actor
NMS

Need method to work regardless of actor size
use dimensionless distance metric

\[ d(A, B) = \frac{d_e(A, B)}{\sqrt[4]{AreaA \ast Area(B)}} \]

Where \( d_e \) is the euclidian distance. If actors are close, treat as same action, if VERY close, same actor
False Negatives very rare
## TrecVID21 Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>team_name</th>
<th>team_abbrev</th>
<th><a href="mailto:nAUDC@0.2fa">nAUDC@0.2fa</a></th>
<th><a href="mailto:p_miss@0.15fa">p_miss@0.15fa</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BUPT-MCPRL</td>
<td>BUPT-MC_26542</td>
<td>0.4085</td>
<td>0.3249</td>
</tr>
<tr>
<td>2</td>
<td>UCF</td>
<td>UCF_26546</td>
<td>0.4306</td>
<td><strong>0.3408</strong></td>
</tr>
<tr>
<td>3</td>
<td>INF</td>
<td>INF_26532</td>
<td>0.4444</td>
<td>0.3508</td>
</tr>
<tr>
<td>4</td>
<td>M4D_2021</td>
<td>M4D_202_2646</td>
<td>0.8466</td>
<td>0.7941</td>
</tr>
<tr>
<td>5</td>
<td>TokyoTech_AIST</td>
<td>TOKYOTE_26508</td>
<td>0.8516</td>
<td>0.8197</td>
</tr>
<tr>
<td>6</td>
<td>Team UEC</td>
<td>TEAMUE_26530</td>
<td>0.964</td>
<td>0.9503</td>
</tr>
</tbody>
</table>
## TrecVID21 Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>team_name</th>
<th>team_abbrev</th>
<th><a href="mailto:nAUDC@0.2tf">nAUDC@0.2tf</a></th>
<th><a href="mailto:p_miss@0.15t">p_miss@0.15t</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BUPT-MCPRL</td>
<td>BUPT-MC_26542</td>
<td>0.4085</td>
<td>0.3249</td>
</tr>
<tr>
<td>2</td>
<td>UCF</td>
<td>UCF_26546</td>
<td>0.4306</td>
<td><strong>0.3408</strong></td>
</tr>
<tr>
<td>3</td>
<td>INF</td>
<td>INF_26532</td>
<td>0.4444</td>
<td>0.3508</td>
</tr>
<tr>
<td>4</td>
<td>M4D_2021</td>
<td>M4D_202_26467</td>
<td>0.8466</td>
<td>0.7941</td>
</tr>
<tr>
<td>5</td>
<td>TokyoTech_AIST</td>
<td>TOKYOTE_26508</td>
<td>0.8516</td>
<td>0.8197</td>
</tr>
<tr>
<td>6</td>
<td>Team UEC</td>
<td>TEAMUE_26530</td>
<td>0.964</td>
<td>0.9503</td>
</tr>
</tbody>
</table>
### TrecVID21 Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>team_name</th>
<th>team_abbrev</th>
<th><a href="mailto:nAUDC@0.2tf">nAUDC@0.2tf</a>α</th>
<th><a href="mailto:p_miss@0.15tfa">p_miss@0.15tfa</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BUPT-MCPRL</td>
<td>BUPT-MC_26542</td>
<td>0.4085</td>
<td>0.3249</td>
</tr>
<tr>
<td>2</td>
<td>UCF</td>
<td>UCF_26546</td>
<td>0.4306</td>
<td><strong>0.3408</strong></td>
</tr>
<tr>
<td>3</td>
<td>INF</td>
<td>INF_26532</td>
<td>0.4444</td>
<td>0.3508</td>
</tr>
<tr>
<td>4</td>
<td>M4D_2021</td>
<td>M4D_202_26467</td>
<td>0.8466</td>
<td>0.7941</td>
</tr>
<tr>
<td>5</td>
<td>TokyoTech_AIST</td>
<td>TOKYOTEM_26508</td>
<td>0.8516</td>
<td>0.8197</td>
</tr>
<tr>
<td>6</td>
<td>Team UEC</td>
<td>TEAMUE_26530</td>
<td>0.964</td>
<td>0.9503</td>
</tr>
</tbody>
</table>

- Based on VIRAT dataset
- Has known cameras
### TrecVID21 Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>team_name</th>
<th>team_abbrev</th>
<th><a href="mailto:nAUDC@0.2ta">nAUDC@0.2ta</a></th>
<th><a href="mailto:p_miss@0.15tfa">p_miss@0.15tfa</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BUPT-MCPRL</td>
<td>BUPT-MC_26542</td>
<td>0.4085</td>
<td>0.3249</td>
</tr>
<tr>
<td>2</td>
<td>UCF</td>
<td>UCF_26546</td>
<td>0.4306</td>
<td><strong>0.3408</strong></td>
</tr>
<tr>
<td>3</td>
<td>INF</td>
<td>INF_26532</td>
<td>0.4444</td>
<td>0.3508</td>
</tr>
<tr>
<td>4</td>
<td>M4D_2021</td>
<td>M4D_202_2646</td>
<td>0.8466</td>
<td>0.7941</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>TokyoTech_AIST</td>
<td>TOKYOTE_26508</td>
<td>0.8516</td>
<td>0.8197</td>
</tr>
<tr>
<td>6</td>
<td>Team UEC</td>
<td>TEAMUE_26530</td>
<td>0.964</td>
<td>0.9503</td>
</tr>
</tbody>
</table>

- based on VIRAT dataset
- has known cameras
- 2nd place
## SDL21 Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>sub_id</th>
<th>mean <a href="mailto:p_miss@0.01tfa">p_miss@0.01tfa</a></th>
<th>mean <a href="mailto:p_miss@0.02tfa">p_miss@0.02tfa</a></th>
<th>mean <a href="mailto:nAUDC@0.2tfa">nAUDC@0.2tfa</a></th>
<th>relative processing time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCF</td>
<td>25908</td>
<td>0.62</td>
<td>0.5372</td>
<td>0.3518</td>
<td>0.684</td>
</tr>
<tr>
<td>2</td>
<td>CMU-DIVA</td>
<td>26095</td>
<td>0.65</td>
<td>0.5438</td>
<td>0.333</td>
<td>0.776</td>
</tr>
<tr>
<td>3</td>
<td>IBM-Purdue</td>
<td>26113</td>
<td>0.65</td>
<td>0.5531</td>
<td>0.3533</td>
<td>0.575</td>
</tr>
<tr>
<td>4</td>
<td>UMD</td>
<td>26619</td>
<td>0.68</td>
<td>0.5938</td>
<td>0.3898</td>
<td>0.515</td>
</tr>
<tr>
<td>5</td>
<td>UMD-Columbia</td>
<td>25031</td>
<td>0.68</td>
<td>0.5975</td>
<td>0.4002</td>
<td>0.52</td>
</tr>
<tr>
<td>6</td>
<td>UMCMU</td>
<td>25576</td>
<td>0.75</td>
<td>0.6861</td>
<td>0.4922</td>
<td>0.614</td>
</tr>
<tr>
<td>7</td>
<td>Purdue</td>
<td>25782</td>
<td>0.8</td>
<td>0.7294</td>
<td>0.4942</td>
<td>0.239</td>
</tr>
<tr>
<td>8</td>
<td>MINDS_JHU</td>
<td>24666</td>
<td>0.84</td>
<td>0.7791</td>
<td>0.6343</td>
<td>0.898</td>
</tr>
</tbody>
</table>
### SDL21 Results

Based on MEVA dataset

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>sub_id</th>
<th><a href="mailto:mean_p_miss@0.01tfa">mean_p_miss@0.01tfa</a></th>
<th><a href="mailto:mean_p_miss@0.02tfa">mean_p_miss@0.02tfa</a></th>
<th><a href="mailto:mean_nAUDC@0.2tfa">mean_nAUDC@0.2tfa</a></th>
<th>relative_processing_time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCF</td>
<td>25908</td>
<td>0.62</td>
<td>0.5372</td>
<td>0.3518</td>
<td>0.684</td>
</tr>
<tr>
<td>2</td>
<td>CMU-DIVA</td>
<td>26095</td>
<td>0.65</td>
<td>0.5438</td>
<td>0.333</td>
<td>0.776</td>
</tr>
<tr>
<td>3</td>
<td>IBM-Purdue</td>
<td>26113</td>
<td>0.65</td>
<td>0.5531</td>
<td>0.3533</td>
<td>0.575</td>
</tr>
<tr>
<td>4</td>
<td>UMD</td>
<td>26619</td>
<td>0.68</td>
<td>0.5938</td>
<td>0.3898</td>
<td>0.515</td>
</tr>
<tr>
<td>5</td>
<td>UMD-Columbia</td>
<td>25031</td>
<td>0.68</td>
<td>0.5975</td>
<td>0.4002</td>
<td>0.52</td>
</tr>
<tr>
<td>6</td>
<td>UMCMU</td>
<td>25576</td>
<td>0.75</td>
<td>0.6861</td>
<td>0.4922</td>
<td>0.614</td>
</tr>
<tr>
<td>7</td>
<td>Purdue</td>
<td>25782</td>
<td>0.8</td>
<td>0.7294</td>
<td>0.4942</td>
<td>0.239</td>
</tr>
<tr>
<td>8</td>
<td>MINDS JHU</td>
<td>24666</td>
<td>0.84</td>
<td>0.7791</td>
<td>0.6343</td>
<td>0.898</td>
</tr>
</tbody>
</table>
## SDL21 Results

- Based on MEVA dataset
- Has unknown cameras

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>sub_id</th>
<th>mean <a href="mailto:p_miss@0.01tfa">p_miss@0.01tfa</a></th>
<th>mean <a href="mailto:p_miss@0.02tfa">p_miss@0.02tfa</a></th>
<th>mean <a href="mailto:nAUDC@0.2tfa">nAUDC@0.2tfa</a></th>
<th>relative processing time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCF</td>
<td>25908</td>
<td>0.62</td>
<td>0.5372</td>
<td>0.3518</td>
<td>0.684</td>
</tr>
<tr>
<td>2</td>
<td>CMU-DIVA</td>
<td>26095</td>
<td>0.65</td>
<td>0.5438</td>
<td>0.333</td>
<td>0.776</td>
</tr>
<tr>
<td>3</td>
<td>IBM-Purdue</td>
<td>26113</td>
<td>0.65</td>
<td>0.5531</td>
<td>0.3533</td>
<td>0.575</td>
</tr>
<tr>
<td>4</td>
<td>UMD</td>
<td>26619</td>
<td>0.68</td>
<td>0.5938</td>
<td>0.3898</td>
<td>0.515</td>
</tr>
<tr>
<td>5</td>
<td>UMD-Columbia</td>
<td>25031</td>
<td>0.68</td>
<td>0.5975</td>
<td>0.4002</td>
<td>0.52</td>
</tr>
<tr>
<td>6</td>
<td>UMCMU</td>
<td>25576</td>
<td>0.75</td>
<td>0.6861</td>
<td>0.4922</td>
<td>0.614</td>
</tr>
<tr>
<td>7</td>
<td>Purdue</td>
<td>25782</td>
<td>0.8</td>
<td>0.7294</td>
<td>0.4942</td>
<td>0.239</td>
</tr>
<tr>
<td>8</td>
<td>MINDS_JHU</td>
<td>24666</td>
<td>0.84</td>
<td>0.7791</td>
<td>0.6343</td>
<td>0.898</td>
</tr>
</tbody>
</table>
## SDL21 Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>sub_id</th>
<th>mean <a href="mailto:p_miss@0.01tfa">p_miss@0.01tfa</a></th>
<th>mean <a href="mailto:p_miss@0.02tfa">p_miss@0.02tfa</a></th>
<th>mean <a href="mailto:nAUDC@0.2tfa">nAUDC@0.2tfa</a></th>
<th>relative processing time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCF</td>
<td>25908</td>
<td>0.62</td>
<td>0.5372</td>
<td>0.3518</td>
<td>0.684</td>
</tr>
<tr>
<td>2</td>
<td>CMU-DIVA</td>
<td>26095</td>
<td>0.65</td>
<td>0.5438</td>
<td>0.333</td>
<td>0.776</td>
</tr>
<tr>
<td>3</td>
<td>IBM-Purdue</td>
<td>26113</td>
<td>0.65</td>
<td>0.5531</td>
<td>0.3533</td>
<td>0.575</td>
</tr>
<tr>
<td>4</td>
<td>UMD</td>
<td>26619</td>
<td>0.68</td>
<td>0.5938</td>
<td>0.3898</td>
<td>0.515</td>
</tr>
<tr>
<td>5</td>
<td>UMD-Columbia</td>
<td>25031</td>
<td>0.68</td>
<td>0.5975</td>
<td>0.4002</td>
<td>0.52</td>
</tr>
<tr>
<td>6</td>
<td>UMCMU</td>
<td>25576</td>
<td>0.75</td>
<td>0.6861</td>
<td>0.4922</td>
<td>0.614</td>
</tr>
<tr>
<td>7</td>
<td>Purdue</td>
<td>25782</td>
<td>0.8</td>
<td>0.7294</td>
<td>0.4942</td>
<td>0.239</td>
</tr>
<tr>
<td>8</td>
<td>MINDS_JHU</td>
<td>24666</td>
<td>0.84</td>
<td>0.7791</td>
<td>0.6343</td>
<td>0.898</td>
</tr>
</tbody>
</table>

- based on MEVA dataset
- has unknown cameras
- 1st place in UF
Generalization

- Drop in performance going from Known Facility to Unknown
- Our system gets best generalization
Generalization

- Drop in performance going from Known Facility to Unknown
Generalization

- Drop in performance going from Known Facility to Unknown
- Our system gets best generalization
Qualitative Results

Many common situations that hurt performance
Qualitative Results

Many common situations that hurt performance

▶ Distant actors
Qualitative Results

Many common situations that hurt performance

► Distant actors

► Actor changes distance over time
Qualitative Results

Many common situations that hurt performance

▶ Distant actors
▶ Actor changes distance over time
▶ Temporal variability
Qualitative Results

Many common situations that hurt performance

- Distant actors
- Actor changes distance over time
- Temporal variability
- Crowded Scenes
Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras.
Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras. Built upon our Gabriella system by

- Replacing localization by tracklet generation,
- Better classifiers: class imbalance, multi-label class correlation, Knowledge distillation
- Improved Post processing using Spatio-temporal deduplication
- Top place for ActEV-SDL21 UF leaderboard and Runners-up for TRECVID21
Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras. Built upon our Gabriella system by

- Replacing localization by tracklet generation,
Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras. Built upon our Gabriella system by

- Replacing localization by tracklet generation,
- Better classifiers: class imbalance, multi-label class correlation, Knowledge distillation
Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras. Built upon our Gabriella system by

- Replacing localization by tracklet generation,
- Better classifiers: class imbalance, multi-label class correlation, Knowledge distillation
- Improved Post processing using Spatio-temporal deduplication
Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras. Built upon our Gabriella system by

- Replacing localization by tracklet generation,
- Better classifiers: class imbalance, multi-label class correlation, Knowledge distillation
- Improved Post processing using Spatio-temporal deduplication
- Top place for ActEV-SDL21 UF leaderboard and Runners-up for TRECVID21