2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Audio-Visual Traffic Light State Detection for Urban Robots

Sagar Gupta, Akansel Cosgun

Abstract— **We present a multimodal traffic light state detection using vision and sound, from the viewpoint of a quadruped robot navigating in urban settings. This is a challenging problem because of the visual occlusions and noise from robot locomotion. Our method combines features from raw audio with the ratios of red and green pixels within bounding boxes, identified by established vision-based detectors. The fusion method aggregates features across multiple frames in a given timeframe, increasing robustness and adaptability. Results show that our approach effectively addresses the challenge of visual occlusion and surpasses the performance of single-modality solutions when the robot is in motion. This study serves as a proof of concept, highlighting the significant, yet often overlooked, potential of multi-modal perception in robotics.**

I. Introduction

The development of urban mobile robots is bringing big changes, especially in areas like package delivery, surveillance, and as robotic helpers. For these robots to move around city streets safely and effectively, they need to be good at navigating autonomously, especially when moving on sidewalks or crossing streets. A key part of this navigation is being able to spot pedestrian traffic lights (PTLs) and knowing when the lights are red or green, to ensure safe and effective robot navigation in urban environments.

Recent advancements in deep learning have notably enhanced computer vision, leading to the creation of sophisticated vision-based object detectors [15], [8], [16]. However, these systems often struggle with occlusions, a significant challenge in urban environments where pedestrians, bicycles, or vehicles can block the robot's camera view. This issue is more pronounced for ground robots with lower camera positions, intensifying the problem of occlusion. To overcome this, exploring other sensor modalities becomes crucial. Sound, for instance, can complement visual data effectively. The integration of audio and visual information has shown promising results in various fields, including deepfake detection [11], robot learning [5], emotion recognition [9] and multimedia analysis [2]. Nonetheless, the application of audio-visual fusion for traffic light state detection remains largely unexplored.

We address this research gap by focusing on PTL detection from the perspective of a quadruped robot navigating real urban environments. Our approach is particularly suited for PTLs that emit sound patterns corresponding to their light state, designed for aiding visually impaired pedestrians. Our vision-based detection (Section III) utilizes a third-party object detector to get the traffic light housing bounding box, which then we use a simple pixel counting approach in the HSV color scale. Our audio-based detection (Section IV) takes raw audio as input, extracts Mel-Frequency Cepstral Coefficients (MFCC) as features [1] and uses a random forest classifier [3]. We propose a feature-level audio-visual fusion model (Section V) that incorporates a sequence of frames in a given timeframe. In Section VI, we report the accuracy of our audio-visual PTL state classifier under the following conditions:

- When the robot is not moving, and under no occlusions
- When the robot is not moving, but its view is blocked
- When the robot is moving, and under no occlusions

Section VII details the implementation of our approach on a Unitree Go1 quadruped robot, showcasing it autonomously crossing the road when the light turns green[1]. The contribution of this paper is three-fold:

- An audio-visual fusion model that uses audio features and color histograms in given bounding boxes in a given sequence of frames.
- The first time audio-visual fusion is applied to traffic light state detection for urban robots.
- Analysis on a dataset taken by a quadruped robot, under varying visual occlusion and robot motion.

Fig. 1: In urban settings, robots often struggle with visual occlusion, hindering their ability to detect traffic lights. Our solution combines auditory cues with vision, utilizing traffic lights' sound patterns to indicate their state. This method ensures robots can navigate effectively, even when visual signals are obscured.

Authors are with Deakin University, Australia

[1]https://www.youtube.com/watch?v=Wm293-VgKzI

II. System Overview

The system, designed for urban robots, classifies PTLs as red or green by integrating both visual and auditory data. For model training, we used two datasets. The visual dataset comprised images from the ImVisible dataset [18], featuring PTLs from Shanghai intersections. This dataset was enriched with manual annotations of PTL bounding boxes in YOLO format, into a single class. Our auditory dataset included two hours of handheld audio from 3 crosswalks in Melbourne, Australia; each for red and green class.

The vision-based detection employs YOLO architecture [15] for traffic light detection, followed by pixel counting for feature extraction. The audio inference part extracts MFCC [1] from 250 millisecond segments to be used as features. Finally, the system employs feature level fusion for final classification.

As shown in Figure 2, the system processes vision and audio data synchronously.

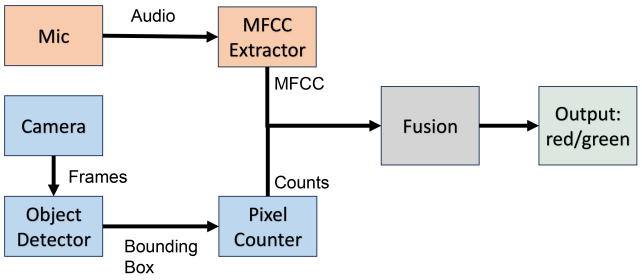


Fig. 2: Proposed Audio-Visual Feature Fusion Architecture

Our approaches were tested a video dataset captured by a Unitree Go1 quadruped robot equipped with a smartphone. This dataset encompassed 50 minutes of unobstructed footage on a stationary robot, 15 minutes of full or partial visual occlusion, and 20 minutes with the robot in motion producing motor and footstep sounds, as instanced in figure 3.

The system's performance was evaluated by comparing the outputs of the best-performing vision, audio, and fusion detection algorithms. Each data point used for this comparison corresponds to the length of the audio frame. Our video dataset operates at 30 FPS with a resolution of 1920x1080, aligning with the majority of consumer-grade cameras.

III. Vision-based Detection

PTLs in images of crosswalks typically represent only a minor fraction of the total pixels. Vision-based traffic light state detection has been extensively implemented in literature. Authors in [12] have proposed single shot detection that uses the same model for PTL detection and state classification. Authors in [17] utilize a combination of object detection and color classification.

The ImVisible dataset [18] provides images of pedestrian crosswalks in Shanghai, China, which we extracted at 1920x1080 resolution.. The images were captured under various conditions of weather, daylight, and traffic. There is

(a) Containing 'green' datapoint (b) Containing 'red' datapoint

Fig. 3: Examples from our video dataset collected onboard Unitree Go1 robot using smartphone camera

extensive similarity between Shanghai and Australian PTLs in terms of shape, structure, and color. The dataset comprised 1477 images with red PTLs, 1303 with green PTLs, and 412 without any PTLs. To facilitate PTL detection training, we found it necessary to manually annotate the PTL housings.

We compared fine-tuning YOLOv8n, YOLO's latest model, for detecting red and green PTLs against two approaches: training YOLO on multiple classes (red, green) and training YOLO on a single class (PTL). YOLO fine-tuned on red/green PTL classes allows single-shot classification, whereas YOLO fine-tuned on a single class (PTL) requires further classification to determine the light state. Since we are using object detection for our vision methods, there is a possibility that no object is detected within an image. In that case, our method is unable to make a final classification, thus, the output of our classification will be 'Unavailable'.

The Red-Green-Blue (RGB) color model is a prevalent method for image representation, which allows for segmentation of light value through distinct red, green, and blue components. However, the RGB model is sensitive to variations in lighting, leading to potential difficulties in detecting varying shades of green or red. To address this limitation, an alternative color space, known as Hue-Saturation-Value (HSV), is employed to classify traffic lights by hue value in [17].

HSV is particularly effective in distinguishing color values under diverse lighting conditions due to its separation of the color component (hue) from the color intensity (saturation) and brightness (value). This method mirrors the approach used in [17] for classifying vehicular traffic lights (VTLs) in Hangzhou, China, where the hue ranges for red and green VTLs were identified as 160-179 and 40-70, respectively.

Determining optimum hue ranges involves utilizing our single-class YOLO model to detect bounding boxes in our robot's dataset, which comprises 50 minutes of footage without occlusions, captured using a stationary robot setup. These bounding boxes were then converted to the HSV color

space, and the pixel count for each hue value was averaged and graphically represented in figure 4.

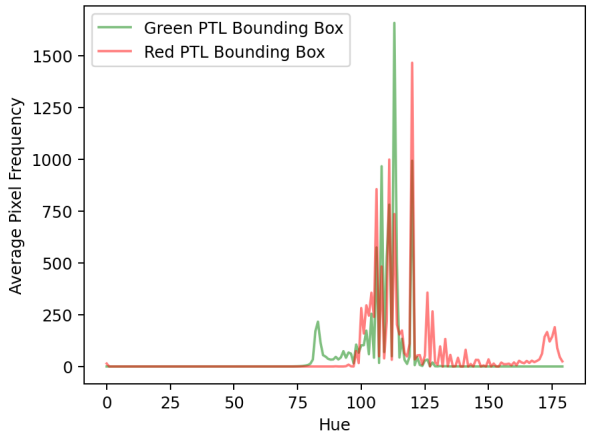
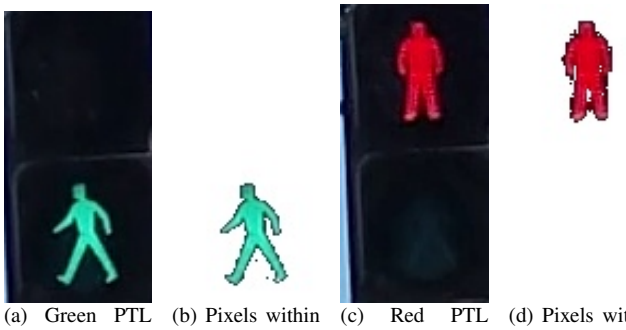


Fig. 4: Hue histogram of red and green PTL bounding boxes, averaged over all frames in our video dataset containing no occlusion captured on-board robot.

Our color analysis of the video dataset yielded distinct hue ranges for green and red PTL bounding boxes, identified as 75-100 for green and 170-180 for red. These findings suggest that these hue ranges are representative of the unique color characteristics inherent to each PTL class. The hue-based filtering results for bounding boxes for an example data point (figures 5a and 5c) are shown in figures 5b and 5d.



(a) Green PTL Bounding Box (b) Pixels within hue range (c) Red PTL Bounding Box (d) Pixels within hue range

Fig. 5: Extracting pixels within hue range of PTL bounding boxes from example frames

For classifying using hue thresholds, Let's define:

- R as the number of red pixels within the bounding box.
- G as the number of green pixels within the bounding box.

Then, the percentage of red pixels (P_{red}) and green pixels (P_{green}) can be calculated as follows:

$$P_{\text{red}} = \frac{R}{R+G} \times 100$$

$$P_{\text{green}} = \frac{G}{R+G} \times 100$$

P_{red} and P_{green} indicate the proportion of pixels classified as green or red, adjusted for varying resolutions. The final classification C is determined by comparing P_{red} and P_{green}:

$$C = \begin{cases} \text{'Unavailable'}, & \text{if } P_{\text{red}} = P_{\text{green}} \text{ or no detections} \\ \text{'Red'}, & \text{if } P_{\text{red}} > P_{\text{green}} \\ \text{'Green'}, & \text{if } P_{\text{red}} < P_{\text{green}} \end{cases}$$

We conducted a comparative analysis of two approaches: object detection for PTL followed by hue-based classifier, and an object detection model fine-tuned for red and green PTL detection. We evaluated the performance of our vision inference methods using all images from ImVisible dataset [18] as well as all frames from our video dataset captured onboard the robot. We only use data points without visual occlusion and without robot movement to analyze accuracy of our strategies in ideal conditions. This exercise contained 1303 images containing green PTLs, 1477 images containing red PTLs from the ImVisible dataset, which was used for training. From our video dataset, we extracted 56,070 frames containing green PTL and 30,780 frames containing red PTL for a comprehensive evaluation of unseen data. Since all of the data points contain either a red or green PTL, an 'unavailable' classification is considered inaccurate in case no detections are made. We firstly split images from our two datasets into red or green sets, and analyze accuracy of our models against all elements in the set, tabulated in table I.

TABLE I: Classification accuracy of vision-based approaches on ImVisible dataset [18] and all frames without visual occlusion from video dataset captured on-board robot

Classification Method	ImVisible Dataset [18]	Robot Video Dataset
Single class YOLO + HSV thresholding	**98.4%**	**96.7%**
Multi class YOLO	97.1%	90.5%

From our analysis, it is revealed that using HSV thresholds proved to be most effective on both datasets, achieving 96.7% accuracy on our video dataset. Comparatively, YOLO fine-tuned on red/green PTL classes performed worse on the dataset it was trained on, but was able to achieve 90.5% accuracy on our video dataset. For integration in our fusion model, we elect our HSV thresholding approach coupled with YOLO trained on a single class due it its outperformance of single shot detection through YOLO.

IV. Audio-based Detection

The integration of vision and audio data in our system necessitates careful consideration of synchronization, audio frame length, and the selection of features that accurately reflect the state of PTL. Our model's performance was evaluated using audio clips of varying lengths: 250ms, 500ms, 750ms, and 1000ms. We found that the audio signals emitted from PTLs typically repeat every 100-200ms, about six times per second. Notably, a delay occurs when the light changes to green, marked by an initial sound effect before the standard pattern begins, resulting in a variable audio delay of 300-600ms during the transition from red to green.

A. Number of Mel Frequency Cepstral Coefficients

MFCC are a cornerstone in sound event and speech recognition, representing the short-term power spectrum of

sound [1]. These coefficients are derived from a linear cosine transform of a log power spectrum on a nonlinear mel scale. The choice of the number of MFCC and audio frame length significantly influences classification accuracy [4]. While traditionally, 13 cepstral coefficients are used, higher order coefficients are often excluded as they represent rapid changes in estimated energies and contain less information [7]. Research by [6] found that 16 MFCC classified using a k-Nearest Neighbor (k-NN) classifier was optimal for an intruder detection system. To get optimum number of MFCC, we mirror the approach in [6], which compared widely used classifiers against a range of MFCC. For our study, we opt for a cross-comparison of two classifiers, k-Next Neighbors (k-NN) and Random Forest (RF). k-NN Implements various distance measures and efficient neighbor search in large datasets [19]. RF uses random vectors to grow an ensemble of trees for voting on the most popular class [3].

We also compared these classifiers against Bayesian Networks [10] but the results were significantly worse, hence we elected to present results from our best performing classifiers. Given the unique strengths of each classifier and considering variables like environmental noise and audio pattern variability, we propose a cross-comparison across N MFCC, with N ranging from 10, 12, 14, 16, 18, 20, 24 and 28. This approach aligns with [6].

Our self-collected audio dataset comprises 2 hours of green PTL data and 2 hours of red PTL data recorded using a handheld smartphone. We extracted 30% of our audio data to be used for testing and used the remainder for training. The audio signals were split according to frame length, and features were extracted from all N MFCC. These features were then trained on the two classifiers, and their accuracies were charted in figure 6.

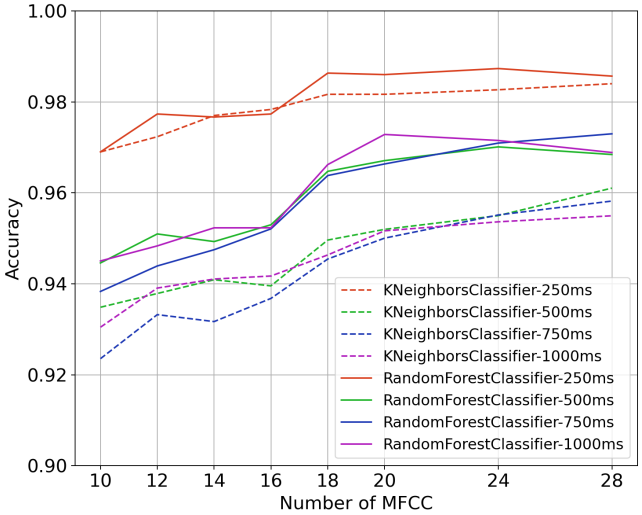


Fig. 6: Cross-Comparison of Random Forest [3] and k-Nearest Neighbor [19] classifiers trained on varying audio lengths and number of MFCC for red/green classification.

The Random Forest classifier trained on 250ms audio frames with 24 MFCC achieved the highest accuracy at 98.7%. With 24 MFCC identified as optimal, we extracted accuracy data for our classifiers trained on this feature set across varying audio frame lengths. The results, presented in Table II, indicate the highest accuracy with 250ms frames. Notably, the Random Forest classifier outperformed k-NN across all frame lengths, leading us to select the Random Forest classifier trained on 250ms frames with 24 MFCC as our model of choice.

Classifier Type	Frame Length	Accuracy
KNeighbors Classifier	250ms	98.3%
	500ms	95.5%
	750ms	95.5%
	1000ms	95.4%
Random Forest Classifier	**250ms**	**98.7%**
	500ms	97.0%
	750ms	97.1%
	1000ms	97.2%

TABLE II: Comparison of Random Forest [3] and k-Nearest Neighbor [19] classifiers trained on varying audio lengths and 24 MFCC for red/green classification.

V. Audio-Visual Fusion

The fusion of data from disparate sources can significantly enhance accuracy, as evidenced in studies like [20]. This particular research successfully fused thermal (RGB-T) and RGB images to improve urban scene understanding. Generally, multimodal fusion in machine perception can be categorized into two primary types: decision-level and feature-level fusion [2], [14]. One notable application of audio-visual fusion is in the field of deepfake detection. The study in [11] combined features from a face detection model with MFCC, relying on just one frame per second for visual inference. Furthermore, [9] demonstrated the use of audio and video features alongside motion capture for emotion recognition, employing decision-level fusion. [13] introduced a transformer-based architecture that utilizes 'attention bottlenecks' for effective multimodal fusion at various layers. In our research, we aimed to explore and compare the effectiveness of these two fusion strategies—feature-level and decision-level fusion—for our classifier.

We performed a comparison of fusing audio and vision features for classification versus combining the decisions of audio and vision-based classifiers to output classification at the decision level. In our fusion approach, we define a single data point as encompassing 250ms of both video frames and audio. Given that the average inference time for our YOLO model is 62ms and a video frame appears every 33.3ms when the FPS is 30, we aim to synchronize our fusion methods by pairing 250ms of audio frames with alternating vision frames to maintain real-time inference.We average the features from every other frames within the audio clip's duration to represent the video segment's features. The overall process of this system for each data point is illustrated in figure 7.

A. Feature Level Fusion

The feature level fusion process involves concatenating the audio features with visual attributesg. For training, we concatenate MFCC features from our handheld audio dataset

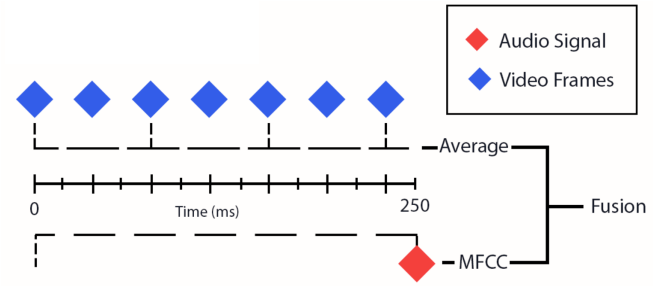


Fig. 7: Audio-Visual feature extraction and synchronization before fusion

with pixel percentage features from the ImVisible dataset [18], incorporating 1303 images of green pedestrian traffic lights (PTLs), 1477 images of red PTLs, and 412 images without PTLs. This inclusion of non-PTL images is aimed at training the model to depend on audio inputs in the absence of visual detections. A Random Forest classifier [3] is subsequently trained on these combined features.

For extracting vision features, we use [P_{red}, P_{green}] that are color percentage values calculated from pixel counts for red and green hue values. These features ensure that the pixel ratio remains consistent for both training and evaluation, independent of the variations in input image resolution. [P_{red}, P_{green}] are [0,0] if no traffic light is detected. For classification, we average these features across all frames selected for analysis. The visual attributes are then combined with the audio features, which consist of 24 MFCC extracted from 250 milliseconds of audio. Random Forest [3] is then used to perform binary classification of these features into red/green.

B. Decision Level Fusion

For decision level fusion, we sum the confidence scores from our vision and audio classification methods for each data class, with the predicted class being the one with the highest cumulative confidence. In terms of vision, the confidence metric is derived from the average detection confidence of the bounding boxes over the 4 frames. For audio, it is based on the class confidence for the respective data class. This method relies on combining the individual strengths of audio and vision classifiers by considering their respective confidence levels in predicting the PTL state, thereby leveraging the advantages of both modalities to enhance overall classification accuracy. When our object detector is unable to make a detection, only audio confidences are used as we don't have confidences for bounding boxes.

VI. Results

The classification accuracy of both our feature level fusion and decision level fusion methods against red and green data points has been thoroughly analyzed against our video dataset captured onboard our robot and documented in Tables III, IV, and V. These tables detail the overall accuracy of each fusion strategy under three distinct conditions: when the robot is stationary when it is in motion, and when the camera view is occluded. Our fusion model, handling 250ms of vision and audio data, averaged an inference time of 242ms for processing 4 frames at a resolution of 1920x1080, using the laptop Nvidia RTX 4060 GPU, on battery power.

A. Under no visual occlusion on stationary robot

Analysis for datapoints containing no visual occlusion and no robot noise is presented in table III. We observed a notable improvement in accuracy compared to the single modality approaches. Specifically, the feature level fusion method achieved an accuracy of 99.1%, while the decision level fusion method reached 98.3%. It is interesting to note that while the decision-level fusion's accuracy was slightly below the vision-only approach (98.5% accurate), the integration of features for training and inference in feature-level fusion exhibited superior performance.

TABLE III: Classification Accuracy of Audio, Vision and Fusion Approaches. Data: Captured on-board robot. Type: No visual occlusion. Green Light Data Points: 7476, Red Light Data Points: 4104, Overall Data Points: 11580

Classification Method	Green Light Accuracy	Red Light Accuracy	Overall Accuracy
Vision-only	99.3%	97.1%	98.5%
Audio-only	98.1%	95.0%	97.0%
Vision + Audio, Feature-level	**99.7%**	**98.0%**	**99.1%**
Vision + Audio, Decision-level	99.2%	96.6%	98.3%

B. Under Visual Occlusion

Analysis for datapoints containing visual occlusion is presented in table IV. In scenarios of complete or partial visual occlusion, vision only method performs very poorly as it is unable to make detections, and made accurate predictions for only 4.9% of the datapoints. The feature level fusion method demonstrated an accuracy of 95.6%, compared to 97.4% for the decision level fusion. This finding indicates that decision-level fusion is more advantageous in situations where visual data is not available, as it relies more on audio confidence values for classification. Our audio classification model achieved the highest accuracy, 97.8%, slightly higher than our decision-level fusion.

TABLE IV: Classification Accuracy of Audio, Vision and Fusion Approaches. Data: Captured on-board robot. Type: Visual occlusion. Green Light Data Points: 3256, Red Light Data Points: 1812, Overall Data Points: 5068

Classification Method	Green Light Accuracy	Red Light Accuracy	Overall Accuracy
Vision-only	5.4%	4.0%	4.9%
Audio-only	**99.7%**	**94.3%**	**97.8%**
Vision + Audio, Feature-level	98.1%	91.2%	95.6%
Vision + Audio, Decision-level	**99.7%**	93.4%	97.4%

C. Under Robot Movement

Analysis for datapoints containing the robot in motion is presented in Table V. Feature-level fusion showed enhanced performance in conditions where the robot was moving, affecting both visual and audio inputs. Trained on a combination of vision and audio data, feature level fusion method

is better equipped to manage the most challenging category. As a result, for data points where the robot is moving, feature level fusion achieves a higher accuracy of 98.8%, whereas decision level fusion shows a slightly lower accuracy, correctly classifying 96.0% of the data points. These results highlight the effectiveness of fusion strategies in diverse and challenging operational environments, underlining their importance in enhancing the robustness and accuracy of multimodal perception systems.

TABLE V: Classification Accuracy of Audio, Vision and Fusion Approaches. Data: Captured on-board robot. Type: Robot moving. Green Light Data Points: 3136, Red Light Data Points: 1716, Overall Data Points: 4852

Classification Method	Green Light Accuracy	Red Light Accuracy	Overall Accuracy
Vision-only	96.9%	95.7%	96.5%
Audio-only	89.7%	91.7%	90.4%
Vision + Audio, Feature-level	**99.6%**	97.3%	**98.8%**
Vision + Audio, Decision-level	94.6%	**98.5%**	96.0%

VII. Robot Implementation

The proposed PTL detection system is implemented on a Unitree Go1 robot with a front-mounted smartphone for video capture, supported by a laptop with an RTX 4060 GPU. The robot, smartphone, and laptop are connected via Wi-Fi, accessing the Robot Operating System (ROS) network.

A FastAPI web server on the laptop, combined with a Python video processing module, uses our fusion model. The web server requests rear-facing video footage at 1920x1080 resolution and 30 FPS every 250 milliseconds. The video processing module analyzes the footage for pixel counting, averaging these features, and integrating them with 24 MFCC from the audio track.

The fusion model's features and the detected bounding box dimensions are published to a ROS topic /traffic_light_state. A ROS node subscribed to this topic determines the robot's actions. If a green light is detected, the bounding box size helps estimate the distance to a goal point using the robot's odometry data. The ROS node guides the robot to this point, maintaining the correct path and orientation. Upon nearing the goal point within one meter, the robot stops, indicating successful navigation across the pedestrian crossing.

VIII. Conclusion

In this study, we introduce an audio-visual fusion model for detecting pedestrian traffic light (PTL) states from the view of a quadruped robot. Our vision model utilizes established vision-based detectors for initial PTL identification and incorporates a simple pixel-counting approach for determining the state as red or green. The audio component of our model extracts Mel-Frequency Cepstral Coefficients (MFCC) features 6, integrating these with visual data at a feature level. This fusion technique can handle data from multiple frames within a set timeframe, enhancing the model's adaptability and performance.

Our fusion method notably achieved a classification accuracy exceeding 95% when the robot's view is under visual partial or full occlusions, substantially surpassing vision-only solutions, albeit slightly behind the audio-only modality. In dynamic scenarios where the robot is in motion, our fusion model performed the best over vision-only and audio-only modalities with over 98% accuracy. With an average inference time of 64ms on a standard consumer-grade GPU, our model is well-suited for real-time processing, meeting a critical requirement in urban robotics.

References

[1] Zrar Kh. Abdul and Abdulbasit K. Al-Talabani. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 2022.

[2] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 2010.

[3] Leo Breiman. Random forests. *Machine Learning*, 2001.

[4] Namrata Dave. Feature extraction methods lpc, plp and mfcc in speech recognition. *International journal for advance research in engineering and technology*, 2013.

[5] Maximilian Du, Olivia Lee, Suraj Nair, and Chelsea Finn. Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning. 2022.

[6] Lacrimioara Grama and Corneliu Rusu. Choosing an accurate number of mel frequency cepstral coefficients for audio classification purpose. In *International Symposium on Image and Signal Processing and Analysis*, 2017.

[7] Shikha Gupta, Jafreezal Jaafar, WF Wan Ahmad, and Arpit Bansal. Feature extraction using mfcc. *Signal & Image Processing: An International Journal*, 2013.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, 2017.

[9] Ning Jia, Chunjun Zheng, and Wei Sun. A multimodal emotion recognition model integrating speech, video and mocap. *Multimedia Tools and Applications*, 2022.

[10] Timo Koski and John Noble. *Bayesian networks: an introduction*. John Wiley & Sons, 2011.

[11] Sneha Muppalla, Shan Jia, and Siwei Lyu. Integrating audio-visual features for multimodal deepfake detection. *arXiv preprint*, 2023.

[12] Julian Müller and Klaus Dietmayer. Detecting traffic lights by single shot detection. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018.

[13] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, 2021.

[14] Juan DS Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal, and Alessandro L Koerich. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv*, 2019.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[16] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[17] Xiaoqiang Wang, Xianghui Cheng, Xue Wu, Huili Zhou, Xiai Chen, and Lin Wang. Design of traffic light identification scheme based on tensorflow and hsv color space. *Journal of Physics: Conference Series*, 2018.

[18] Samuel Yu, Heon Lee, and John Kim. Lytnet: A convolutional neural network for real-time pedestrian traffic lights and zebra crossing recognition for the visually impaired. In *Computer Analysis of Images and Patterns (CAIP)*, 2019.

[19] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 2007.

[20] Wujie Zhou, Shaohua Dong, Jingsheng Lei, and Lu Yu. Mtanet: Multitask-aware network with hierarchical multimodal fusion for rgb-t urban scene understanding. *IEEE Transactions on Intelligent Vehicles*, 2023.