Audio-Visual Disaster Rescue

David Duggan, Cole Lewis, Daniella Mallari

Problem Statement

As the Earth's climate changes, our species is experiencing a rapid increase of volatile and dangerous weather that often tragically conclude in disaster events in which a great loss or injury of life and property is experienced. This is happening more frequently in more areas now than ever before.

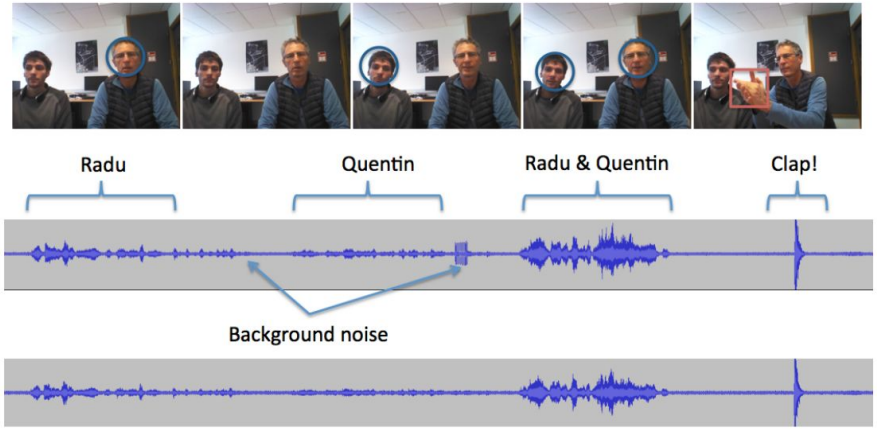
The problem at hand is how to effectively navigate disaster-stricken environments safely, effectively, and quickly to determine where rescue resources should be allocated most efficiently towards locating and saving victims.

Technical Challenges

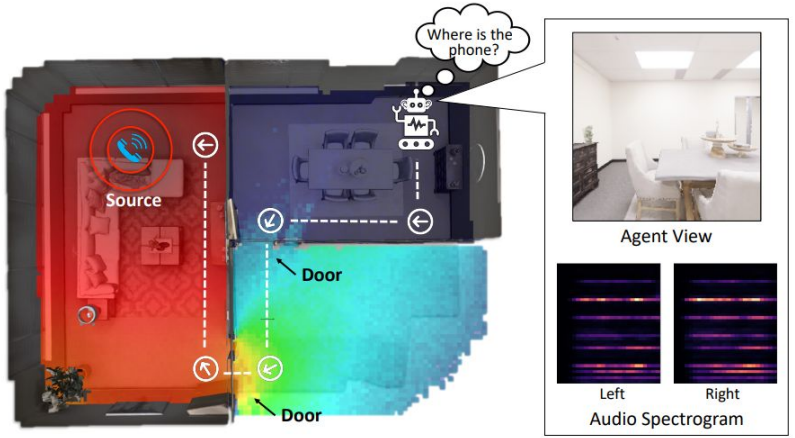
Our project presented multiple technical challenges to overcome:

- Integrating multiple types of stimuli to determine if there are victims in the immediate vicinity as well as making inferences to the victims present within range of the microphones.
- Designing and implementing models that accurately detect humans, whether by voice or camera view.
- Making and implementing the models in as close to real-time as possible, which became an impossibility when figuring in time to capture audio input.

Related Works



Radu Horaud, PERCEPTION team at INRIA Grenoble Rhône-Alpes working with Audio-Visual Fusion for Human-Robot Interaction



Audio source in an unmapped 3D environment, C. Chen & U. Jain et al., SoundSpaces: Audio-Visual Navigation in 3D Environments

Camera and Microphone Array: PlayStation Eye

- Camera used was the PlayStation Eye, a PlayStation camera accessory with built-in microphone array.
- Created in 2007 by Sony as an accessory to the PlayStation 3.
- The built in microphone array allows us to do sound-source localization.
    - 4-microphone broadside linear array

Person Recognition Model

- The visual model used was the YOLOv4-tiny object-detection network pre-trained on detecting humans.
- Detection is done in real-time and so a model with fast inference was needed.
  - Yolo is known as one of the fastest object detection networks because it only looks at the frame once.
  - We chose the tiny version to give it an even faster inference speed, which was necessary on CPU.
- Model outputs bounding box/class predictions of humans detected in a frame.

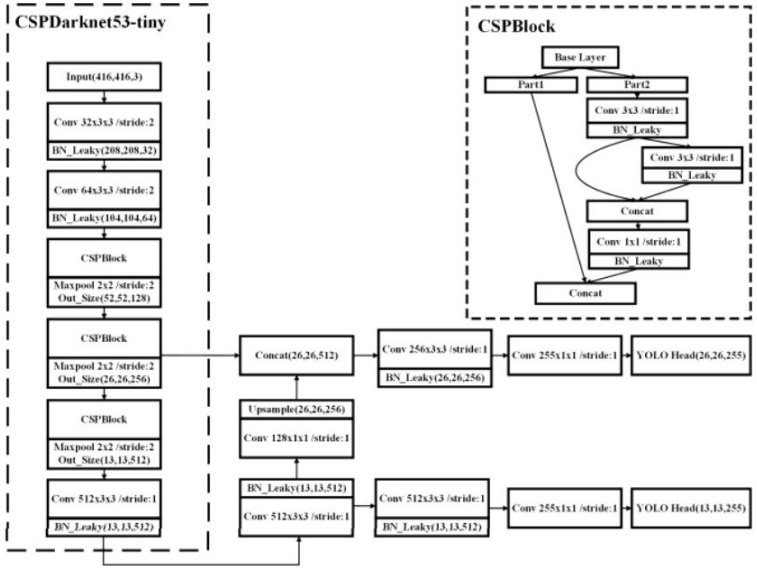
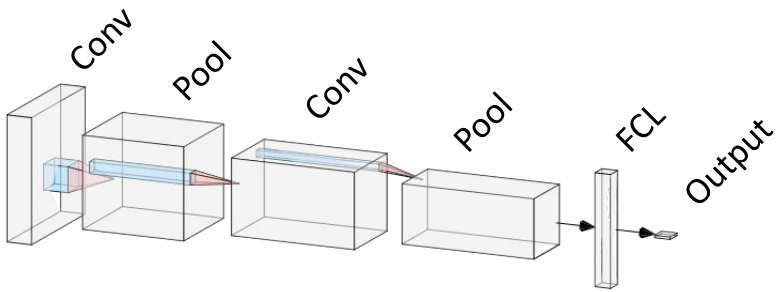


FIGURE 1. YOLOv4-tiny network structure.

Human-audio Detection Model

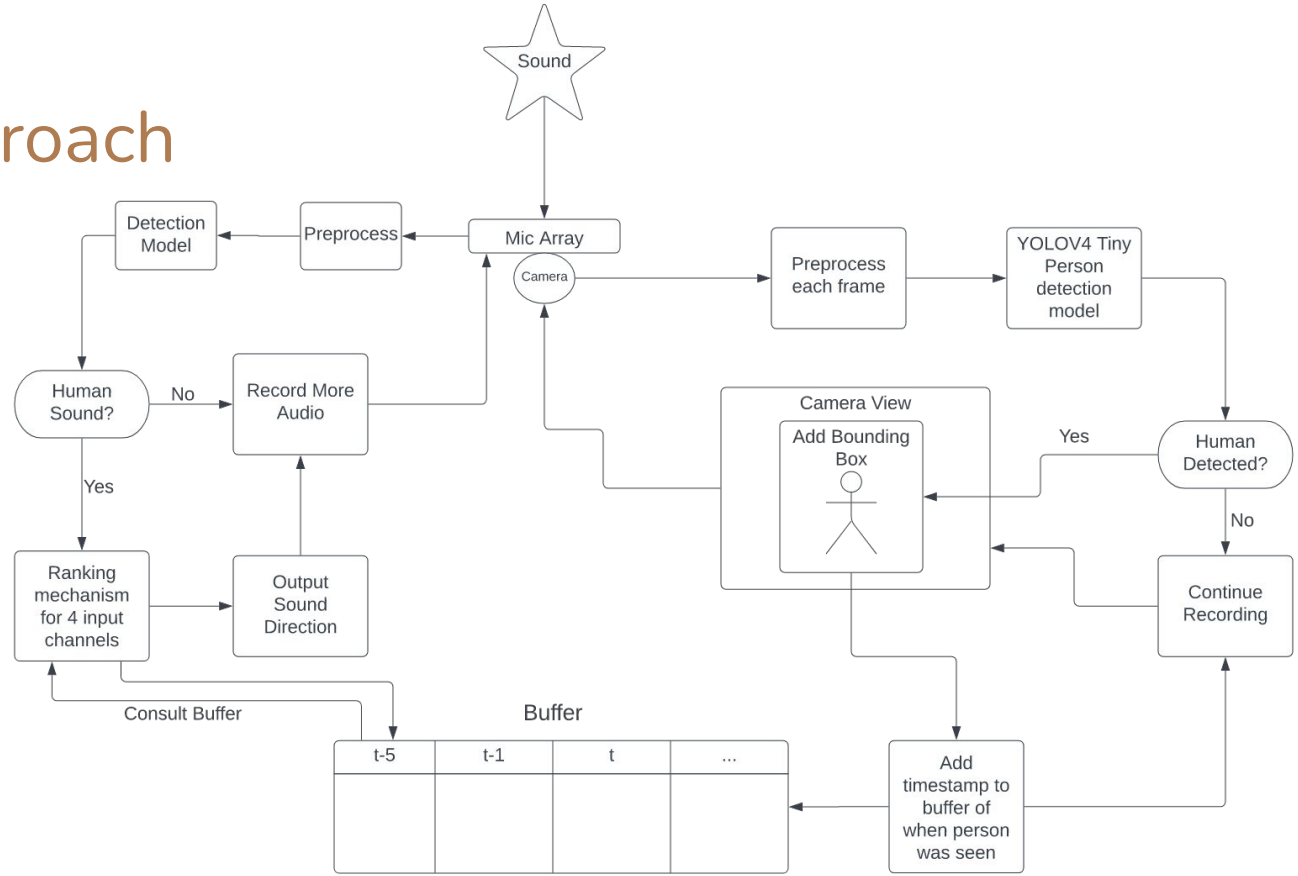
- The human-audio detection model is a Convolutional Neural Network.
- Audio was converted into spectrograms and then passed to network to classify whether it was a human sound or not.
    - Uses 5-second audio recording.
- The idea behind this model was to make the sound-source localization tolerant to other background noises.



Sound-source Localization

- Want to determine which direction human sound-source came from.
  - Directions desired are front, back, left & right.
- Created a ranking system to determine where sound came from based on different methods.
  - Two methods tested:
    - Which channel had the loudest instantaneous amplitude over the 5 second recording period.
    - Which channel had the loudest average amplitude over the 5 second recording period.
- Easy to determine left and right using a linear microphone array, but difficult to determine front and back because all microphones are in a line.
- If the highest ranked channel was one of the middle two, the tie was broken using the visual model.
  - Ranking system checks if vision model detected a human during the time when sound was recorded.
  - If human was seen, the sound came from in front of the PSEye.
  - If no human was seen, the sound came from behind the PSEye.
- Looked into beamforming, but it requires a static environment.

Approach


Sound
Mic Array
Camera
Preprocess
Detection Model
Preprocess each frame
YOLOV4 Tiny Person detection model
Human Sound?
No
Record More Audio
Yes
Ranking mechanism for 4 input channels
Output Sound Direction
Camera View
Add Bounding Box
Human Detected?
Yes
No
Continue Recording
Consult Buffer
Buffer
t-5
t-1
t
...
Add timestamp to buffer of when person was seen

Results - Sound Model

- Tests of hyperparameter tuning of the sound model are shown below. Best model was sound_model3, as it had the lowest number of False Negatives (Type II error).
- In practice however, the sound model did not perform nearly as well.
- Sound model did not perform well in setting mentioned on previous slide, only recognizing human voice 37.5% of the time over 40 trials.
    - We believe that poor predictive accuracy could be due to the model's sensitivity to noise and being tested on a different microphone then it was trained on.
    - Average inference time is 0.20 ± 0.02 seconds

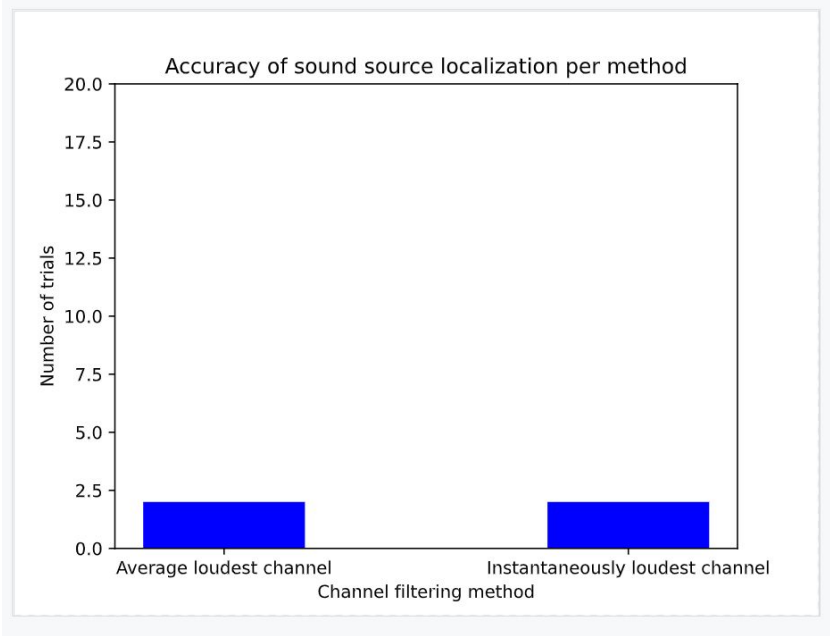
Model name	Epochs	Batch size	Learning rate	TP	FP	FN	TN	Accuracy
sound_model2	150	32	0.001	303	14	109	210	0.806
sound_model3	300	32	0.001	178	139	23	296	0.74
sound_model4	200	16	0.001	296	19	118	204	0.784
sound_model5	300	16	0.0001	309	6	198	124	0.67

Results - Visual Model

- Our visual, human-recognition model averages an IoU score of 0.44 over an Open Images Dataset of people and its frame rate is dependent upon the hardware it is being deployed on.
- Using our laptops, we average about 16-17 FPS with the visual model.
    - We believe inference time speedup is possible using a GPU to parallelize inference and perform more efficiently on FLOPS.
- We used a Tensorflow Lite implementation of the model to speed up its inference.

Results - Sound Localization

- Our sound localization system performed similarly with both methods, averaging around 10% accuracy.
- There are a few potential reasons for this:
  - Microphone array is 76mm across, so microphones are at most 19mm apart. This can make it hard for them to distinguish sounds.
  - Weak sound model inference in real-world setting means many human sounds never make it to localization stage.
  - PlayStation Eye microphone is less than today's fidelity standards.
  - Many moving parts in system, so some efficiency is lost at each step.



Results - Discussion

- Our results were skewed by multiple factors which do not include the quality of programming including:
  - aging/incompetent hardware
  - lack of access to the training dataset used with the visual human-recognition model, and more
  - Testing distribution is different than training distribution (different hardware, noisy input, etc.)
- In the future we could make potential improvements by:
  - Looking at more advanced methods of sound-source localization that work in changing environments.
  - Further quantize vision model.
  - Perform data augmentation to create noisy samples for sound model.
  - Use more advanced hardware to improve sound-source localization abilities.

Broader Impact

We expect the impact of our project to be limited to instances in which any sort of general person detection in a given area is sought after. We present a basic form of person detection that utilizes multiple sensory faculties in order to make inferences, as well as a means of integrating them together.

Ultimately, our project is limited by its lack of specificity. It can’t specifically locate people by sound and only makes vague inferences dependent on raw channel input translating to cardinal directions relative to the microphone’s position.

The greatest potential for future improvement lies in sound localization, which this iteration was limited by our available hardware.