A Multimodal Dynamic Neural Network for Call for Help Recognition in Elevators

Ran Ju
Data Science Research Center, Duke Kunshan University
Suzhou, China
ran.ju@dukekunshan.edu.cn

Huangrui Chu
Data Science Research Center, Duke Kunshan University
Suzhou, China
huangrui.chu@dukekunshan.edu.cn

Yechen Wang
Data Science Research Center, Duke Kunshan University
Suzhou, China
yechen.wang@dukekunshan.edu.cn

Qi Deng
Technology Asia and Escalator, KONE Elevators Co., Ltd
Suzhou, China
qi.deng@kone.com

Ming Cheng
Data Science Research Center, Duke Kunshan University
Suzhou, China
ming.cheng@dukekunshan.edu.cn

Ming Li*
Data Science Research Center, Duke Kunshan University
Suzhou, China
ming.li369@dukekunshan.edu.cn

ABSTRACT

As elevator accidents do great damage to people's lives and property, taking immediate responses to emergent calls for help is necessary. In most emergency cases, passengers must use the "SOS" button to contact the remote safety guard. However, this method is unreliable when passengers lose the ability of body movement. To address this problem, we define a novel task of identifying real and fake calls for help in elevator scenes. Given that the limited call for help dataset collected in elevators contains multimodal data of real and fake categories, we collected and constructed an audiovisual dataset dedicated to the proposed task. Moreover, we present a novel instance-modality-wise dynamic framework to efficiently use the information from each modality and make inferences. Experimental results show that our multimodal network improves the performance on the call for help multimodal dataset by 2.66% (accuracy) and 1.25% (F1 Score) with respect to the pure audio model. Besides, our method outperforms other methods on our dataset.

CCS CONCEPTS

• **Human-centered computing**;

KEYWORDS

Multimodal action recognition, abnormal event recognition, dynamic neural network

1 INTRODUCTION

Elevators are an important part of our life for daily transportation. As technology advances, so do the risk factors. Many hazardous situations are possible to happen while traveling in an elevator. Therefore, an efficient call for help system is needed to ensure life and property safety. Nowadays, emergency calls in elevators rely on manually pressing the "SOS" button. This hardware-dependent method has several disadvantages. It is hard for passengers to get an immediate rescue if the button is damaged or the passengers have limited mobility. However, due to the limited application and challenging nature, the specific task of intelligent calls for help recognition in elevators receives little attention in the past decades. With cutting-edge sensors to capture audio and video signals becoming common, advanced video and speech processing techniques and machine learning methods play key roles in various fields, like action recognition and classification [Wu et al. 2020; Zhang and Xiang 2020], speech recognition [Povey et al. 2011; Wolf and Nadeu 2014]. An intelligent call for help detection system improves rescue efficiency becomes possible.

For the reason that research in this area is limited, we proposed a novel framework of the intelligent system. After being waken up by the keywords, the intelligent system will only send warnings when it identifies a real call for help instead of giving out information every time it detects the keyword. In this case, there will be fewer misjudgments and the efficiency will be improved in return. Meanwhile, as speech-based keyword spotting [Chen et al. 2014] has been widely applied in many smart applications, this paper excludes the front-end keyword spotting and focuses on developing an efficient model that can identify real or fake call for help events after keyword detection.

Furthermore, we incorporate both audio and video information to address the real and fake call for help recognition. Multimodal input has a better performance compared to unimodal input. Although in most cases, prime audio signals are able to discriminate the two categories, they will fail in the situation that the passengers are imitating the real call for help in their conversations. Complementary video information providing the actions and states of the passengers can separate the vocally ambiguous cases. Therefore, the fusion of audio and video features takes advantage of complementary information. We are certainly not the first to attempt to fuse information from different modalities, there were works for multimodal event recognition before, such as [Gemmeke et al. 2017; Pareek and Thakkar 2021]. However, the above methods have a common drawback that unimodal information is sufficient for most samples, for those cases, rigid multimodal networks waste time and efficiency. Unlike these, we are intended to design a reliable dynamic neural network based on our dataset. The structure of our multimodal dynamic network is introduced in Section 4.

To support our research toward utilizing multimodal information to recognize the real or fake call for help in elevators, we first defined the meaning of real and fake categories, as specified in

*Corresponding Author.

Section 3.1. In our definition, the real category means that passengers need emergency rescues due to hazardous events of elevators, urgent accidents, or personal health problems. In contrast, the fake category means that the passengers mention the keyword in conversations in the absence of emergencies. Then, we construct an elevator multimodal call for help dataset, the Elevator Help Dataset, consisting of 3724 pairs of audio and video data captured by 266 actors. The detail of the dataset is in Section 3.

To summarize, the contributions of our paper are threefold.

- We define a new task of real and fake call for help recognition in an elevator to ensure the safety and property of humans.
- We introduce the Elevator Help Dataset fulfilling the gap of the dataset in the task we proposed.
- We propose a dynamic inference network with multimodal inputs to balance the work efficiency and accuracy of the model.

2 RELATED WORKS

In the ideal case, a complete call for help detection system needs two stages: keyword spotting [Higgins and Wohlford 1985] that can detect the presence of "help" keywords in continuous speech signals, followed by an audio-visual classifier to recognize the real or fake call for help.

Speech-based keyword spotting is currently a common and mature technique widely deployed in many applications (e.g., voice assistant [Hoy 2018]). Thus, there is no urgent need to study keyword spotting in this specified task. In this paper, we focus on addressing recognizing real or fake calls for help in elevators. Our proposed method mainly involves two concepts: event recognition and dynamic neural networks. In this section, we will discuss the related works, respectively, and summarize how they are associated with our work.

Event Recognition. Event recognition by various types of data has been a long-standing research interest. As audio and video are the most common multimedia content currently, many researchers have studied event recognition based on audio and video data [Gemmeke et al. 2017; Pareek and Thakkar 2021]. In the area of video-based event recognition, methods are plentiful and range from using hand-crafted descriptors (e.g., iDT [Wang and Schmid 2013]) with machine learning models to deep neural networks (e.g., two-stream [Simonyan and Zisserman 2014], C3D [Tran et al. 2015], I3D [Carreira and Zisserman 2017]). In the area of audio-based event recognition, Warren et al. [Warren and Verbrugge 1984] first explore the connection between perceptual properties and acoustic features. Gemmeke et al. [Gemmeke et al. 2017] propose the AudioSet dataset, which is a large-scale human-labeled dataset for audio event recognition. Also, a series of competitions funded by IEEE (e.g., DCASE 2013 [Giannoulis et al. 2013], DCASE 2017 [Mesaros et al. 2017]) all promote the development of this research area.

Multimodal Fusion. Recently, multimodal event recognition is becoming prevalent. The inputs of neural networks develop from single audio [McLoughlin et al. 2015] or video [Zhang and Xiang 2020] data to multimodal data including RGB-Depth, audio-visual signals [Cippitelli et al. 2017; Wu et al. 2020], etc. As audio and video data are different types of information, they have certain complementary capabilities in some cases. Therefore, combining information from different modalities takes advantage of integrated information. Many previous works related to multimodal event recognition have studied how to fuse information from different modalities efficiently. One is the joint representation, also known as early fusion [D'mello and Kory 2015; Nojavanasghari et al. 2016; Ramachandram and Taylor 2017], which aims at fusing multimodal features in the middle parts of neural networks. Another fusion scheme is late fusion. Contrary to early fusion, late fusion is based on the results of the decision from each modality. The advantage of late fusion is that it allows different models for different modalities, which build better models for specific modality [Baltrušaitis et al. 2018]. Additionally, hybrid fusion [Atrey et al. 2010; D'mello and Kory 2015] combines these fusion methods, and it has been proved successful in multimedia event recognition [Lan et al. 2014].

Dynamic Neural Networks. Most popular deep networks are rigid. They have a static inference paradigm [Han et al. 2021]. Once the training is completed, the parameters and structure of the model remain unchanged, which limits the models' capabilities [Graves 2016; Huang et al. 2017; Sabour et al. 2017; Yang et al. 2019]. The more difficult the task, the more a neural network with a larger size and stronger representation ability is needed. As the hard samples account for a few parts of the whole dataset, a large computational model adopted for all inputs is inefficient [Han et al. 2021]. The key idea of dynamic neural networks is to solve easy input data by smaller models and hard input data by larger models. There are three categories of dynamic networks: spatial-wise, temporal-wise, and instance-wise. Spatial-wise dynamic networks choose different locations of the image as the input. The relevant approaches can be further divided into three levels: pixel-level, region-level, and resolution-level. These methods are suitable for cases where the global background has no crucial clues. Temporal-wise dynamic networks solve the problem of redundancy in temporal dimensions. For the video as the input, there are two methods: RNN-based adaptive networks and a dynamic pre-sampling procedure for key frames [Han et al. 2021]. Instance-wise dynamic networks select different structures or parameters when processing different samples. The main goal of choosing structures is to improve computational efficiency, while dynamic parameters desire to improve the representation power with minimal computational cost [Han et al. 2021]. A natural way for dynamic architecture is early exiting, which adds routers in the middle layers and decides output results when the confidence score reaches the threshold. Evidence [Huang et al. 2017] shows that the ports in multiple layers of a network can interview each other and weaken the capabilities of the model. Multi-scale dense network (MSDNet) [Huang et al. 2017] adopts a multi-scale architecture with dense connections, solving the problem of intermediate router effectively. Considering the background and length of our data, we use an instance-wise dynamic inference scheme.

In all, for our task of call for help recognition, the definition of the categories, the application setting in elevators, and limited views are unavailable in the existing public datasets. Therefore, we construct a dataset for our specified task. Moreover, we propose a modality and depth dynamic inference framework using information from different modalities efficiently. This way, the proposed model is able to learn both the low-level and high-level cross-modal information,

has a better representation power, and makes accurate predictions with higher efficiency compared to unimodal methods.

3 TASK DEFINITION AND PROPOSED DATASET

3.1 Task Definition

Our goal is to build an automatic system that can recognize the trueness of calls for help in elevator scenes and improve the efficiency of the safety monitoring system. For developing this framework, the intuitive way is regarding the proposed task as a para-linguistic recognition task after keyword spotting that can identify real or fake calls for help in elevators. However, two potential cases will confuse speech-only models. First, some common conversations also include the keyword, while speakers are casually talking about a related topic. Second, people imitate movie scenes or accidents with keywords spoken in a vivid tone. In the two cases above, only audio features cannot tell whether people are really in need of help.

To mitigate this problem, we consider the cases in daily life to define the real and fake call for help categories. Both categories have the same requirement that the keyword "Help" is included in the sentence. Briefly speaking, a real call for help happens when passengers need rescues in certain urgent emergencies (e.g., elevator malfunctions, physical discomfort), while a fake call for help is the situation where passengers have no need for instant rescues even if they try to fool the system(e.g., talking about related topics, imitating jokes).

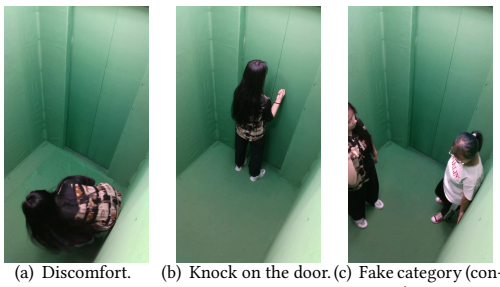
The following are examples of the real category. Imagine a lady is trapped in the elevator due to the malfunctioning elevator door. She shouts "Help! Help! The door is broken. Someone, please help me!" flapping the elevator door to attract others' attention. Her tone is frightening during the whole process. Another example is that a man sitting on the floor holds his stomach with a painful face. He tries to shout but his voice sounds feeble. He says "Help! There is something wrong with me. Please help me!"

Fake call for help examples is shown below. A woman says to another passenger that " I heard someone shouting ' Help! Help!' outside. Could you hear that?" In this example, the woman is imitating another person's call for help. Although her tone when she said "Help! Help!" is frightening which sounds like real samples, she does not need rescue. Another example is that a man says, "I watched the film 'The Help!' yesterday." The keyword is in the sentence, but it is not a sign of a call for help. The keyword is only part of a normal word.

3.2 Elevator Help Dataset

Data Collection. Our dataset is collected in a simulated elevator environment with the green screen as the background. The green screen is used for image segmentation in the future. The video and audio data are recorded by one camera in the top corner of the cabin ceiling and one microphone in the center of the ceiling, as shown in Figure 2(a). Due to the redundancy of adjacent frames in video data, the video FPS (frames per second) does not need to be too high, and the frame rate is downsampled and saved to 5 fps. Additionally, the audios are collected by a high-fidelity microphone with a sampling rate of 16 kHz.

To acquire sufficient high-quality data, we employ 266 actors of all ages and provide them with a series of selective scripts for reference to perform various scenarios (shown in Table 1). In each call for help event, the actors can arbitrarily choose a pair of dialogues and actions and freely play according to their understanding of the call for help events. For the real call for help samples, the actors will imitate the real urgent situations, such as elevator accidents or physical discomfort. For the fake call for help, actors will freely talk about some topics with "Help." or imitating some real call for help scenes in their conversations. Figure 1 shows the examples of the visaul dataset. The entire dataset is completed by a combination of given scripts for reference and free performances. Considering the actors are strangers to each other and the duration of collecting all the data is up to over half a year, the independence and diversity of our data samples can be guaranteed.



(a) Discomfort. (b) Knock on the door. (c) Fake category (conversation).

Figure 1: Examples of the visual data.

Data Annotation. To improve the data quality, we manually annotate all the synchronized audio and video data. The reason we have to manually annotate all the data is that there is no existing dataset for us to train a clipper on the keyword. Finally, 4168 audio-visual clips are cut out, 2076 of which are labeled as the real call for help events, and the other 2092 clips are labeled as the fake call for help events. Each clip has a one-second audio data and a five-second video data, with the keyword "Help" in the recorded conversation. We set the length of audio as one second for the reason that a one-second audio is guaranteed to contain the keyword we set and some contexts of the conversation. In contrast, a one-second video has limited information. Therefore, we use a five-second video history (include the keyword) to ensure enough video information. Through this proposed dataset, we transfer the proposed task to a binary classification problem based on audio-visual data.

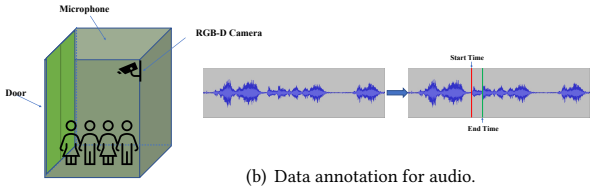
Data Statistics. We shuffled the data in pairs and split the training, validation, and test set with the ratio of $7 : 1 : 2$. At the same time, we control that one identity presents only in one set. The distribution of two categories is shown in Figure 3(a) and the distribution of duration of untrimmed data is shown in Figure 3(b).

4 PROPOSED METHODS

As our proposed task is specifically a multimodal learning task, we design the neural network following the idea of making full use of information from different modalities. Many researchers have studied multimodal fusion and it has become a classical problem [Atrey et al. 2010; D'mello and Kory 2015; Snoek et al. 2005].

Table 1: Selective Scripts for performing call for help events.

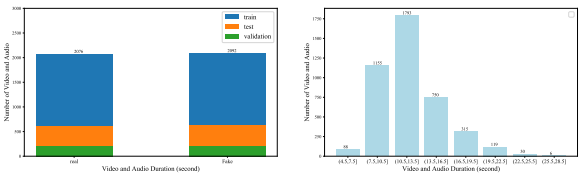
Contents of the Conversation (speaking in Mandarin)	Actions
"Help, please help me."	Stomp the feet, scratch the head.
"The elevator is broken, help!"	Dizziness.
"I'm trapped. Who can save me? help!"	Shake to left and right.
"Oh my god, someone fainted, help!"	Walk around in panic.
"Come and save her. Help me!"	Cry.
"I'm dying. Help me!"	Stand motionless.
"Do not hurt me! Help! Help!"	Flap the elevator doors and walls.
"Stay away from me! Otherwise I will call for help!"	Lay weakly on the floor.
...	...
Contents of the Conversation (speaking in Mandarin)	**Actions**
"When I came here, I saw someone fainted, the others were calling for help."	Stand calmly.
"I watched a movie called 'The Help!' yesterday."	Confused expressions.
"Last week, I was trapped in the elevator. I shouted 'Help! Help!' "	Happy.
"Nancy is so annoying. She always turns someone for help."	Worried.
"Can you show me how the actress call for help in the drama? "	Walk around in panic.
"Nancy is so nice. She helped the old man and saved his life."	Exaggerated expression.
"Do you know what to say when you want to call for help? "	Show other people a picture on the phone.
"Do you know what is 'help' in other languages? "	Wave hands in the air.
...	...



(a) Data collection.

(b) Data annotation for audio.

Figure 2: Data Processing. Figure 2(a) shows the elevator environment where we collect data. Figure 2(b) shows the data annotation pipeline for audios. We recorded start and end times for each keyword in the audios.



(a) Distribution of our data.

(b) Distribution of untrimmed data.

Figure 3: Statistics of our data.

However, due to the heterogeneity of different modalities, separate backbone networks are often required to extract features from different modalities and then fused together, which is usually time-consuming and results in a waste of the large model size. To achieve the balance between model accuracy and model efficiency, we propose a modality-wise dynamic architecture that adopts the merits of multimodal fusion and dynamic model inference in a flexible way. Figure 7 summarizes the framework of our instance-modality-wise dynamic neural networks. There are three components of the complete model: pre-trained VideoNet, pre-trained AudioNet, and the Final DecisionNet with a fusion block to fuse extracted audio-visual features. The Final DecisionNet will make a decision with the fused features obtained from two backbone networks.

4.1 Single Modal Backbone Networks

In order to better extract audio and video features, we build two backbone networks for audio and video inputs.

AudioNet. AudioNet aims to extract high-level audio features through the input of low-level audio features. Particularly, we treat the proposed task as an audio-based event recognition problem. Consider that we have a training set of audios, we denote the audio as A and the corresponding label as y, where $y \in \{0, 1\}$, $y = 0$ denotes A covers real call for help. With the original audio signals A in hands, we first compute MFCC (Mel Frequency Cepstral Coefficients) A^M with the number of filters in the filterbank as 40 and the number of cepstrum to return as 40 as the pre-processing method. Then we use feature extractor F^A to extract high-level audio feature matrix X^A whose dimension is 512. The feature extractor F^A we use is ResNet-18 [He et al. 2016].

VideoNet. Similar to AudioNet, VideoNet aims to extract high-level video features through the input of video signals. In this part, we treat the proposed task as a video-based event recognition problem. We denote the video as V and the corresponding label as y,

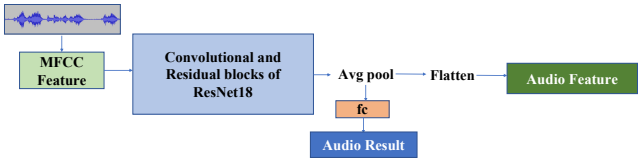


Figure 4: Pipeline of AudioNet. The feature matrix X^A is the flatten output of global average pooling layer.

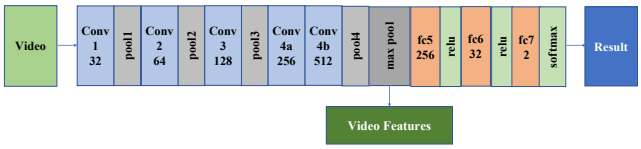


Figure 5: Pipeline of VideoNet. The feature matrix X^V whose dimension is the same as X^A is the flatten output of global average pooling layer.

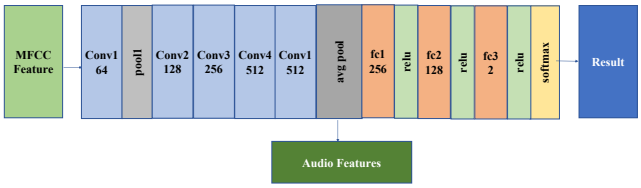


Figure 6: Struture of the modified VGGish. The input MFCC Features are the same as our AudioNet.

where $y \in \{0, 1\}$, $y = 0$ denotes V covers real call for help. We first compute normalized video signals V^N as the pre-processing method. Then we use feature extractor F^V to extract high-level audio feature matrix X^V whose dimension is 512. As 3D convolutions have shown competitive performance in tasks related to video-based action recognition [Feichtenhofer et al. 2019; Qiu et al. 2017], we build a 3D convolutional neural network to be the feature extractor F^V. Figure 5 shows the basic components of the proposed VideoNet.

4.2 Fusion Schemes.

Feature-level Fusion. Many feature-level fusion manners have been raised for multimodal input [Atrey et al. 2010]. We choose one simple but effective fusion manner that concatenates the audio features X^A and video features X^V into the fused features X^C whose dimension is 1024. More precisely, we obtain X^C whose shape is $(1, 1024)$ from two vectors whose shapes are both $(1, 512)$. In Figure 7, we denote the multimodal information fusion process as C.

4.3 Modality-wise Dynamic Network

As mentioned before, aiming at achieving the balance between model accuracy and model efficiency, we proposed an instance-wise dynamic neural network that flexibly adopts information from different modalities according to the difficulty of the input data. Due to the property of our proposed task of call for help recognition, the audio feature is necessary. Thus, it is mainly a speech-related task with video data as auxiliary information. Since there are two modalities in our dataset, we set a two-stage dynamic inference in the proposed model. For a given pair of audio and video data A, V, the AudioNet will first give out the primary predictions $Output_1$ based on audio MFCC signals A^M and the high-level audio features X^A.

At this stage, there exist two different cases:

- Case 1: The AudioNet is confident of its output.
- Case 2: The AudioNet is not confident of its output.

Being confident means for $Output_1$, the highest probability of two categories obtained by using a Softmax function is higher than a set threshold T, and vice versa.

The following steps for the two cases are different:

- Case 1: The final output is $Output_1$ from the AudioNet. There will be no following steps.
- Case 2: Video Features X^V will be extracted from the VideoNet. The Final DecisionNet will first fuse X^A, X^V and then get the output $Output_2$ after four fully connected layers.

The pseudo-code 1 shows the pipeline for the dynamic inference process.

Algorithm 1: pseudo algorithm for our two stage instance-modality-wise dynamic inference

```
Data: A, V
Result: Output_1 or Output_2
Compute X^A and Output_1 using the AudioNet ;
if AudioNet is confident of Output_1 then
    return Output_1;
else
    Compute X^V using the VideoNet ;
    X^C ← Fuse X^V and X^A ;
    Input X^C into Final DecisionNet;
    Output_2 ← Final DecisionNet(X^C);
    return Output_2 ;
end
```

4.4 Network Training.

The two backbone networks have been well trained on our dataset separately. Since the main function of two backbone networks is extracting features, the weights and parameters of the pre-trained feature extractors should not be updated during the training process of the Final DecisionNet. In the training process of the Final DecisionNet, the two backbone networks are frozen, and the network at this stage only learns how to get effective information from the fused features X^C.

We implement the networks based on PyTorch. Without otherwise stated, we use cross-entropy as the criterion function. For network optimization, Adadelta with weight decay by 10^{-5} is used as the optimizer for the AudioNet and Adam with initial learning rate 10^{-4} and weight decay by 10^{-5} as the learning rate for the VideoNet and Final DecisionNet.

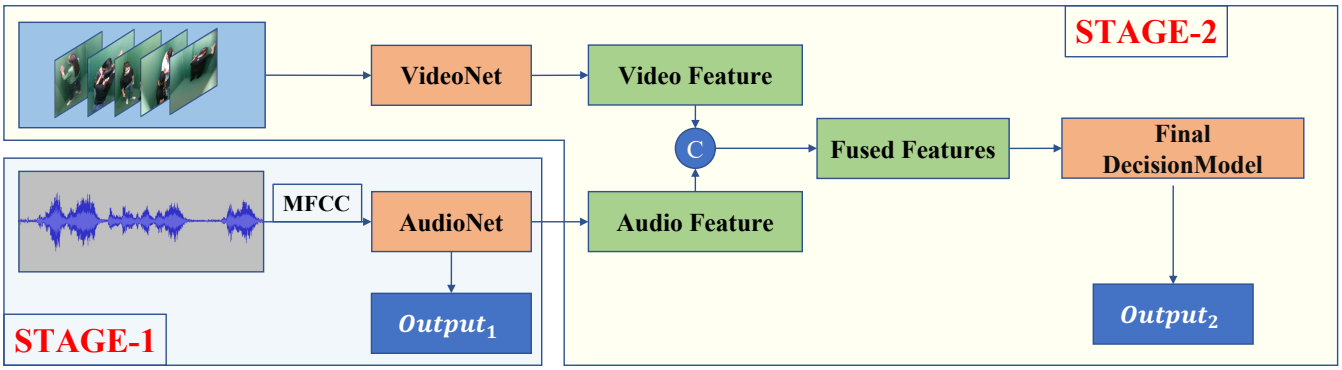


Figure 7: Proposed Instance-Modality-wise Dynamic Neural Network. The operator C represents the concatenation of feature vectors.

5 EXPERIMENTS

This section evaluates our method trained on the Elevator Help Dataset for the task of real and fake calls for help recognition in elevators. We evaluate the accuracy and F1 score for each model. We also compare the performances of each part of our model against other methods, such as ResNet 3D [Tran et al. 2018], VGGish [Hershey et al. 2017]. All the experiments are conducted with GPU Nvidia GTX1080ti.

5.1 Ablation Studies

The effect of modality. Most of the work still focuses on unimodality event classification. One advantage of our work is using audio-visual multimodal input. Therefore, we conducted experiments on our Elevator Help Dataset to verify the superiority of multimodal inputs. We tried three different inputs: audio, video, the fusion of audio and video. We also use a baseline model SVM with the fused feature X^C as the input. The result of them is reported in table 3.

The model with audio as the only input is the same as AudioNet in Figure 4, ResNet 18 [He et al. 2016]. The model with video as the only input is shown in Figure 5. And the model with multimodal input we use to evaluate is the structure shown in Figure 4 without the dynamic inference mechanism which only has $Output_2$ as the result.

We also provide a table to show the necessity of fusion of multimodal features (Table 2). In Table 2, "c" means the set of correct predictions while "f" means the set of false prediction. "cc" represents the number of samples in the set that both Stage 1 and Stage 2 gives out correct predictions; "ff" represents the number of samples in the set that both Stage 1 and Stage 2 gives out false predictions; "cf" represents the number of samples in the set that Stage 1 gives out correct predictions while Stage 2 gives out false predictions; "fc" represents the number of samples in the set that Stage 1 gives out false predictions while Stage 2 gives out correct predictions.

A key observation is that the audio modality performs better than the visual modality. It is not hard to understand: people's tone, pitch, and volume according to their emotions in emergencies will be more distinct in normal situations compared with visual signals.

Table 2: Correction of failure in Stage 1 using the multimodal network.

Method	cc	cf	fc	ff
Our Method of separate training	787	3	25	12
End-to-end training	352	38	425	12

Additionally, actions and gestures can change on purpose in both real and fake scenarios. For example, in the Elevator Help Dataset, a real call for help case can be that a person stands motionless in the cabin and shout for help; and a fake call for help case can be that person imitating the movie scene with his hands shaking in the air and laughing at the silly actions. Therefore, the multimodal model is better than the unimodal model.

The effect of fusion schemes. The fusion scheme we use is early fusion. To verify the superiority of early fusion, we experiment with the other two fusion schemes: late fusion and hybrid fusion. For late fusion, we do the result-level fusion. We simply add the output from the AudioNet and the VideoNet together and re-range them as the new results. One example for understanding is assuming $Output_1 = [0.4,\ 0.6]$, $Temporary\ Output = [0.3,\ 0.7]$. After the sum operation, the result will be $[0.7, 1.3]$ and the final re-ranged $Output_2$ will be $[0.35,\ 0.65]$. For hybrid fusion, we add a late fusion of $Output_1$ and $Output_2$ to our original model. The comparison of the results is shown in Table 3. It is clear that in our case, early fusion is the most efficient fusion scheme.

The superiority of the strategy of training the backbone networks separately. We stressed training the backbones separately instead of training the complete network in an end-to-end way. The potential problem of end-to-end training is that one backbone has been overfitting while the other one is still underfitting. Since we have little knowledge of the training detail of each part of the network, it is hard for us to adapt the optimization parameters and functions flexibly and instantly. Another trouble brought by the end-to-end training is that a well-designed weighted loss is needed for the training process.

Table 3: Performances of models with different modalities and fusion schemes.

Modality	Accuracy	F1 score	# Parameters	FLOPS(G)
Audio	95.53	95.54	11, 171, 266	0.19
Video	88.03	87.29	4, 659, 362	39.51
SVM (baseline)	49.46	49.46		
Multimodality (Our model early fusion)	98.19	98.19	16, 495, 046	39.71
Multimodality (late fusion)	95.77	95.74	15, 830, 628	39.71
Multimodality (hybrid fusion)	97.58	97.59	16, 495, 046	39.71

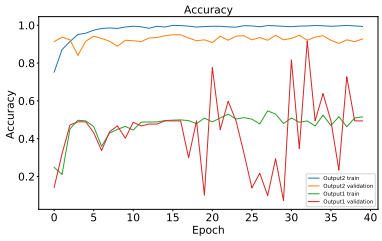
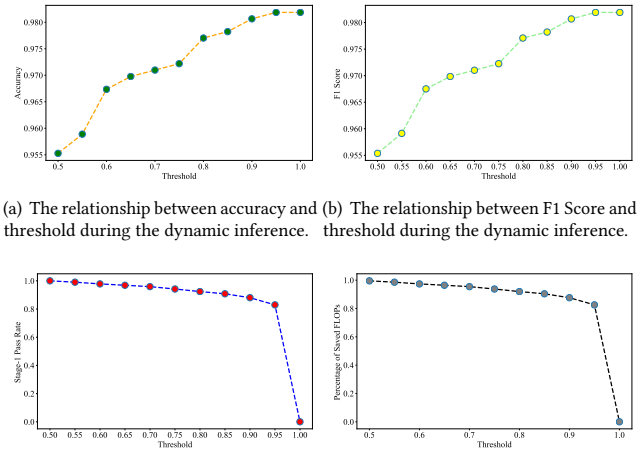


Figure 8: End-to-End Training Process.



(a) The relationship between accuracy and threshold during the dynamic inference.

(b) The relationship between F1 Score and threshold during the dynamic inference.

(c) The relationship between stage 1 pass rate and threshold during the dynamic inference.

(d) The relationship between FLOPS and threshold during the dynamic inference.

Figure 9: Effects of different parameters in dynamic inference.

In our experiments of end-to-end training, we design the weighted loss in equation 1: Assume for a given pair of audio and video A, V, the AudioNet gives out the output O_A with cross-entropy loss l_A, the VideoNet gives out the output O_V with cross-entropy loss l_V and the Final DecisionNet gives out the output O_D whose loss is denoted as l_D.

$$l_D = \alpha_A l_A + \alpha_V l_V \tag{1}$$

where α_A denotes the weight for l_A and α_V denotes the weight for l_V, and $\alpha_A + \alpha_V = 1$.

As it is hard to evaluate the impact of which feature has dominant positions in the determining process, we set α_A and α_V as 0.5, 0.5 in our experiments.

The two-stage data is shown in Table 2. And the change of accuracy of the two outputs during the training process is shown in Figure 8.

As shown in Table 2, it is clear that Stage 1 in the end-to-end training model is not as powerful as it is in the model trained with our strategy. And from Figure 6, we can tell that the AudioNet is barely working as the accuracy is only 0.5 which equals a random guess in a two-class classification. We think the reason is that the model tends to learn how to get effective information from the combined features instead of learning how to extracting effective high-level features from the multimodal inputs. The failure of Stage 1 will go through Stage 2 with larger amounts of computations and results in a decrease in efficiency.

The choice of our backbone network. We compare the performances of our backbone networks with other feature extractors to verify the efficiency of our AudioNet and VideoNet. To control the effect of dimensions of inputs, we use the same size of MFCC features and audio as the input. And during the process of training, we use the same optimizer and optimization parameters and train the models with the same number of epochs. We also let the models give out the same dimension of high-level audio and video features to control the effect of dimensions of the features. The result is shown in Table 4.

Table 4: The performance of different backbone networks

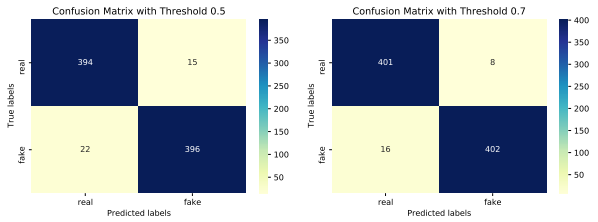
Modality	Model	Accuracy	F1 score	# Parameters
Audio	ResNet18	95.53	95.54	11,171,266
	Modified VGGish	89.12	89.36	7,989,634
Video	Our VideoNet	88.03	87.29	4,659,362
	ResNet 3D	55.82	22.46	33,204,930

For audio, we choose a slightly modified VGGish as a comparison. The network structure is shown in Figure 6. Although the number of parameters of our AudioNet is slightly larger than the modified VGGish, the performance of our model is much better. For video, we choose the ResNet 3D [Tran et al. 2018] as a comparison of our VideoNet. Our VideoNet is obviously better than ResNet 3D. One reason for the result is that the network structure of our VideoNet is much simpler than ResNet 3D who has a serious overfitting problem.

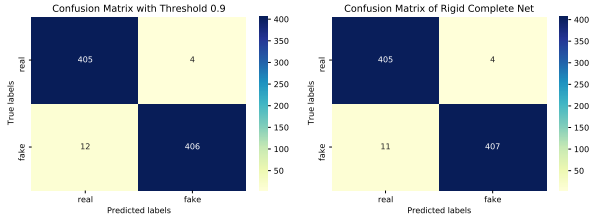
The efficiency of dynamic structures. As introduced in Section 4.3, our model chooses modalities wisely for different samples

Table 5: Results of model with different inference ways. The time we recorded is for our testing dataset which has 827 samples.

Inference way	Accuracy	F1 Score	Stage1 Passrate(%)	Time
Rigid AudioNet	95.53	95.53	100	10*s*
Dynamic with threshold 0.55	95.89	95.91	99.03	10*s*
Dynamic with threshold 0.65	96.98	96.98	97.82	10*s*
Dynamic with threshold 0.75	97.22	97.22	94.20	11*s*
Dynamic with threshold 0.85	98.07	98.07	90.81	12*s*
Dynamic with threshold 0.95	98.19	98.19	82.95	12*s*
Rigid Complete Network	98.19	98.19	0.00	25*s*



(a) The confusion matrix of AudioNet. (b) The confusion matrix of dynamic inference with threshold 0.7.



(c) The confusion matrix of dynamic inference with threshold 0.9. (d) The confusion matrix of complete net.

Figure 10: Confusion matrix of selected different networks.

in order to improve the efficiency of inference. To verify the validity of our dynamic inference structure, we use parameters such as accuracy, F1 score, pass rate, inference time, percentage of saved parameters to evaluate the model.

Table 5 and Figure 9 shows performances of the dynamic inference with different threshold. With the threshold higher, there will be more samples to be sent to Stage 2 and thus takes more computations, time with higher accuracy. The time that the rigid complete network cost is over twice than the dynamic network with the threshold as 0.95 whose accuracy is the same as the rigid network. It is clear that the dynamic network uses less time and memory without a huge loss of accuracy compared with a rigid multimodal network.

The confusion matrix of the dynamic networks with different threshold also verify our assumption that the instance-modality-wise dynamic network is efficient (See Figure 10).

6 CONCLUSION

In this paper, we study the call for help recognition with multimodal input under a dynamic inference structure. We define a new problem of identifying real and fake calls for help in elevators. And due to the lack of an applicable dataset, we collect a new dataset Elevator Help Dataset to fill the gap. Then we propose a method to better learn the combined multimodal features and make an instance-modality-wise prediction efficiently. Extensive experiments show, 1) our dataset is applicable; 2) multimodal features improve the performances of the model; 3) the dynamic inference method improves the efficiency of the model. Further improvements are expected by finding more sophisticated evaluation metrics of the dynamic network as well as exploring a policy decision method for dynamic inference. For future work, we plan to extend our work to include a system that enables to do multimodal training and unimodal inference.

7 ACKNOWLEDGEMENT

This research was funded by Kunshan Government Research (KGR) Funding in AY 2020/2021.

REFERENCES

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16, 6 (2010), 345–379.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

Guoguo Chen, Carolina Parada, and Georg Heigold. 2014. Small-footprint keyword spotting using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4087–4091.

Enea Cippitelli, Francesco Fioranelli, Ennio Gambi, and Susanna Spinsante. 2017. Radar and RGB-depth sensors for fall detection: A review. *IEEE Sensors Journal* 17, 12 (2017), 3585–3604.

Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 1–36.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6202–6211.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D Plumbley. 2013. Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 1–4.

Alex Graves. 2016. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983* (2016).

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. 2021. Dynamic neural networks: A survey. *arXiv preprint arXiv:2102.04906* (2021).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

Alan Higgins and R Wohlford. 1985. Keyword recognition using template concatenation. In *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 10. IEEE, 1233–1236.

Matthew B Hoy. 2018. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37, 1 (2018), 81–88.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q Weinberger. 2017. Multi-scale dense networks for resource efficient image classification. *arXiv preprint arXiv:1703.09844* (2017).

Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. 2014. Multimedia classification and event detection using double fusion. *Multimedia tools and applications* 71, 1 (2014), 333–347.

Ian McLoughlin, Haomin Zhang, Zhipeng Xie, Yan Song, and Wei Xiao. 2015. Robust sound event classification using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 3 (2015), 540–552.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2017. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 284–288.

Preksha Pareek and Ankit Thakkar. 2021. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review* 54, 3 (2021), 2259–2322.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.

Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine* 34, 6 (2017), 96–108.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829* (2017).

Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199* (2014).

Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*. 399–402.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*. 3551–3558.

William H Warren and Robert R Verbrugge. 1984. Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *Journal of Experimental Psychology: Human perception and performance* 10, 5 (1984), 704.

Martin Wolf and Climent Nadeu. 2014. Channel selection measures for multi-microphone speech recognition. *Speech Communication* 57 (2014), 170–180.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. In *European Conference on Computer Vision*. Springer, 322–339.

Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. 2019. Condconv: Conditionally parameterized convolutions for efficient inference. *arXiv preprint arXiv:1904.04971* (2019).

Lei Zhang and Xuezhi Xiang. 2020. Video event classification based on two-stage neural network. *Multimedia Tools and Applications* 79, 29 (2020), 21471–21486.