Disaster Scene Description and Indexing (DSDI) Task Overview

Asad Anwar Butt, National Institute of Standards and Technology; Johns Hopkins University
George Awad, National Institute of Standards and Technology
Jeffrey Liu, MIT Lincoln Laboratory
William Drew, Office of Homeland Security and Preparedness

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

TRECVID 2022

Introduction - DSDI

- Video and imagery data can be extremely helpful for public safety operations.
- Natural Disasters, e.g.,
  - Wildfires
  - Hurricanes
  - Earthquakes
  - Floods
- Man-made Disasters, e.g.,
  - Hazardous material spills
  - Mining accidents
  - Explosions

Introduction - DSDI

- Prior knowledge about affected areas can be very useful for the first responders.
- Oftentimes, the communication systems go down in major disasters, which makes it very difficult to get any information regarding the damage.
- Automated systems to gather information before rescue workers enter the area can be very helpful.

Introduction - DSDI

- Computer vision capabilities have rapidly advanced recently with the popularity of deep learning.
  - Research groups have access to large image and video datasets for various tasks.
- However, the capabilities do not meet public safety needs.
  - Lack of relevant training data.
- Most current image and video datasets have no public safety hazard labels.
  - State-of-the-art systems trained on such datasets fail to provide helpful labels.

Training Dataset

NIST

- In response, the MIT Lincoln Lab developed a dataset of images collected by the Civil Air Patrol of various natural disasters.
- The Low Altitude Disaster Imagery (LADI) dataset was developed as part of a larger NIST Public Safety Innovator Accelerator Program (PSIAP) grant.
- Two key properties of the dataset are:
  - Low altitude
  - Oblique perspective of the imagery and disaster-related features.
- The DSDI test data and ground truth from 2020 & 2021 are also available for teams to use as training data.

Training Dataset

- LADI Dataset:
  - Hosted as part of the AWS Public Dataset program.
  - Consists of 20,000+ annotated images.
  - The images are from locations with FEMA (Federal Emergency Management Agency) major disaster declarations for a hurricane or flooding.
  - Lower altitude criteria distinguish the LADI dataset from satellite datasets to support the development of computer vision capabilities with small drones operating at low altitudes.
  - A minimum image size was selected to maximize the efficiency of the crowd source workers; lower resolution images are harder to annotate.
- 2020 - 2021 DSDI Test Set:
  - ~ 11 hours of video.
  - Segmented into small video clips (shots) of maximum 20 sec.
  - Videos are from earthquake, hurricane, and flood affected areas.
  - Total number of shots: 4626

Test Dataset

- A test dataset of about 6 hours of video was distributed for this task.
- Collected by FEMA as individual images after disaster events.
- Individual images were stitched to form videos with reasonable speed.
- The test dataset was segmented into small video clips (shots) of a maximum of 16 sec, with a mean length of 10 sec.
- Subset was selected by NIST taking into account diversity
- In addition, a small set of videos was collected from Defense Visual Information Distribution Service (DVIDS): https://www.dvidshub.net/
- Total number of shots: 2157

Testing Data: Example Videos

* All videos are under public domain licenses

Testing Dataset - Categories

NIST

- Hierarchical labeling scheme: 5 coarse categories, each with 4 to 9 more specific annotations.

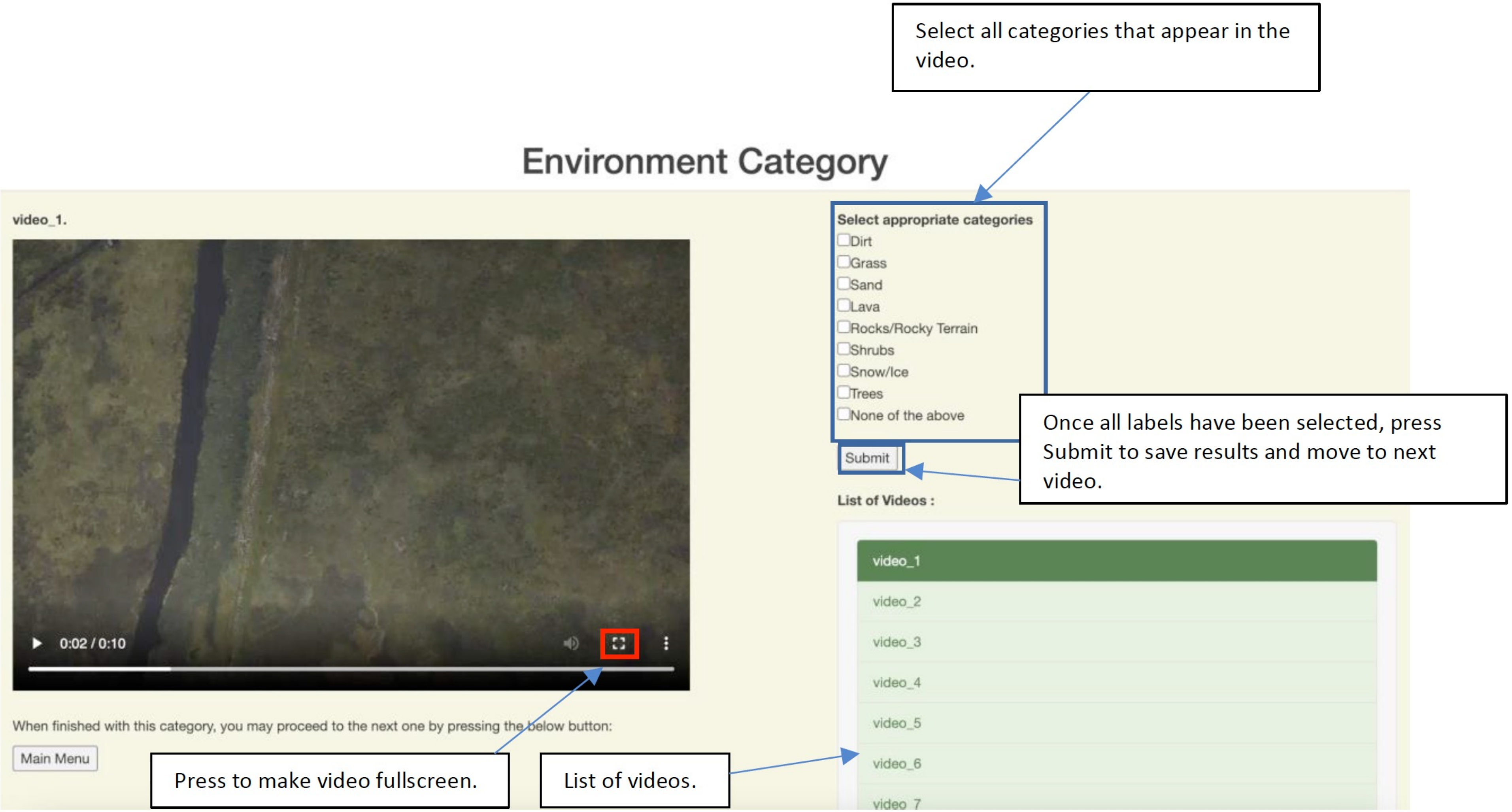
Damage	Environment	Infrastructure	Vehicles	Water
Misc. Damage	Dirt	Bridge	Aircraft	Flooding
Flooding/Water Damage	Grass	Building	Boat	Lake/Pond
Landslide	Lava	Dam/Levee	Car	Ocean
Road Washout	Rocks	Pipes	Truck	Puddle
Rubble/Debris	Sand	Utility Or Power Lines/Electric Towers		River/Stream
Smoke/Fire	Shrubs	Railway		
	Snow/Ice	Wireless/Radio Communication Towers		
	Trees	Water Tower		
		Road		

Annotation

NIST

- We had 2 full time annotators (**same** annotators since 2020) instead of crowdsourcing.
- For each category, a practice web page was created with multiple examples and sample test videos.
- This allowed the annotators to become familiarized with the task and labels before starting a category.
- The annotators worked independently on each category.
- For each coarse category, they marked all the specific labels that were present in the video.
- To create the final ground truth, for each shot, the union of labels was used.

Annotation Tool



The annotators watch the video and mark the categories that are visible in the video.

DSDI System Task

- Systems are required to return a ranked list of up to 1000 shots for each of the 32 features.
- Each submitted run specified its training type:
  - LADI-based (L): The run only used the supplied LADI dataset for development of its system.
  - Non-LADI (N): The run did not use the LADI dataset, but only trained using other dataset(s).
  - LADI + Others (O): The run used the LADI dataset in addition to any other dataset(s) for training purposes.

Evaluation Metrics

NIST

- The following evaluation metrics were used to compare the submissions:

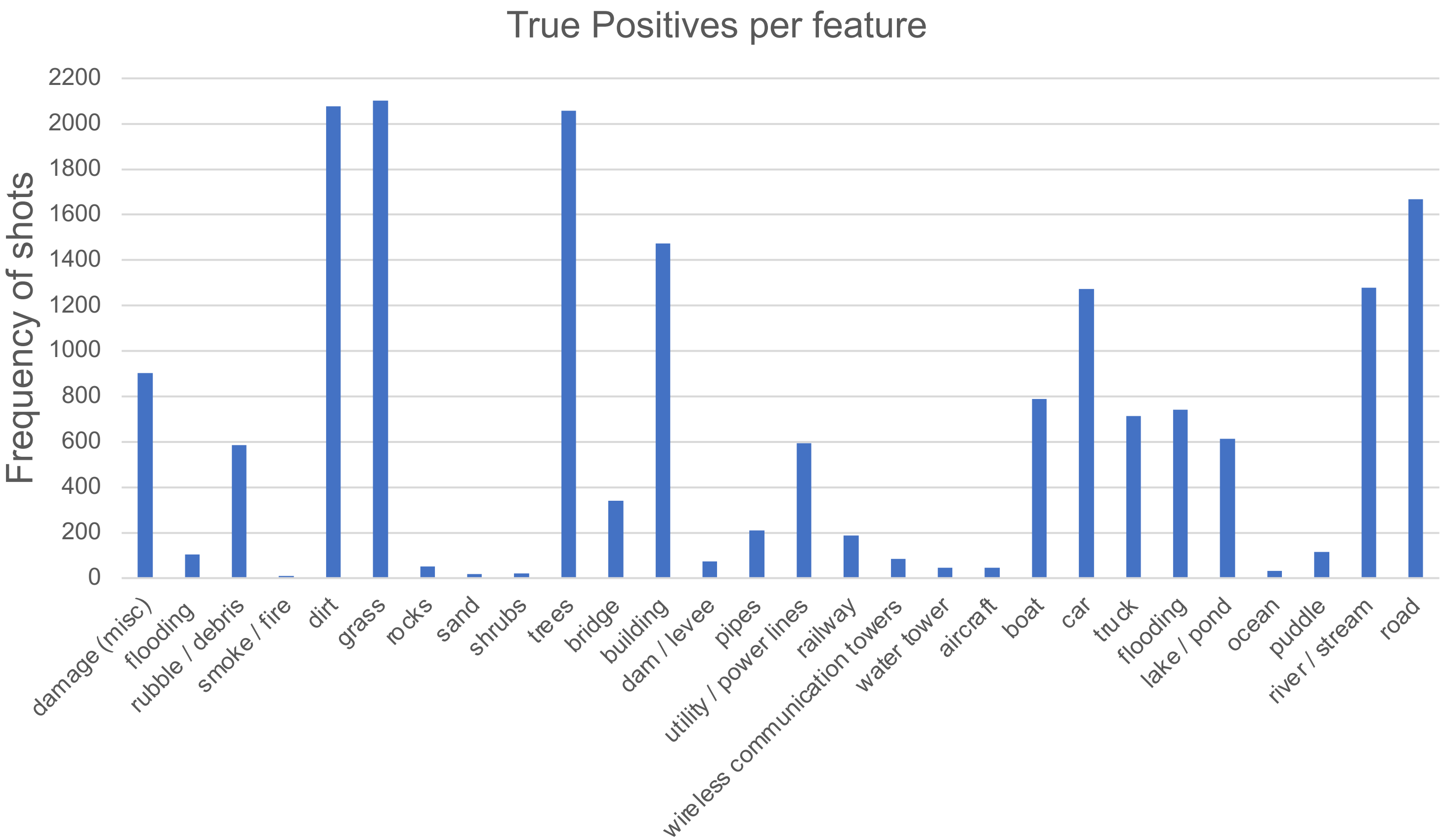
Metric	Description
Speed	Clock time per inference (reported by participants).
Mean Average Precision (MAP)	Average precision is calculated for each feature, and the mean average precision reported for a submission.
Recall	True positive, true negative, false positive, and false negative rates.

Submissions

Run Type	Run Id
O	PKU_WICT_7
O	PKU_WICT_2
O	PKU_WICT_4
O	PKU_WICT_5
L	PKU_WICT_6
L	PKU_WICT_3
L	PKU_WICT_1
L	PKU_WICT_8
L	UMKC_1
O	UMKC_1

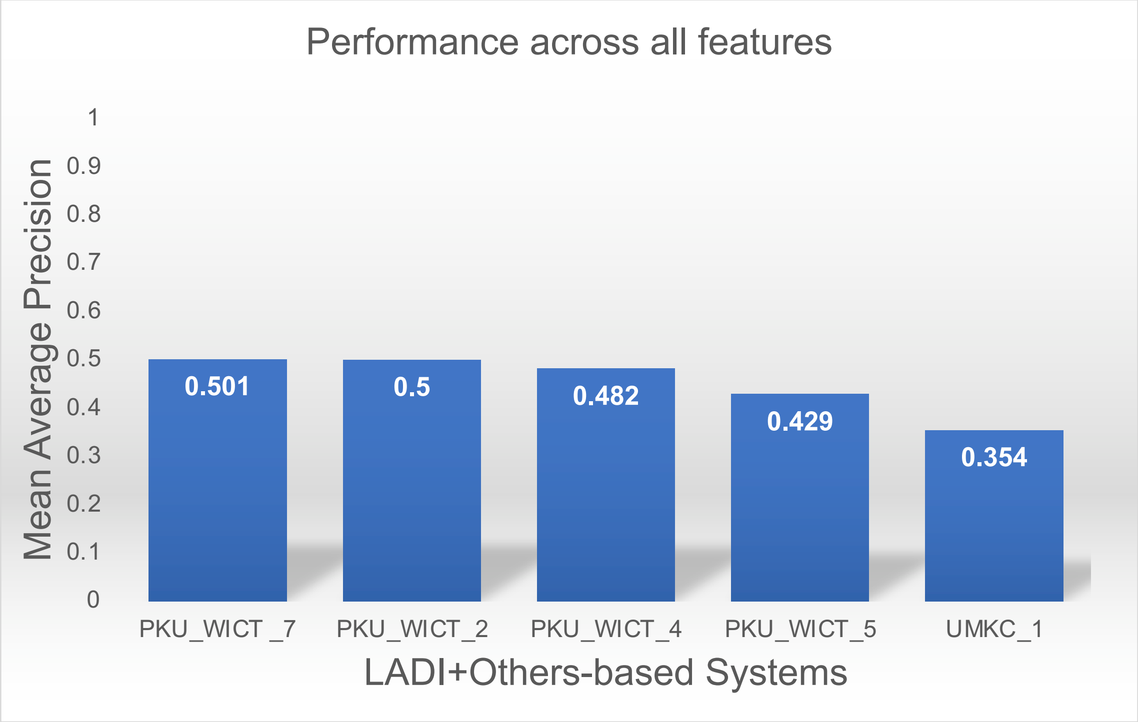
PKU_WICT : Peking University
UMKC : University of Missouri-Kansas City

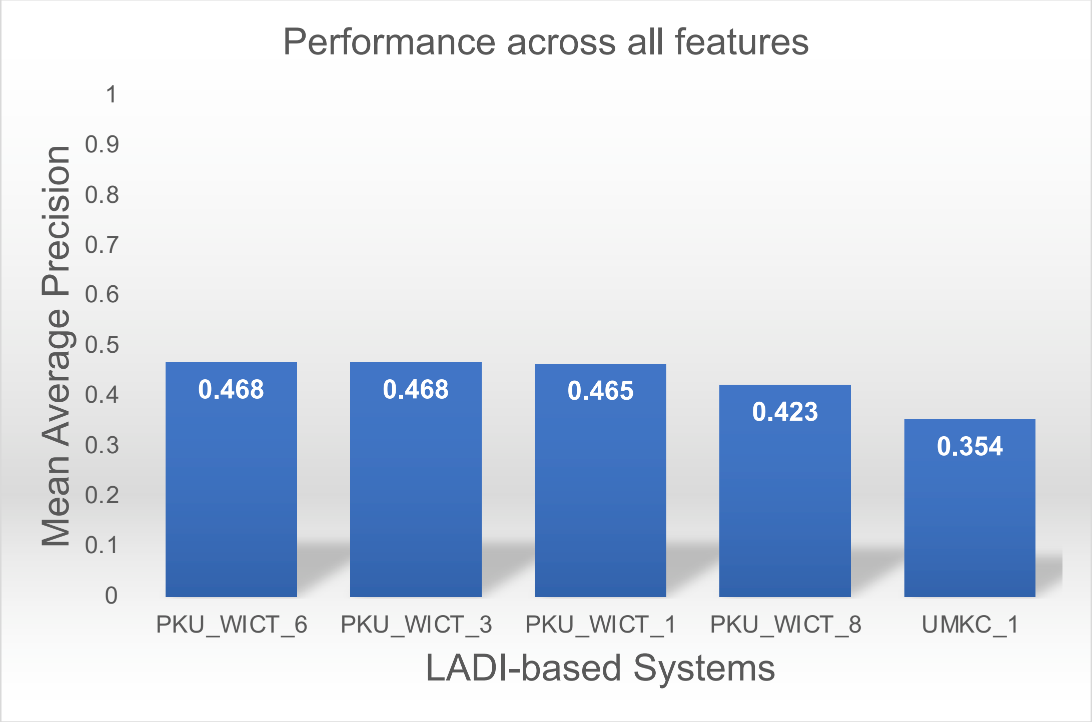
Frequency of Features



- Graph shows number of shots containing each feature.
- Some features (e.g. grass, trees, buildings, roads, etc.) occur much more frequently than others.
- 4 features were dropped due to the rare occurrence in ground truth (lava, snow/ice, landslide, road washout)

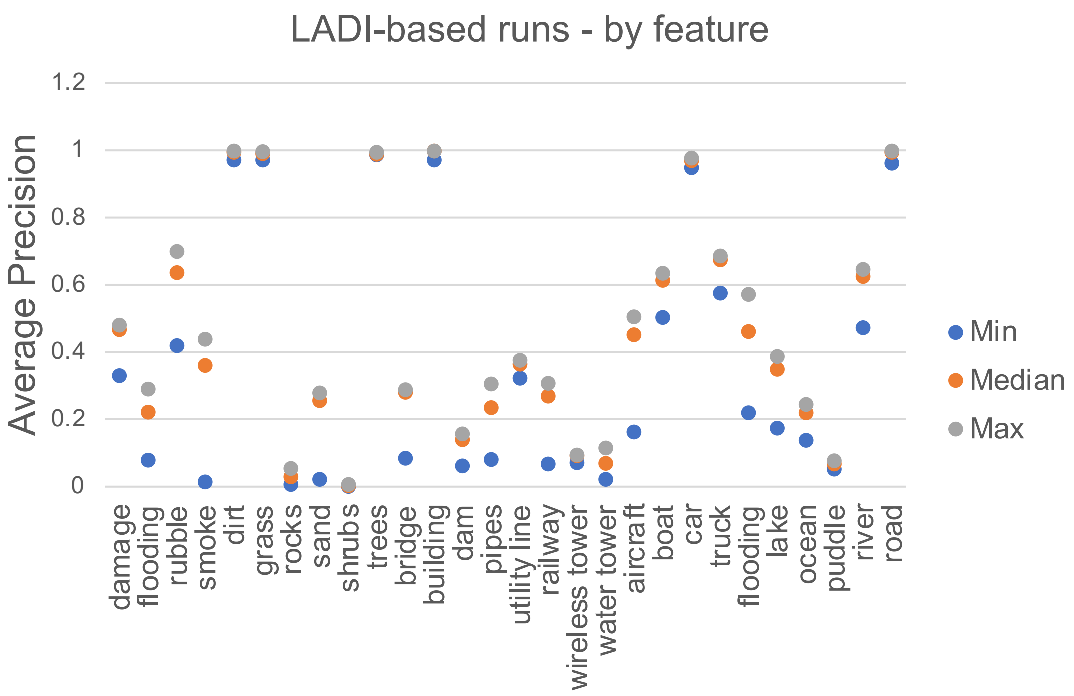
Results by Teams

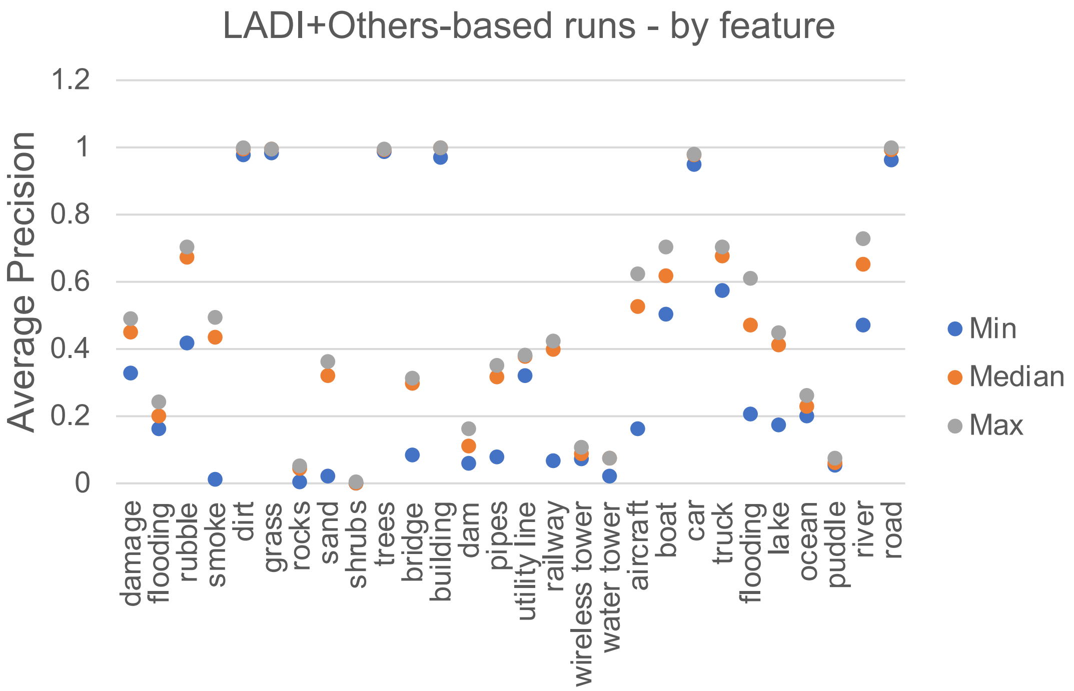




Results by Categories

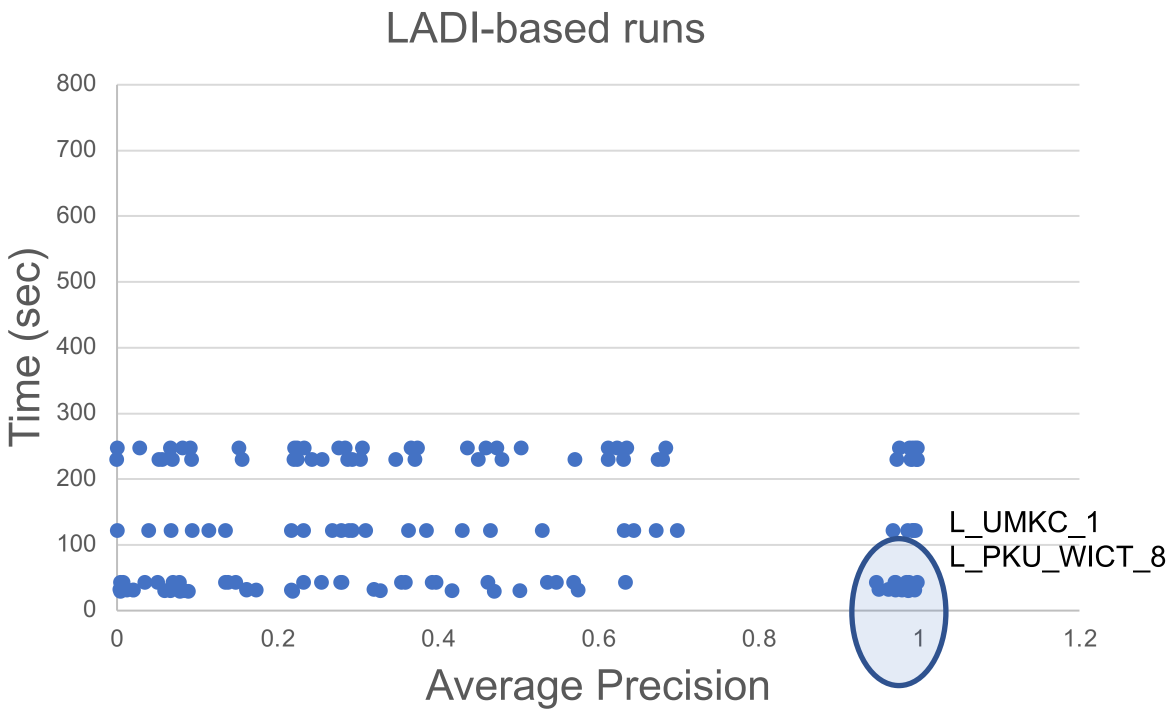
Results by Features

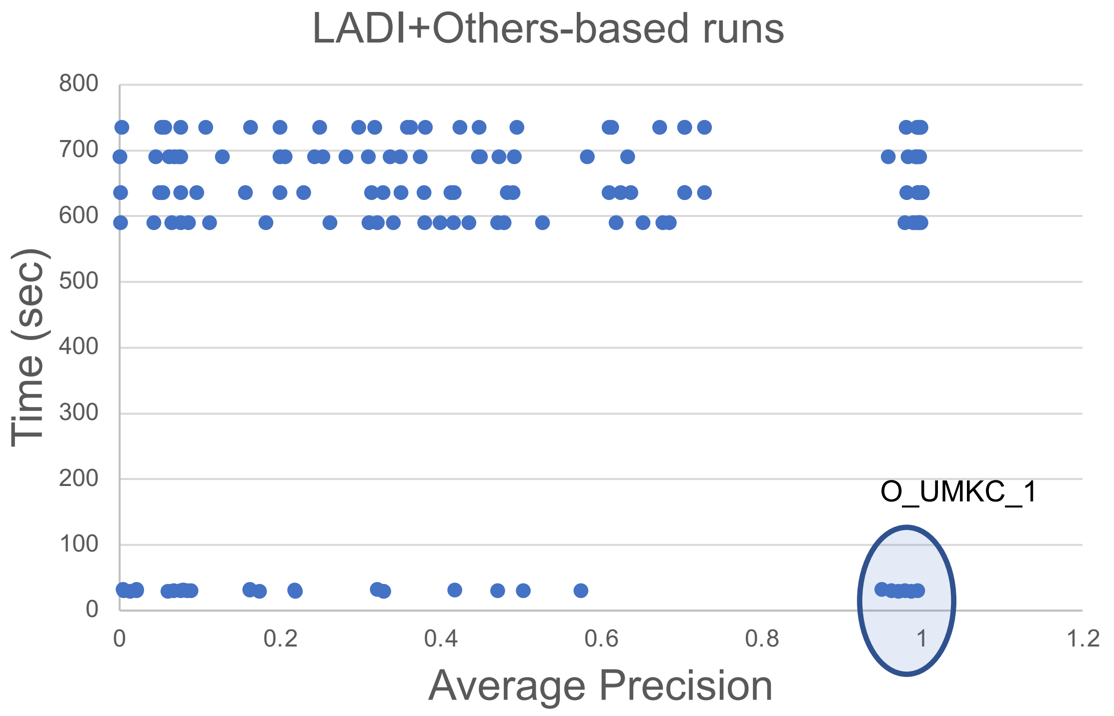




- Average precision values for each feature categorized by training type.
- 5 LADI-based runs ; 5 LADI+Others-based runs

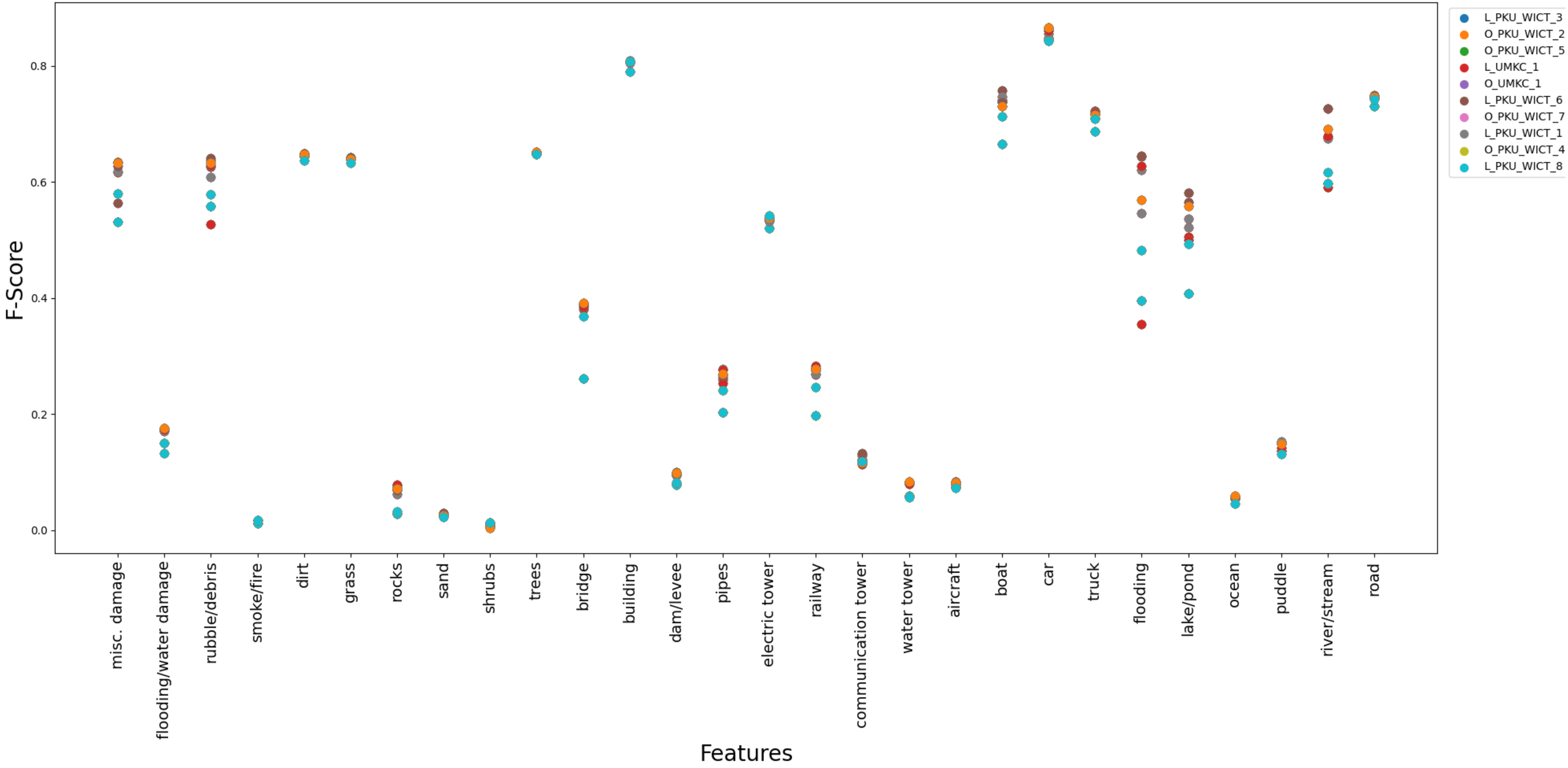
Efficiency





- LADI-based systems reported less processing time
- Majority of systems consumed more time but without more gain in performance
- Lowest processing time was ~30 sec at max performance

F-Measure per feature



Conclusion and Future Work

NIST

- A new test dataset from various event sources were employed representing more diversity.
- Performance varies by feature.
- L+O runs performed higher than L-based runs.
- Few runs/features are good and efficient.
- Challenges include:
  - Small dataset and limited resources for annotation.
  - Training and testing dataset should be from the same distribution. Hard to do with different nature of calamities.
- The task had little participation compared to the last 2 years.
- Question to teams about continuation of task.