HUMAN PRESENCE DETECTION IN EMERGENCY SITUATIONS USING DEEP LEARNING BASED AUDIO-VISUAL SYSTEMS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

İZLEN GENECİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

AUGUST 2022

HUMAN PRESENCE DETECTION IN EMERGENCY SITUATIONS USING DEEP LEARNING BASED AUDIO-VISUAL SYSTEMS

submitted by **İZLEN GENECİ** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin _____
Dean, **Graduate School of Informatics**

Dr. Ceyhan Temürcü _____
Head of Department, **Cognitive Science**

Prof. Dr. Banu Günel Kılıç _____
Supervisor, **Information Systems Dept., METU**

Prof. Dr. Hüseyin Cem Bozşahin _____
Co-supervisor, **Cognitive Science Dept., METU**

Examining Committee Members:

Assoc. Prof. Dr. Barbaros Yet
Cognitive Science Dept., METU _____

Prof. Dr. Banu Günel Kılıç
Information Systems Dept., METU _____

Prof. Dr. Hüseyin Cem Bozşahin
Cognitive Science Dept., METU _____

Assist. Prof. Dr. Umut Özge
Information Systems Dept., METU _____

Assoc. Prof. Dr. Gökçe Nur Yılmaz
Computer Engineering Dept., Ted University _____

Date: 24.08.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: İZLEN GENECİ

Signature :

ABSTRACT

HUMAN PRESENCE DETECTION IN EMERGENCY SITUATIONS USING DEEP LEARNING BASED AUDIO-VISUAL SYSTEMS

GENECİ, İZLEN

M.S., Department of Cognitive Science

Supervisor: Prof. Dr. Banu Günel Kılıç

Co-Supervisor: Prof. Dr. Hüseyin Cem Bozşahin

August 2022, 94 pages

The significance of emergency event detection in surveillance systems has drawn the attention of researchers in recent years. Existing methods mostly depend on visual data to identify any abnormal events since only visual sensors are frequently put in public settings. On the other hand, in an emergency, sound information may be exploited. When eyesight is occluded, audio waves can penetrate to some extent. Applications for visual analysis may be helpful when there is noise in the audio and the scene is congested. Thus, the shift from single-modality to multimodality learning has become crucial given the recent rapid growth of deep learning. Both the audio analysis and the visual analysis were performed separately. In audio-based analysis, audio was transformed into samples using sliding window technique to capture the brief window of a target audio class. Therefore, in a real-time operating system, emergency circumstances can be recognized when the target sound happens briefly. For human sound classes of "Speech", "Scream", "Cry", the minimum sliding window sizes were 0.25 s, 1 s and 0.30 s, respectively. In visual analysis, face detection was conducted along with facial alignment using five facial landmarks. The AP for face detection was 77% on WIDER Face dataset (IoU=0.5). Using the detected faces, facial expression recognition (FER) was performed as well as age and gender estimations by employing an attention-based method. For seven basic emotions, 64.14% accuracy was achieved on AffectNet dataset. The combination of these audio and visual-based systems eliminates the limitations of perceptual tasks in both modalities.

Keywords: audio event detection, face detection, facial expression recognition

ÖZ

DERİN ÖĞRENME TABANLI İŞİTSEL-GÖRSEL SİSTEMLER İLE TEHLİKE DURUMUNDA İNSAN TESPİTİ

GENECİ, İZLEN

Yüksek Lisans, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Prof. Dr. Banu Günel Kılıç

Ortak Tez Yöneticisi: Prof. Dr. Hüseyin Cem Bozşahin

Ağustos 2022, 94 sayfa

Gözetleme ve arama kurtarma sistemlerinde acil durum tespitinin önemi son yıllarda araştırmacıların dikkatini çekmiştir. Mevcut yöntemler, genellikle ortamlara yalnızca görsel sensörler yerleştirildiğinden, herhangi bir anomali durumunu tanımlamak için çoğunlukla görsel verilere dayanır. Diğer yandan, acil bir durumda, ses bilgileri ayırt edilebilir anomali tespitine yardımcı olabilir. Görsel bilginin sınırlı olduğu durumlarda, ses dalgaları bir dereceye kadar nüfuz edebilir. Ayrıca, seste gürültü olduğu ve faaliyet alanının yoğun olduğu durumlarda görsel analiz uygulamaları yararlı olabilir. Bu nedenle, derin öğrenmenin son zamanlardaki hızlı büyümesiyle birlikte, tek modaliteden çok modlu öğrenmeye geçiş çok önemli hale gelmiştir. Görsel-işitsel bir sistem oluşturmak amacıyla hem ses analizi hem de görsel analiz ayrı ayrı gerçekleştirilmiştir. Ses tabanlı analizde, çeşitli ses olaylarının aynı anda gerçekleştiği gerçekçi ortamlarda hedef bir ses sınıfının kısa penceresini yakalamak için kayan pencere tekniği kullanılarak ses örneklere dönüştürülmüştür. Bu nedenle, gerçek zamanlı bir işletim sisteminde, hedef ses kısa bir süreliğine gerçekleştiğinde acil durumların tanınması hedeflenmiştir. İnsan sesi sınıfları "Konuşma", "Çığlık", "Ağlama" için minimum kayan pencere boyutları sırasıyla 0.25 s, 1 s ve 0.30 s olarak belirlenmiştir. " Görsel analizde, beş yüz işaret noktası kullanılarak yüz hizalaması ile birlikte yüz algılama gerçekleştirilmiştir. Yüz tespiti için Average Precision (AP) değeri WIDER Face veri setinde %77 olarak belirlenmiştir (IoU=0,5). Tespit edilen yüzler kullanılarak, dikkat temelli bir yöntemle yaş ve cinsiyet tahminlerinin yanı sıra yüz ifadesi tanıma (FER) gerçekleştirilmiştir. Yedi temel duygu için, AffectNet doğrulama veri setinde model tarafından %64.14 doğruluk elde edilmiştir. Bu işitsel ve görsel tabanlı

sistemlerin kombinasyonu, her iki modalitedeki algılama görevlerinin limitlerini ortadan kaldırmak için kullanılabilir.

Anahtar Kelimeler: ses olayı tespiti, yüz tanıma, duygu durumu tespiti

To Özlen, Süleyman and Berk Geneci

ACKNOWLEDGMENTS

I would like to start by expressing my gratitude to my supervisor Prof. Dr. Banu Günel Kılıç who made this work possible. Her advice and motivation helped me to a great extend to accomplish my thesis. I will always be grateful to her for giving me this opportunity, for guiding me and encouraging me throughout the way. I would also like to thank my co-supervisor Prof. Dr. Hüseyin Cem Bozşahin for supporting me, inspiring me and broadening my vision.

I had been a part of METU Center for Image Analysis (OGAM) for my studies where every person helped me and supported me. I would like to thank Prof. Dr. A. Aydın Alatan for providing me the opportunity to work in OGAM which made me discover my passion for research. I would also like to express my gratitude to Dr. Alper Koz, who helped me immensely, patiently answered all my questions and concerns and was the best colleague. I am grateful for my time in OGAM where I made the best friends and learned so much. I would like to thank Ogul Can, for listening to me, understanding me and always cheering me up. I thank Aybüke Erol who was the best friend I could have asked for and was always there to help me.

There are three people in my life that I have to thank for my every accomplishment in life, my mother, father and my brother. I thank them for the endless love and support they gave me throughout my life. They believed in me even when I did not. I thank my father for teaching me that I can do anything I set my mind to. I would like to thank my mother for always being there when I needed, in everything I do. I thank my brother for giving me a different view of life and his words of motivation. I would like to express my gratitude to my uncle Atalay Uçar who has been like a brother to me. I appreciate him for teaching me so much, expanding my perspective on life and for his sense of humor. I would also like to thank Hakan Emel for his support and for inspiring me to be the best version of myself. I appreciate his constant encouragement and words of wisdom.

Lastly, I would like to express my gratitude to my friends in TOGG, who I worked alongside with. I thank Özsel Kılınç for his advice and sharing his technical knowledge with me. I thank Güneş Çepiç for always listening to me, motivating me and providing me support in this process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AVL	Audio-visual Learning
AV	Audio-Visual
AV	Action Vector
AEC	Audio Event Classification
AED	Audio Event Detection
FT	Fourier Transform
STFT	Short-time Fourier Transform
MFCC	Mel-frequency Cepstral Coefficients
SVM	Support Vector Machine
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
CNN	Convolutional Neural Network
FCN	Fully Convolutional Network
FPN	Feature Pyramid Network
RPN	Region Proposal Network
R-CNN	Region-based Convolutional Neural Network
YOLO	You Only Look Once
SSD	Single-shot Detector
FER	Facial Expression Recognition
DAN	Distract Your Attention
MAN	Multi-head Cross Attention Network
AP	Average Precision
IoU	Intersection over Union

CHAPTER 1

INTRODUCTION

In order to protect individuals and ensure public safety, adequate surveillance must be provided. Recently, audio-visual learning, which aims to take advantage of the link between auditory and visual modalities, has attracted a lot of interest in surveillance applications. Among the notable accomplishments of the two modalities are voice recognition [9], facial recognition [10], and analysis of age, gender [11], and expression [8] on the face. The limitation of perceptual tasks in each modality has been solved by the introduction of audio-visual learning (AVL) using both senses. Additionally, examining the connections between auditory and visual information opens up additional relevant and fascinating study subjects, which in turn improves our understanding of machine learning [12].

1.1 Applications and Scenarios

The two distinct systems that make up the proposed system are the branches for audio and image processing. By working independently and fusing their results, both branches are able to overcome the constraints of a single modality problem. The system may be modified for different tasks and for certain use cases. Security surveillance systems can be one of the most important usage areas of the system. In surveillance systems, crowd analysis and anomaly event detection in crowds plays a significant role [13] because of its importance to public safety. A face detection and analysis system can help detect the people in a crowd as well as their emotions. When an abnormal occurrence takes place, it frequently comes with some unique sounds [14]. Thus, with the audio event detection, anomalies in a crowd such as fighting, fleeing, chasing, chaos can be detected [15].

A major area of use for this technology is in search and rescue situations, where people in distress need to be promptly located in inaccessible locations [16]. While visual information is commonly used with UAVs for detecting people in search and rescue operations, video sensors may be constrained by a lack of visual feedback owing to poor lighting conditions or obstructions restricting the field of vision. UAVs equipped with a microphone array may be of crucial assistance in locating individuals in emergency circumstances [16].

Another area the audio-visual perception system can be utilized is robotics. It takes common sense knowledge to handle a wide range of situations with complex semantics to interpret and understand when communicating with humans, as well as socially acceptable responses. Different AI techniques are required in the context of emotional design in order to enable robots to comprehend emotions as a part of the interaction process [17]. Robots will likely need a stronger awareness of the environment they operate in, to traverse and interact with it more fully. Both audio and visual information can be beneficial for the human-machine interaction.

Furthermore, the system can be used to monitor occupant behaviors and identify potential emergency events in SPH (single person household) settings. Emergency occurrences have a significant influence on the health of the occupant especially for older occupants [18]. According to [19], in the US, a lot of elderly individuals get injuries from falls every year, some of which can be deadly. According to a research, among adults 65 and over, one out of every five falls might result in serious injuries, such as fractured bones or brain trauma [19]. When an emergency arises, the system may recognize a person yelling and pleading for assistance, notify their emergency contacts, and also summon the emergency services. It can also detect the negative facial expressions in an emergency situation when audio information is noisy. Moreover, the system can be utilized to develop an intelligent baby caring/watching system. It can be used to prevent the infant from various harms including accidents that can occur such as falling or suffocation by covering mouth and nose [20]. Detection of the face and analysis of the facial expression can be useful in such situations. Auditory analysis can let the parents know of the sounds the baby making including crying or screaming.

1.2 Aims and Objectives

The goal of the study is to detect emergency situations using the advantage of both audio and visual information. We aim to increase the robustness by eliminating the limitations of perceptual tasks in both modalities. For this purpose, a detailed analysis in both audio-based and visual-based deep learning systems was conducted. For the audio-based system, a novel method using sliding window technique to capture the brief window of a human voice was proposed. For the visual-based system, a detailed facial analysis of facial expressions, age and gender was developed. For emergency cases, it is important to obtain high recall values, as false positives are more acceptable in these scenarios. Thus, detection of an emergency in either one of the two modalities is the focus of the study.

1.3 Scope of the Study

The drawbacks of the two modalities were cumulatively made up for using an audio-visual system, which tries to combine these two factors. For this purpose, audio and visual parts of the system were developed separately. The objective is to examine and

demonstrate which of the two modalities performs better in various settings. Since the system has a variety of the applications and usage areas, fusion of the visual and audio-based modules can be performed in various ways as needed for the specific application. Thus, a detailed analysis on audio and visual systems of the study was conducted separately and the fusion of the audio and visual modules was left at the conceptual level.

1.4 Outline of Thesis

The thesis is focused on audio-based and vision-based analysis. The rest of the thesis is organized as follows:

In Chapter 2, literature review was split into three parts. First part focuses on audio-visual systems, second part is audio-based systems and the third part is vision-based systems. The entire research in literature is focused on human detection in emergency situations.

In Chapter 3, System Overview was introduced. It consists of possible use cases and scenarios of the system as well as possible fusion methods of the separate audio and visual systems.

Chapter 4 focuses solely on the audio-based detection of humans in emergency situations. First the proposed methods, secondly, the performance metrics, then the results of the experiments were provided.

Chapter 5 focuses on the vision-based human detection in emergency situations. First the proposed methods, secondly, the performance metrics, finally, the experimental results were provided.

Chapter 6 focuses on the detailed discussions on the results of both modalities as well as the usage of the system for various use cases.

The thesis is concluded with highlights of the important points of the proposed methods and drawn conclusions from the experiments in Chapter 7.

CHAPTER 2

BACKGROUND INFORMATION

2.1 Audio-Visual Systems for Surveillance

Recent research has focused on the importance of emergency event detection in surveillance systems. It is possible for audio waves to partially pass through obstructions. On the other hand, when the audio is noisy and the scene is crowded, visual analysis could be more beneficial. Thus, given the recent explosive rise of artificial intelligence, the transition from single-modality to multimodality learning has become essential. Both audio and visual information are employed in audio-visual systems. The data being image or audio is important in determining the neural network model to be selected. Consideration of the audio-visual analysis initially, followed by consideration of the visual and auditory bases and their respective approaches separately, will make the research more understandable. In this title, audio-visual-based analysis systems and how these systems can be used in surveillance for detecting emergency situations will be discussed. An example multimodal events detection system architecture as explained in [21] can be seen in Figure 1.

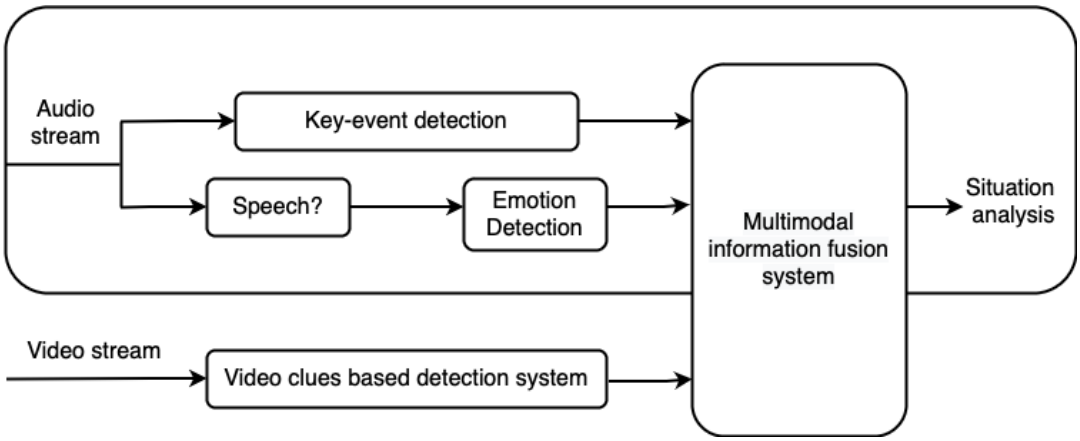


Figure 1: Multimodal event detection system architecture.

The use of only visual or only auditory variables in detection studies that are expected to contribute to safety and emergencies can create significant deficiencies. It can be more difficult to hide and disguise audio signals as opposed to visual data that can be easily covered or hidden. On the other hand, interpreting only on com-

plex auditory signals complicates visual imagination. Therefore, there is a need for studies in which visual and auditory signals can be evaluated simultaneously. Audio-visual learning (AVL), which aims to use these two variables together, was used to cumulatively compensate for the limitations of the two modalities. Discovering the relationship between these two variables and ensuring consistency is a promising area for machine learning [22]. When technologies that enable sound separation are combined with image-based detection technologies, they have begun to provide effective audio-visual separation [23]; [12]; [24].

Going beyond the unsupervised learning approach to make semantic connections healthier, AVL studies require a competent understanding of machine learning because of the extensive modalities that need to be synthesized. Unfortunately, real-world data such as images, videos, and audio do not have certain algorithmically defined properties [25]. Therefore, an efficient representation of data determines the success of machine learning algorithms. Recent studies seeking better representation have designed various tasks such as audiovisual correspondence (AVC) [26] and audio-visual temporal synchronization (AVTS) [12]. By making use of such a learned representation, the audio-visual tasks mentioned at the beginning can be solved more easily.

In [13], it is aimed to combine these two features by providing a different use of CNN in order to encode the Log Mel-Spectrogram for audio as well as the image data extracted from the videos with 3D CNN. In this study, which was carried out on the SHADE dataset, it was found that the use of voice performed significantly better in detecting abnormal events.

Although there are many studies that detect face and voice separately, there are very few studies on systems that can detect these two variables simultaneously. Although there are studies on speaker detection, which is generally important for video conferencing software, in the literature, the main purpose of these studies is detecting the relationship between the speech and faces [27]. This multi-stage method, which is based on the detection of head and facial movements, associating them with sounds coming from the microphone and facial expression analysis, has been tried to be used for emergencies in the following years and to be able to detect in environments outside the meeting interface.

[28] introduces an audio-visual speaker recognition technique that prioritizes the most salient features using feature-level audio-visual fusion and reinforcement. Any small flaw in the audio tracker could potentially cause feature-level defragmentation to drastically reduce tracker performance. They suggest a brand-new initialization technique that restricts the search region for the visual audience by employing the audio Direction of Arrival (DOA) angle for each speaker. This sophisticated fusion of audio and visual data minimizes the effect of noise. They demonstrate that even DOA from a noisy audio tracker can be effectively localized using the first face in a monitoring system when combined with our broad dictionary learning based classifier. They use a discriminate SVM classifier and a visual detector based on dictionary learning to detect faces. Audible DOA is used to limit the search region for visual face detection in order to strengthen the visual detection.

In a 2008 study [29], it is seen that the spatio-temporal-sound words created by combining the visual and auditory features of video clips with the Latent Semantic Analysis (LSA) model were used for event classification. Probability distributions of spatio-temporal-sound words were learned from training examples that include a series of videos representing different types of audiovisual events. This study, which is based on the concepts of space-time interest points ([30] , object class recognition and human action categorization, has made significant contributions to audio-visual event classification.

In the study of Nair et al. in 2019 [19], an alarm system with a low margin of error was tried to be established by detecting possible emergencies on the basis of sound and then confirming or rejecting them with the SCALE application, which was activated based on the event. The basic principle of the system, which is intended to be a security and telecare application for the age of 65 and above, especially in the United States with an increasing elderly population, is aimed at separating words such as help, ouch, hurt, and sounds such as groaning, shouting, crying, which are defined by the image and sound received from the Raspberry Pi. In future studies, it is aimed to transfer this system to smart phones, which are always with us, as well as all of us, with an energy-efficient application. In the study, in which machine learning was provided with the SVM model, Google AudioSet and Urban Sound Dataset were used, and up to 50% prediction accuracy was obtained in the first studies.

2.2 Audio Event Classification and Emergency Detection

In this study, the main focus of audio-based analysis is to detect the presence of people in emergency situations, such as surveillance and search and rescue operations. Their reactions, and the connection of the sounds with their emotions. Detecting the situations of people such as crying, shouting, and using exclamations to ask for help when they encounter any emergency situation by deep learning can provide significant benefits in cases where the person does not have the chance to call human operators and emergency hotlines to ask for help. These analyses can be used in the field of health. For example, an elderly living alone at home can send audible warnings to an integrated system in the case of falls and accidents, the separation of keywords in the case of an attack or violence, and even the automatic auscultation by identifying the organ sounds and the detection of vital anomalies. However, the main interest of the audio part of this thesis is to determine emergencies by focusing on human voices. The aim is to contribute in order to fill the lack of literature on this subject to detect humans in need of help more precisely by using both visual and audio information.

2.2.1 Audio Event Classification

Audio Event Classification (AEC) involves parsing sounds identified in a sound recording, just as with the analysis of images and object and face detection tasks. In this respect, AEC allows distinguishing what the stimulus sound is and with which label it should be identified, rather than analyzing what the content is. In another workspace, the Audio Event Detection (AED) tasks include, in addition to AEC, detecting the temporal onset and bias of each sound event in an audio recording. In both tasks, it is seen that the labels defined as inclusive have the purpose of controlling and distinguishing. The difference is that AEC focuses on the more spontaneous and superficial distinctions of active sound events and what is heard, while AED requires an additional explanation of the onset and offset times of the detected event. Although there are valuable new studies and developments in both areas, the focus of the thesis will be on the systems under the AEC title.

Important developments in audio detection/classification studies can be seen in the past. In [31], 81% recall rate for classifying gunshot sounds, 51% recall rate for classifying crying sounds, and 93% recall rate for classifying explosion sounds have been obtained with an approach that uses some fundamental properties, such as spectrum centroid and MFCC. The study [32] created a system for categorizing speech, music, and environmental sound events. Based on morphological features and key statistics, this model focuses on computable key features. Then, using Gaussian Mixture Models, they classified the environmental sounds into other classes, which were rain, birds, and applause. Google introduced WaveNet [33] in 2016 to generate raw audio data. Salamon and Bello presented the classification method with deep learning architectures by taking the environmental sound data in [34]. The dataset included urban sounds including 10 low-level classes:air conditioner, car horn, children playing, dog bark, drilling,engine idling, gun shot, jackhammer, siren and street music.

There are also some sub-techniques used in this respect. These are tasks that aim to convert data representation such as the "Fourier Transform", "Spectogram", and "Mel spectogram". Fourier transform (FT) is one of the transformation techniques to convert the time-domain signal into the frequency-domain signal. In the study [35], this method, which was used to understand the hidden anomalies in ECG values and to match the corresponding sounds with time series, provided significant advantages. Spectrograms, on the other hand, provide benefits, such as multi-label classification to identify more than one species that occur together in the same sound recording, and to determine the spectrum regions that are suitable for feature extraction. These approaches treat spectrograms as an image and find areas of interest suitable for extraction of statistical features.

Mel-Spectrograms are a different idea that entails a short-time Fourier transform (STFT) for each frame of the spectrum (energy/amplitude spectrum), from the linear frequency scale to the logarithmic Mel scale. Conversion of audio data into Mel-Spectrograms after processing is very important for deep learning trainings programmed to recognize neural networks. The overall complex and overlapping nature of audio data increases the importance of Mel-Spectrograms for the recognition of useful and distinctive features [36].

The study [37] proposed a new technique for classifying patterns called the nearest feature line (NFL). The NFL uses data from several prototypes for each class, as opposed to the commonly used NN (Neural Network), which does classification by comparing the query to each prototype separately. This study, which prioritizes particular and cepstral voice characteristics and combinations, has successfully resulted in an error rate of 10%. The study [38] conducted a comparison research using different audio characteristics and similarity measures. MFCC, linear predictive coding coefficients (LPC), sub-band energy distribution, and a few additional temporal/spectral properties were among the audio features that were compared. The study [39] suggested a technique for detecting acoustic occurrences in recordings made in the real world. They segmented the audio data and performed both classification and also found the positioning of the audio event. They achieved 24% accuracy rate classifying actual audio recordings which include background noise into 61 classes. The study [40] created a method that was put to the test in CLEAR 2007, which is an assessment workshop sponsored by NIST. CLEAR database consists of 18 audio event classes including speech, cough and laugh. The system relies on SVM classifiers and multi-microphone decision fusion, and it employs a collection of characteristics made up of frequency-filtered band energies and perceptual information. They achieved 30% recall and 20% accuracy.

The AEC tasks are based on the Convolutional Neural Network (CNN) architecture, just like its relative counterparts. There is confusion and high probability of overlapping of audio data, especially in the wild, which is attributed to poorly labeled AEC and AED systems. Some methods have been proposed to combat the high margin of error created by labels that are defined too tightly. A Fully Convolutional Network (FCN) based method, which makes it possible to learn from the weakly labeled data set without preconceptions, is recommended, and studies based on this method have taken their place in the literature. This system is called Weak Label Assumption

Training (WLAT). Likewise, the study [41] found that the FCN system can reveal better mAP outputs for both AED and AEC.

There are also different techniques used when detecting sound events and classifying them according to events. Most techniques rely primarily on tools to extract spectral-based features and then distinguish classes. It follows speech processing techniques with simple approaches, MFCC features and GMM (Gaussian Mixture Model)-based classification [42]. Other used properties are spectral properties, LPC (Linear Predictive Coding), perceptual properties and mixtures of properties. Common classifiers such as nearest neighbors, SVM, RBFNN and random forest are also used. DNN (Deep Neural Network) classifier consists of a multilayer feed-forward perceptron network. The study [43] compared two classifiers using an audio corpus extracted from the FreeSound dataset which contains the crowd, traffic, applause and music classes. In the research, SVM and restricted Boltzmann machine (RBM) was used, it was found that DNN was able to detect at a higher rate compared to SVM. This result supports that deep learning makes important contributions in sound event classification.

In addition to its technical inclusiveness, AEC's basic paradigm is based on matching the sounds that people use as guides while perceiving and living in the outside world with their psychological meaning counterparts. In other words, sounds that have a psychological meaning (laughter, siren, horn, barking, alarm, scream, cry, etc.) aim to label sounds as closely as a person can interpret cognitively when they hear it. In order to enable the reflection of the human perspective in the process of developing the machine learning capacity with methods such as deep learning, the study [44] tried to integrate the questions about the labeling and associations of the sounds played to the human participants into the training process in machine learning. Based on the 50 sound events in the ESC-50 data set, the Action Vectors (AVs) specified according to the aforementioned study added a new dimension to AEC studies by trying to find new alternatives for matching sounds with tags. According to results AEC accuracy using AV depends on the selection and number of actions can distinguish sound events. The better AVs can differentiate between such sound events by incorporating more actions that separate vocalizations. How we organize sound events is also influenced by the actions we choose.

2.2.2 Audio-based Detection in Emergency Situations

It is stated in the previous sections that the usage areas of audio-based event detection can vary from public transportation to health, education, and emergencies. In this title, studies on the importance of detecting sound classes such as screaming, speech, crying and environmental sounds in emergency situations will be included. In the 2016 study [45], it is aimed to automatically detect the screams and shouts that occur in abnormal situations in the subway train. In this study, audio data was collected from the real metro sound recordings of the Paris metro in order to isolate the noise of the noisy railway environment and the passengers. State-of-the-art Deep Neural Networks (DNNs), which are a combination of Restricted Boltzman Machines (RBMs)

and Deep Belief Networks (DBNs) applied to acoustic MFCC properties, were used as classifiers. Machine learning has been identified as problematic with 4 categories: screams, loud talking, conversation and noise environment.

On the other hand, [46] conducted a study in 2021 aiming to detect victims through sound in fire disasters that caused significant losses worldwide in the said year. Two machine learning (ML) methods were used in this study: support vector machines (SVM) and long short-term memory (LSTM). According to the findings of the study, although both methods showed superior performance in the detection of screams, it was found that SVM performed slightly higher than LSTM. Due to its low complexity, according to the researchers, in their autonomous embedded system tool, SVM is a preferable option for real-time implementation.

The data in the study [47] in 2020 to detect the cries of children with behavioral disorders are also noteworthy. A subset of the accessible AudioSet dataset and a set of audio data from the TV show Supernanny containing sounds close to the status of the selected disadvantaged group were used for the research, and these two sets were integrated with validations by adding manual explanations for the data. In the extended AudioSet clips, the model achieved a receiver operating characteristic (ROC) – area under the curve (AUC) of 0.86. In this respect, although there are technical and ethical dilemmas regarding the use of audio-based event detection technologies in homes and public places, it has been concluded that the use of such technologies for emergencies in clinical studies is quite suitable.

When we look at the studies in the literature, it is seen that various studies have been carried out in order to separate the sounds inside the house from the environmental sounds, especially in emergency situations. In solving this problem, the Gaussian Mixture Model for classification with Discrete Wavelet Transform (DWT), MFCC, LPC, Linear Frequency Cepstral Coefficients (LFCC) and Zero Crossing Ratio (ZCR) were the frequently used acoustic parameters in the recognition of speech sounds, for feature vector extraction. GMMs, Hidden Markov Model (HMM), Support Vector Machines (SVM) and k Nearest Neighbor (kNN) algorithms have been widely used. While MFCC and LPC algorithms provide advantages in word recognition, especially LPC algorithm has high performance in speaker recognition, modeling and compression of voice signals. However, for non-speech signals, the success of LPC drops dramatically, especially in noisy sounds [48].

In the paper [49], which was presented at the International Conference on Ubiquitous Robots and Ambient Intelligence (URAl) in 2016, the perception sensor network (PSN) model was introduced. In this model, there is audio-visual fusion to detect a single speaking person among multiple ones for multi-person scenarios. Introduced PSN can monitor an entire room with a low margin of error in multi-person scenarios to answer Who, What, Where questions. More specifically, emergencies consist of alarm sound (fire detector), human screaming, crying (may be violent). The system's two core modules for audio processing are sound source categorization (SSC) and sound source localization (SSL). The SSC must acknowledge that the sound source is either one of the emergency classes or ordinary speech, which can be disregarded.

Then SSL makes an educated guess as to where that audio source is, for example, when someone is yelling, various people in the room must figure out which is SSL.

In the study of [19], an alarm system with a low margin of error was tried to be established by detecting possible emergencies on the basis of sound and then confirming or rejecting them with the SCALE application, which was activated based on the event. The basic principle of the system, which is intended to be a security and telecare application for the age of 65 and above, especially in the United States with an increasing elderly population, is aimed at separating words such as help, ouch, hurt, and sounds such as groaning, shouting, crying, which are defined by the image and sound received from the Raspberry Pi. In future studies, it is aimed to transfer this system to smart phones, which are always with us, as well as all of us, with an energy-efficient application. In the study, in which machine learning was provided with the SVM model, Google AudioSet [50] and Urban Sound Dataset [51] were used, and up to 50% prediction accuracy was obtained in the first studies.

2.3 Face Detection and Facial Expression Analysis

Recent developments in the disciplines of image analysis and detection have made it possible to create useful video surveillance systems with integrated facial analysis features that are precise and helpful in emergency scenarios. It is crucial to use automatic detection and analysis techniques because manual surveillance is more difficult and time-consuming. Depending on the situation, surveillance systems used for security applications can be classified in a variety of ways, such as detecting theft, detecting aggression, detecting a person in need of assistance, and so forth. Due to this, in this title, which constitutes an important area in the analysis of the thesis, the results in the literature related to face detection, the importance of the alignment and mapping methods to be used in the analysis of the face, the important points in the analysis of facial expressions and new designs that can be revealed by associating all these concepts with deep learning will be included.

Accumulating facial data require instant comparison with precision and security. Face detection, which emerged in the field of computer vision, seeks a solution to overcome this big data-related problem by algorithms as these numerous amounts of data is impossible to be identified manually when both time and precision are addressed. The first data that is introduced to face detection applications try to differentiate the face(s) in the given image as input as the first task. Distinguishing the face from its environment successfully in a binary manner as a face/not face forms the first stage of detection. Since face detection is an object detection form, it would be more descriptive to talk about it first.

2.3.1 Object Detection

It would be more descriptive to briefly discuss object detection first since face detection is a type of object detection. Typically, a backbone is employed in many computer vision tasks, which is usually already trained on ImageNet [52]. The backbone is utilized in this manner as a feature extractor and it provides a feature map representation of the input. It can be difficult to find objects of various sizes, especially small ones. For this purpose, necks like Feature Pyramid Networks (FPN) [3] are employed. For object recognition, FPN takes the place of feature extractor in some detectors like Faster R-CNN and produces many feature map layers with higher quality information. The spatial resolution decrease in a feature pyramid as we go upward. On the other hand, the semantic value of each layer increases. The head in an object detection network gives the output as classification scores and bounding box coordinates.

In modern object detection systems, 4 main structures can be mentioned in general. These are, in order:

1. Input: Image, Patches, Image Pyramid.
2. Backbones: VGG16, ResNet-50, RegNet.
3. Neck:
    - Additional blocks: SPP, ASPP, RFB, SAM.
    - Path-aggregation blocks: FPN, PAN, BiFPN.
4. Heads:
    - Dense Prediction (one-stage):
        - RPN, SSD, YOLO, RetinaNet (anchor based).
        - CornerNet, CenterNet, MatrixNet, FCOS (anchor free).
    - Sparse Prediction (two-stage):
        - Faster R-CNN, R-FCN, Mask R- CNN (anchor based).
        - RepPoints (anchor free).

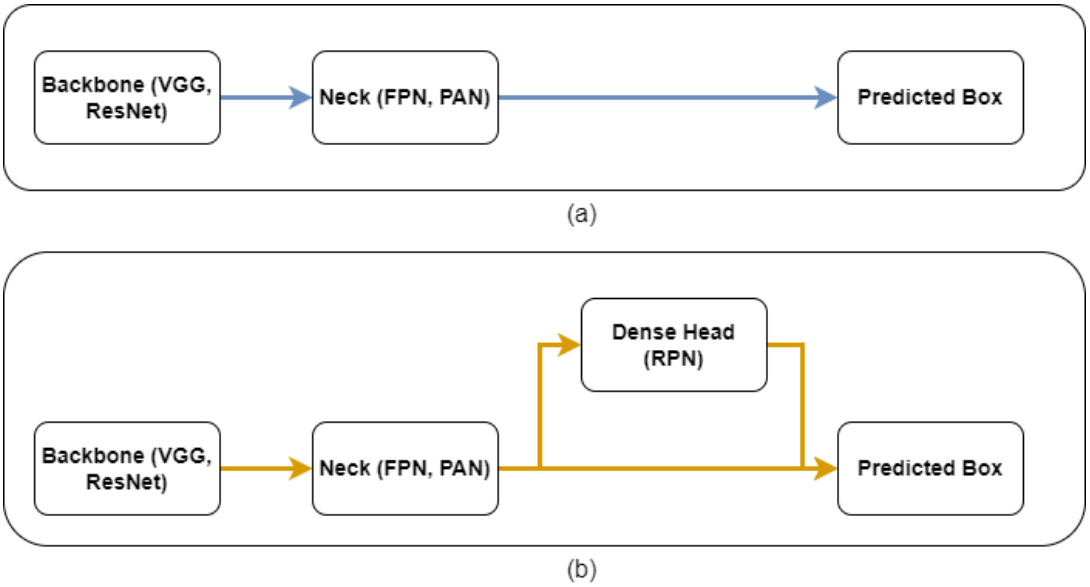


Figure 2: Example block diagram of (a) one-stage and (b) two-stage object detectors.

There are one stage and two stage object detection methods as seen in Figure 2. In one-stage object detectors, without employing pre-generated region proposals, object categorization and bounding-box regression are performed directly. In two-stage object detectors, region proposals are generated using selective search as in R-CNN [53] or a Region Proposal Network (RPN) as in Faster R-CNN [54].

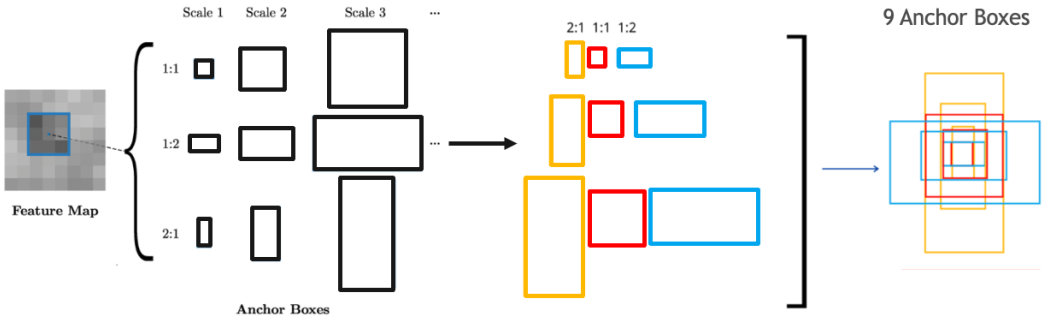


Figure 3: Generated anchor boxes for each position

Object detection studies are generally divided into two, anchor-based and anchor-free methods. A group of predetermined bounding boxes with a specific height and width are known as anchor boxes. These boxes are often selected based on the object sizes in the training datasets and are constructed to capture the scale and aspect ratio of particular object classes to be detected. In two-stage detection methods, a particular algorithm extracts region suggestions before CNN is used to identify and fine-tune the bounding box's location. Region suggestions are not necessary for single-stage detectors. They convert the position directly to the regression of the bounding box, typical methods include the YOLO [55] and SSD [56] series [57].

R-CNN [53] selects 2000 regions from the image to be extracted using selective search. These regions are known as region proposals. Therefore, we may now deal with 2000 regions rather than trying to categorize a large number of regions in an image. A CNN receives these region suggestions and outputs a 4096-dimensional feature vector. The CNN serves as a feature extractor, and the features it extracts are passed into an SVM to determine whether the object is present inside the candidate region suggestion. Fast R-CNN feeds the input image to the CNN instead of the region proposals. Faster R-CNN uses a different network to generate regions instead of using selective search algorithm, therefore it is faster. On the other hand, YOLO (You Only Look Once) performs class probability and boundary box detection from pixels, unlike R-CNN. The image is first separated into several grids. For each grid it generates two bounding boxes and class probabilities for those bounding boxes. YOLO directly optimizes detection performance while training on complete images. Being simultaneous and fast and having an estimation interface that will reduce the margin of error are its advanced features compared to R-CNN. However, the risks associated with positioning and localization errors are YOLO's weaknesses [55].

Another important object detection method is SSD. SSD also uses the idea of bounding box regression as in YOLO. However, this approach made it difficult to detect especially small objects [58]. In 2017, DSSD was sampled to replace the SSD reference network and moved the dataset from VGG to Resnet-101. In the end, it was possible to increase the target detection accuracy, particularly for small objects, although the detection speed drastically fell in the process. In 2018, Focal Loss (RetinaNet) [1]

was developed to alleviate the class imbalance issue. In the one-stage object detection situation, there is a severe imbalance between foreground and background classes during training. Rather than addressing outliers, focal loss is designed to address class imbalance by down-weighting inliers, making RetinaNet a good detection network. Since face detection task has a class imbalance problem, RetinaNet can be a good alternative as a face detector as well.

Anchor-free methods on the other hand, aim to break away from the limitations of the anchor-based methods. Direct object detection is accomplished by anchor-free detectors in two separate ways. First one is finding numerous pre-defined or self-learned keypoints and then bounding the spatial extent of the objects. These are called the keypoint methods. The other way is predicting the four distances from positives to the object border, using the center point or region of the item as the definition of positives [59]. CornerNet [60] introduces the corner pooling concept in order to localize the corners of a bounding box. The negative positions are punished for each corner depending on how close the projected corner is to the actual corner. In the event that the distance is below a threshold, a bounding box is constructed. CenterNet [61] introduces center pooling in which it takes the maximum values in both horizontal and vertical directions. Center pooling locates the largest value in the pixel's horizontal and vertical directions and adds them to determine whether the pixel is a center keypoint. To avoid complex anchor calculation, FCOS (Fully Convolutional One-Stage Object Detection) detects objects by pixel-by-pixel prediction based on the idea of semantic segmentation [62]. It uses pixel-wise prediction to find objects. It produces results that are cutting edge for one-stage detectors. FCOS sees locations directly as training samples. In other words, every place is either a positive sample or a negative sample. It is a positive sample if it fits inside a ground truth box and is identified as the ground truth box label. These developments in object detection have led to an increased interest in face detection studies as well.

2.3.2 Face Detection

Modern CNN-based object detectors, such as RCNN [53], SSD [56], YOLO [55], FocalLoss [1], and their extensions, have enabled significant advancements in face detection as well as object detection. Two kinds of face detection techniques are currently in use: single-stage techniques (e.g. SSD and RetinaNet) and two-stage techniques, including faster R-CNN. Detecting hard faces in uncontrolled environments is the goal of anchor-based detection frameworks. WIDER FACE [7] dataset is commonly used in this respect as it consists of faces in different occlusions, backgrounds and poses.

In a 2017 study [5], a Single Stage Headless (SSH) face detector was developed. SSH [5] combines detection and classification in a single step using simply convolutions, which speeds up inference time. It is also intended to be scale invariant. It detects faces from different depths of the network rather than depending on an external multi-scale pyramid as input for detecting faces of various scales. It also develops a context module which aims to utilize the context information surrounding the face for detect-

ing the faces. It simultaneously detects faces of different scales from different layers. In this way, it provides an advantage against its state-of-the-art peers in the WIDER FACE dataset [7]. SSH also produces cutting-edge results on the FDDB and Pascal-Faces datasets and provides 50 ms/image on a GPU. The reason for giving priority to speed in the mentioned study is the main problem of detecting multiple, complex and moving faces, especially in crowded environments. Unlike its previous counterparts, which performed the detection in two stages, SSH accelerates the detection by reducing it to a single stage. This is accomplished by layering a powerful convolutional detection module on top of layers with various strides, each of which has been trained for a suitable range of face scales. In a single network forward pass, SSH can handle many face scales at once, making it an effective detector of small faces. SSH also uses feature pyramids with context modules to expand the receptive field from Euclidean grids. The context module helps utilizing the information around the face such as hair and shoulders. Using context modules showed to be effective in face detection. S3FD [63] also develops scale-invariant networks to detect faces with different scales.

By emphasizing the features from the face area to find the faces that are obscured, FAN [64] suggests a brand-new anchor-level attention that will emphasize the characteristics of the face region. It considerably increases recall while maintaining speed for the occluded case face detection challenge. It achieves cutting-edge performance on public face detection benchmarks like WiderFace and MAFA when the anchor assign strategy and data augmentation techniques are integrated. MTCNN [65] uses three levels of deep convolutional networks in a cascaded framework to predict face and landmark locations from coarse to fine. It offers a solution for both face alignment and face detection. Convolutional networks are used in three steps of the process to identify faces and facial landmarks such the eyes, nose, and mouth. Mask-RCNN [66] improves Faster R-CNN by adding a branch for object mask prediction in addition to the one already there for bounding box identification. The backbone and a region proposal network make up the initial stage. These networks execute once for each image to provide a list of area suggestions. For each proposed region identified in the first stage, the network predicts bounding boxes and object classes in the second step.

One of the productions in the field is SRN (Selective Refinement Network), with Selective Two-step Classification (STC) module and the Selective Two-step Regression (STR) submodules, which use different level information. As a new module that achieves greater distinguishability and robustness features, three new contributions are made that address the three key aspects of face detection, namely better feature learning, progressive loss design, and link assignment-based data augmentation [67]. In the Selective Enhancement Network for High-Performance Face Detection study, a new one-step face detector is proposed that selectively introduces two-step classification and regression processes to an anchor-based face detector to reduce false positives and improve simultaneous localization. One of the critical points to consider with this approach is the risk of anchoring and reducing the detection speed of processes outside of neural networks such as NMS (Non-maximum Suppression). Double branch central face detector (DBCFace) as a pure convolutional neural network face detection method promises fast face detection without extra link design and NMS [68].

Another one of the most important face detection studies is Pyramidbox [69]. Pyramid Anchors, a semi-supervised technique, was presented to supervise the learning of high-level contextual features. To combine high-level semantic features and low-level facial features, low-level Feature Pyramid Network (FPN) is used to attempt to predict all faces simultaneously. A context-sensitive structure is added to the prediction network to expand its capacity and boost output accuracy. In this way, the Pyramid box made the contribution of targeted velocity and multiple face detection. Additionally, it helped with data-anchor-sampling, which was used to augment the training data and broaden the variety of training data for small faces.

With the article published in 2019, Deng et al. developed a single-stage face detector that performs face localization on a pixel basis at various face scales with the RetinaFace application [6]. RetinaFace performs both face detection and facial alignment. Utilizing both extra-supervised and self-supervised multitask learning, RetinaFace is effective. With the manual addition of five facial keypoints to the WIDER FACE dataset, RetinaFace's first contribution significantly enhances hard face detection [7]. This extra inspection signal improved the detection performance significantly. RetinaFace also performs 3D face reconstruction. RetinaFace uses a Feature Pyramid Network as neck, as well as a context module similar to the ones in SSD and PyramidBox for using the context information surrounding the face. In this module, instead of the usual 3x3 convolution, a deformation convolutional network (DCN) is employed over the feature maps to improve the context modeling capability. It also uses Cascade Multi Task Loss with multi-task loss. The first context module uses the anchors to predict the bounding box, while further modules use the regressed anchors to predict a more precise bounding box [6].

2.3.3 Facial Alignment

The majority of facial landmarks, often referred to as facial key points or facial feature points, are found around the eyes, mouth, nose, and chin. Face alignment or facial landmark detection are then used to find these facial landmarks once the faces in the photos are discovered by the face detectors. Figure 4 shows an example face bounding box and also five facial landmarks used for face detection and alignment. For the majority of facial applications, such as face verification [70] and recognition [71], expression recognition [72], facial attribution analysis [73], and solutions to other computer vision issues, we can obtain a significant amount of corresponding shape and texture information based on these landmarks with semantic meaning. So for facial analysis applications, face alignment is a fundamental and crucial task [74].

The effect of face alignment on facial analysis performance can be significant. Scaling differences, rotation angle or curvature of the face in the images captured from the camera, which vary according to the distance of the face from the camera, affect the performance. One way to achieve high recognition rates is to identify a canonical face template and map the various landmarks on individuals' faces, such as eyes, nose, or mouth, using basic linear conformal transformations such as scaling, shifting, and rotating to the locations designed for each individual in the template. In some studies,

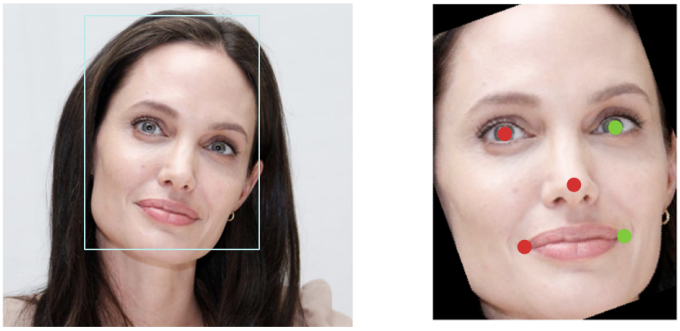


Figure 4: An example of a face bounding box and five facial landmarks

the importance of this issue has been tried and reported in studies in the literature [75] [76] [77]. In our study we use 5 facial landmarks; 2 for the eyes, 1 for the nose, 2 for the sides of the mouth.

Facial alignment is aiming for a canonical face alignment based on translation, size, and rotation. There are various ways of performing facial alignment. Some techniques merely use the facial landmarks themselves to obtain a normalized rotation, translation, and scale representation of the face, particularly the eye regions. According to this techniques, first aim is for all of the images in the dataset to be centered. The second aim is that the eyes to lie on a horizontal line. The final aim is that the sizes of faces are roughly the same.

2.3.4 Facial Expression Analysis

Sentiment analysis studies in the field of computer vision are called Automatic Facial Expression Recognition, shortly FER (Face Expression Recognition). Among the features of face detection, facial expression recognition is one of the most critical components that reveal people's emotions, intentions, and responses to stressors within the limits of their control. Voluntary and involuntary movements and micro mimics in facial expressions provide essential information about the mood and intentions of individuals. In this aspect, image and video-based facial expression recognition technologies can contribute to many conceivable terms such as preventing crimes and recognizing perpetrators, preventing suicide, helping victims in hostage-taking and violent crimes, and facilitating the management of other emergencies such as safety driving. FER applications are also used to build a more natural interaction between humans and computers, especially the humanoids as they can recognize and analyze the human emotions. As well as being used by security forces in the field of surveillance and behavior analysis, FER applications are used in different areas such as automatic smile detection in digital cameras [78].

People communicate and reflect to stressors by their expressions. Body language and facial expressions shapes the basis of face-to-face communication between people. The effect of facial expressions within communication is approximately 55% in the whole of communication. Due to this, facial expression detection and analysis have a great role to understand people's emotional situation [79]. Facial expression recognition makes use of distinctive expressions on the face. According to the study of Ekman and Friesen, seven types of facial expression of emotions in human face are universal [80]. These emotions are happy, sad, anger, surprise, fear, disgust and neutral, and in some cases contempt [78].

The study of facial expressions was once a topic of study for psychologists. In [**?**] preliminary findings were presented from a study on the automatic analysis of facial expressions from an image sequence. Then, as the technology advanced, studies on facial expression analysis were hastened by breakthroughs in the disciplines of face detection, face tracking, and face identification [81]. Human emotions are the result of a variety of circumstances, whereas facial expression recognition is related to the analysis of facial movements based solely on visual input. Human emotions can also be expressed through a variety of channels, including gestures, voice, gaze direction, posture, and facial expressions. To measure the facial expressions correctly, there is a critical need for determining the landmarks of the face. Eye and mouth regions are taken into account in obtaining the features of facial expressions. While determining facial expression, the regions where the changes caused by the expressions are most evident are very important in correctly recognizing the expression. Therefore, it is necessary to determine the eyes and mouth area correctly.

There are three fundamental phases in facial expression analysis. The first phase is the face detection. Then, feature extraction of facial expressions from the image, and the last phase is recognizing the facial expression. Changes, particularly in some areas of the face, cause facial emotions to appear. The foundation of facial expression recog-

nition investigations is the identification of these fleeting alterations in the eyebrows, eye circles, nose, lips, and chin regions and facial skin caused by the contraction of one or more facial muscles. The process of feature extraction is crucial for recognizing face expressions. There are two forms of feature extraction: appearance-based and geometric-based. The above-mentioned expressions, such as grin, sad, anger, disgust, surprise, and fear, are classified through a variety of important procedures. Eyes, mouths, noses, eyebrows, and other facial features are included in the geometrically based feature extraction, while the exact part of the face is included in the appearance-based feature extraction [82].

In the survey [83], it explains that the two types of FER systems are static and dynamic. While only spatial information is used to encode characteristics in static FER, temporal relationships among continuous frames are taken into account in dynamic approaches. Overfitting is common in small facial expression datasets. In order to overcome this, pre-trained networks like AlexNet [84], VGG [85], and VGG-face [86] were improved. Performance can be boosted by fine-tuning using more FER datasets. However, because the network was trained independently of FER, the learnt features still contain information that was dominated by faces. Two-stage FaceNet2ExpNet [87] method was suggested as a solution for this. As a result, the fine-tuned face net serves as an initialization and is solely utilized to direct the learning of the convolutional layers. Finally, FER data is used to train the fully linked layers from scratch. Some techniques cut off the elements of the face that are not essential. For example, the eyebrows, eyes, and mouth are three crucial regions of interest (ROI) for tasks involving expression identification. Additionally, several studies suggested automatically learning the essential components for FER. This is essentially what attention networks accomplish. Similar to cognitive attention, attention technique is the ability to concentrate on the relevant or important part of an image. Attention modules were first introduced in NLP in the paper [88]. Distract Your Attention (DAN) [8] consists of three components, FCN (Feature Clustering Network), MAN (Multi-head cross Attention Network) and AFN (Attention Fusion Network). In FCN, an affinity loss is proposed to maximize the inter-class distance and minimize the intra-class distance. There are parallel cross attention heads in MAN. These are distinct. Spatial attention and channel attention are the two components of a cross attention module. The attention maps are scaled by AFN using log-softmax. Then the section that is most intriguing is highlighted [8].

In addition to the expressions in which the emotions are evident, there are also emotions called micro mimics that can be read with certain vague indicators. 2019 study [89] on FER-2013 [90], CK+ [91] and JAFFE [92] datasets focused on different parts of the face and tried to contribute to the explanation of complex emotions. In this respect, it tried to put forward models with high sensitivity in order to map and describe the local manifestations of certain emotions in the eye and mouth landmarks. Starting in the top-left corner of a picture, researchers zero out a square section of size NxN inside the image each time, then use the trained model on the occluded image to predict the outcome [89].

In the study of Wen et al. in 2021 [8], a Multi-head cross Attention Network (MAN) approach is presented, a method that focuses on different parts of the face and makes

it easier to reach the most accurate result with both in-class and in-class comparisons. The prediction accuracy is over 60% in AffectNet-8, AffectNet-7 [93], RAF-DB [94] and SFEW 2.0 datasets, providing multiple attention areas in terms of basic logic and independent simultaneous reading of different parts of the face without overlapping. Further development of this system could promise significant benefits, especially in detecting the correct underlying affect in conflicting facial expressions [8].

2.3.5 Age and Gender Analysis

In intelligent applications like access control, human-computer interaction, enforcement, marketing intelligence, and visual surveillance, age and gender estimation from a single face image is a crucial task because age and gender, the two key facial attributes, are fundamental to social interactions [11]. Age may be represented by a single variable accepting non-negative real values, but gender can be represented by a single variable that can take binary values (male or female) [95].

Estimation of age and gender in relation to facial detection technologies is also among the fields of study. Many larger and deeper CNNs have been proposed with promising performance, such as AlexNet [84], VggNet [96], GoogLeNet [97] and ResNet [2]. However, the methods mentioned are not mobile and fast. In a 2019 study by Zhang et al., age estimation was attempted with the Compact yet efficient Cascade Context-based Age Estimation model (C3AE) [98]. It uses a two-point age representation. A distribution over two neighboring bins serves as the representation for any age point. It simultaneously uses classification and regression, and distribution learning. It also uses a context module which uses three-scale images of the face. One of the main problems in estimating age is that it is a parameter that is individual and difficult to generalize. In this study, cascade training was used alongside deep learning. The KL-Divergence method is adopted. It was found that the applied training and learning created a statistically significant difference compared to the experimental applications in which these procedures were not followed.

CHAPTER 3

SYSTEM OVERVIEW

Emergency event detection in surveillance systems has piqued the interest of researchers in recent years because of its importance to public safety. Because only visual sensors are commonly installed in public spaces, existing approaches mostly rely on visual information to determine there are any abnormal events. Sound information, on the other hand, may be discriminating in an emergency situation, assisting the surveillance system in determining whether there is an abnormality. Audio signals have a degree of penetration when vision information is easily blocked [13]. Visual analysis applications, on the other hand, can be more useful when the scene is crowded and audio information contains noise. Thus, with the rapid advancement of artificial intelligence in recent years, the shift from single-modality to multimodality learning has become critical for improved machine perception.

To eliminate the limitations of perceptual tasks in both modalities, audio-visual learning (AVL) using both modalities has been introduced. Furthermore, investigating the link between auditory and visual information leads to more intriguing and valuable study subjects, as well as improved machine learning perspectives [12]. Using only one camera and one monaural microphone, this thesis proposes an approach for integrating audio and visual information for scene analysis in a surveillance scenario. The block diagram of the proposed architecture is shown in Figure 5

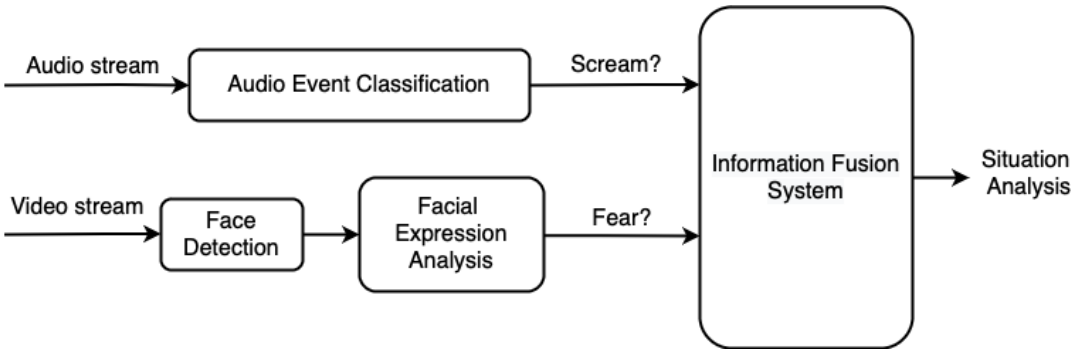


Figure 5: Block diagram of the proposed system.

The proposed system consists of two independent systems which are audio and image processing branches. To eliminate the limitations of a single modality task, both branches work separately and their outcomes are fused. The system can be adjusted

for specific use cases and for various tasks. The different settings for the usage of the system are discussed in this section.

For the security surveillance systems and search and rescue operations, an emergency detection from only one of the audio or visual analysis branches can be identified as an emergency situation. Since the important thing is to detect the emergency with higher precision is the aim in these use cases, false positives are more acceptable. The aim is to detect the emergency situation using audio or visual information. These two systems can support each other when one of them is more efficient in detecting emergencies. This also provide assurance when both of the systems detect an abnormality. For various other cases, The situation analysis can be performed assigning weights to the scores of audio and visual systems, combining them and setting a threshold for the output of the system to be classified as emergency.

In the following chapter of the thesis, human presence detection is performed in audio-based analysis part. The class which the audio data belongs to was analyzed using audio event classification. If the audio belongs to "Scream" or "Cry" classes, there is a possibility that there is an emergency, when only audio data is considered. If a human voice is to be detected in search and rescue operations or surveillance systems, the audio data being "Speech" can be helpful as well. Other than human sound classes, some environmental sound classes were chosen to perform the classification. The chosen sound classes are the ones which can be easily found in the background of a surveillance system. These are "Alarm", "Dog", "Engine", "Phone", "Crash" and "Footsteps" classes. The audio data is processed using sliding windows. For a real time audio data, the system can give an output once in a determined time frame. The audio is split into minimum window sizes and analyzed for the audio analysis section. The largest determined minimum window size for an audio clip to be classified is 1 s, which is explained in detail in Chapter 3. Therefore, the system can produce an output every second or in a specified time interval longer than 1 s.

In the audio-based analysis of the thesis in Chapter 4, for human detection task, face detection is performed. After detecting the faces, facial alignment was employed before the facial analysis. The analysis include facial expression recognition (FER) as the most important part. An attention based network was utilized for the expression recognition task. The model was trained and tested using public datasets in both face detection and facial expression recognition. Furthermore, the FER dataset was labeled automatically using a public pre-trained model for age and genders. The facial expression recognition model was also trained for age and gender using these labels. Information about age and gender can be helpful in a variety of circumstances in determining whether an emergency situation exists. A baby can be given as an example where crying and screaming may not be a sign of emergency, however for an adult it can be. As the output, a live demo using a webcam was implemented which first performs face detection, draws a bounding box and shows the determined age group, gender, and facial expression of the detected person. The demo shows the working system in real time.

CHAPTER 4

AUDIO-BASED ANALYSIS

In surveillance or security systems, automatically detecting emergency circumstances which rely on audio and visual indications are commonly utilized. In some cases, audio information may be more beneficial than visual systems in detecting specific incidents. As a result, when visual systems fail to detect such situations, audio-based solutions can be useful. The learning capabilities of deep learning architectures can be used to build sound classification systems that can be used in various fields including surveillance systems.

The purpose of the audio-based analysis section of this thesis is to distinguish emergency circumstances from ordinary sound events using deep learning networks. Classification of human sounds and environmental sounds is also another aim of this part. A detailed examination of how a deep learning-based audio system can operate in a real time audio-visual system was carried out. A pre-trained network was employed for the classification of the audio data. The network and the dataset that is used for the audio classification task were explained in detail. A novel method was proposed, which uses sample-based analysis of the audio data.

Audio analysis utilizing deep learning algorithms makes a prediction for the entire audio file, according to the literature. For circumstances where more than one audio event occurs, target sound detection is more difficult. For this purpose, audio was converted into samples to capture the small window of a target audio class. Thus, when the target sound occurs for a brief period of time in a real-time operating system, emergency situations can be detected. An analysis on the classification of environmental sounds as an addition to human sounds was also conducted. The proposed methods and results of the experiments for the audio-based analysis are discussed in this section.

4.1 Proposed Methods

In the audio-based identification of emergency circumstances, the classification of audio event categories plays a significant role. A Deep Neural Network-based model was used in the audio-based analysis of the thesis to classify the sound event classes that are important for this purpose. This section discusses the model and dataset that was used, as well as the proposed methodologies for the study.

For the detection of the target sound events, audio event classification (AEC) was performed in small windows of audio data. For this audio classification task, a network pre-trained on Google AudioSet [50] was employed. Google AudioSet is made up of 2,084,320 human-labeled 10-second sound clips that were taken from YouTube videos and an increasing ontology of 632 audio event classes (cite: Yamnetpaper). The ontology is described as a hierarchical graph of event categories that includes a variety of sounds made by people and animals, musical instruments and genres, and typical everyday environmental sounds. It can be said that sounds are quite diverse in different categories and provide rich sources for many studies in this field. In the AudioSet ontology, human sounds include human voice, digestive sounds, heartbeat sounds, respiratory sounds. However, only human voice classes are used in Section 4.1.1 and Section 4.1.2 for analysis since the primary aim is to detect them. In Section 4.1.3, in addition to human voice classes, some environmental sound classes were also employed. Human voice classes in the AudioSet ontology can be seen in Figure 6.

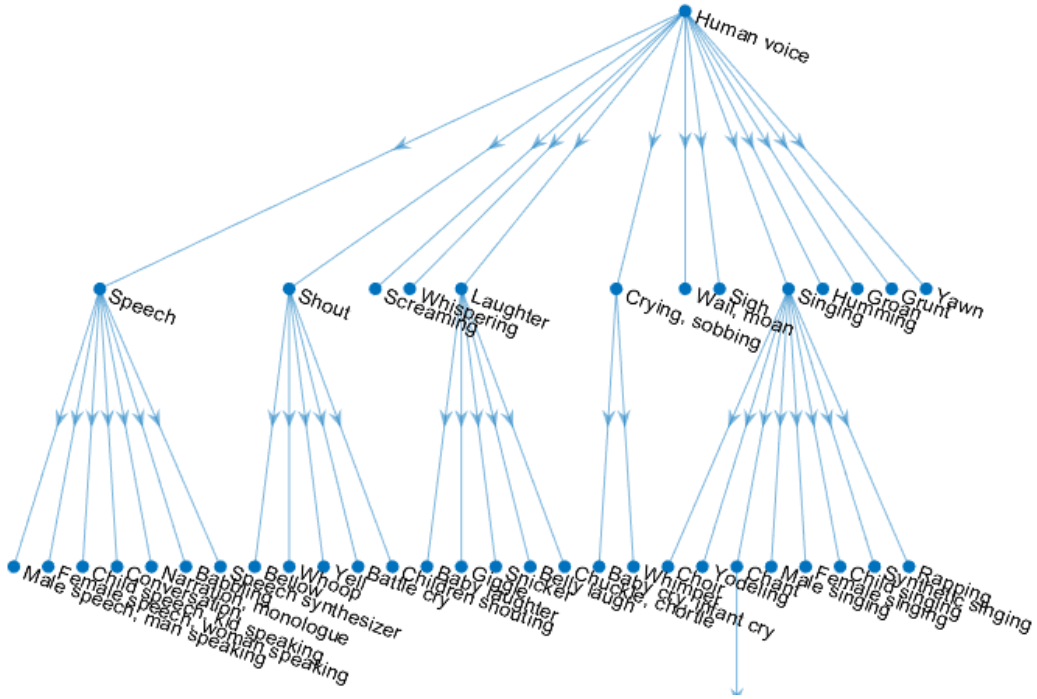


Figure 6: "Human voice" sound classes in the AudioSet ontology.

YAMNet pre-trained deep network [99] was employed for the classification of audio data. YAMNet is based on Mobilenet v1 [100] CNN architecture. It predicts 521 sound classes based on Google AudioSet and was released by Google's Sound Understanding team for large scale audio event detection applications. The YAMNet model's features are calculated by splitting audio into 960 ms frames with a hop of 480 ms and extracting the embeddings on a batch of these frames. A short-time Fourier transform (STFT) is used to break down the frames, with 25 ms windows applied every 10 ms. STFT is computing Fast Fourier Transforms (FFT) sequentially on windowed segments of an audio signal. The audio signal is converted from the time domain to the frequency domain using FFT. A spectogram is the outcome. After that, the spectrogram is divided into 64 mel-spaced frequency bins, each of which is log-transformed. This produces 96 64 bin log-mel spectrogram patches, which are used as input to the YAMNet model [99]. The YAMNet model returns 3 outputs, which are the class scores, embeddings and the log mel spectrogram. The class scores are then sorted and the top-scoring classes are the predictions of the model. YAMNet process is shown in Figure 7.

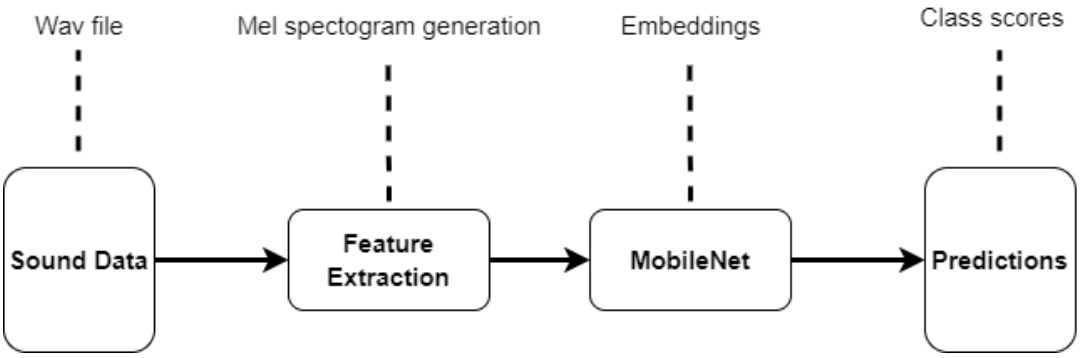


Figure 7: YAMNet running process.

Visualization of the waveforms of audio signals which belong to the "Female Speech", "Baby Cry" and "Female Scream" class from the NIGENS dataset, the log-mel spectogram and top-scoring classes predicted by the YAMNet model can be seen in Figure 8, Figure 9 and 10 respectively.

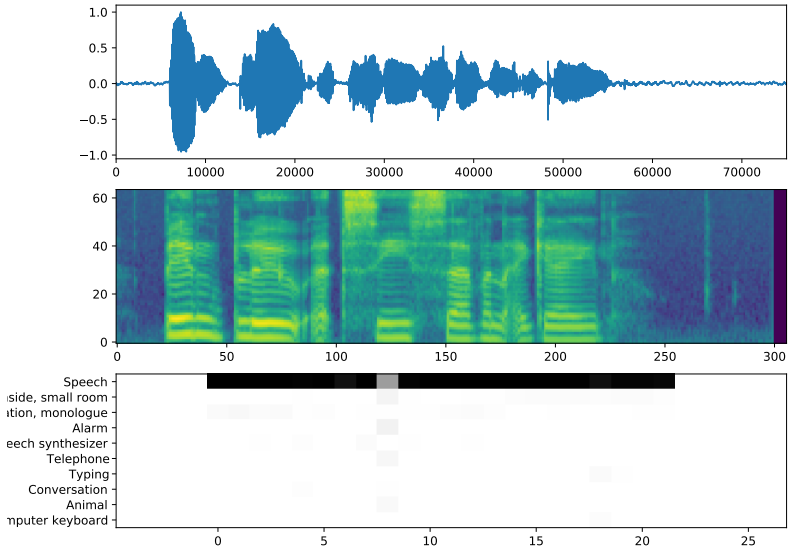


Figure 8: Visualization of (top) waveform, (middle) log-mel spectogram and (bottom) top-scoring classes of an audio signal from "Female Speech" class inferred by the YAMNet model.

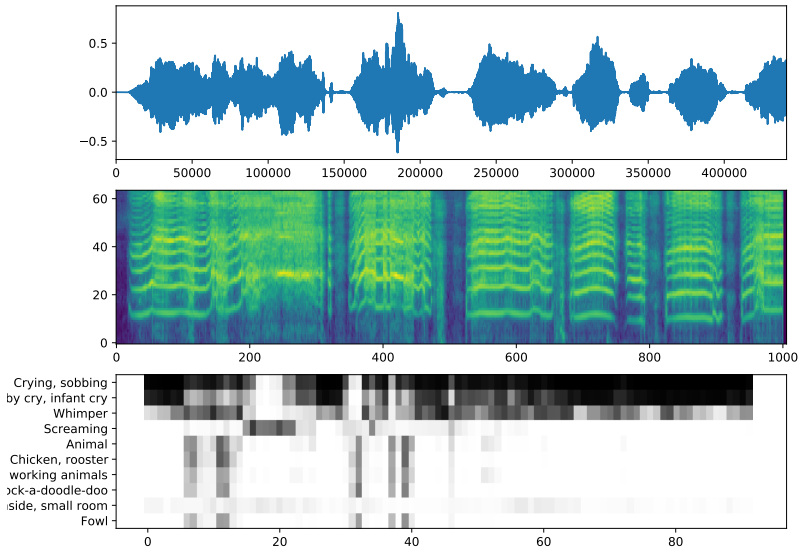


Figure 9: Visualization of (top) waveform, (middle) log-mel spectogram and (bottom) top-scoring classes of an audio signal from "Baby Cry" class inferred by the YAMNet model.

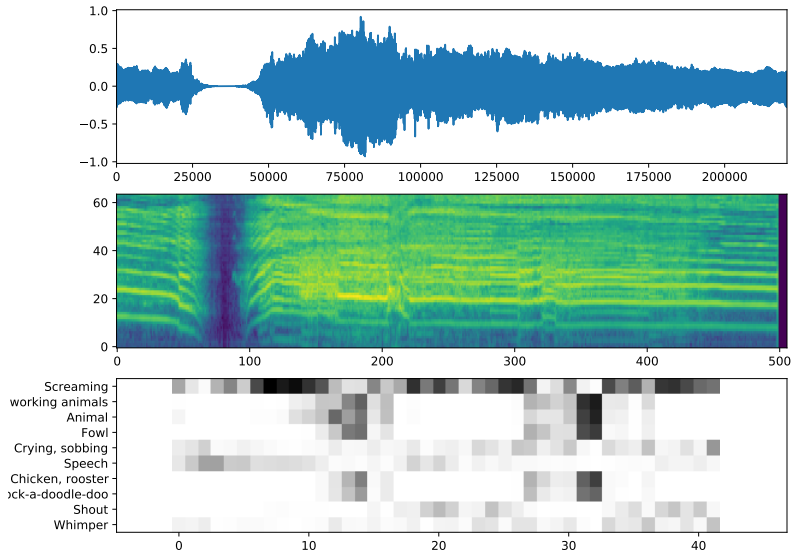


Figure 10: Visualization of (top) waveform, (middle) log-mel spectogram and (bottom) top-scoring classes of an audio signal from "Female Scream" class inferred by the YAMNet model.

For the analysis, audio data is gathered from The NIGENS General Sound Events Database [101]. NIGENS is a publicly available dataset which contains fourteen distinct sound classes. The classes it contains are "Alarm", "Baby Cry", "Crash", "Barking Dog", "Running Engine", "Burning Fire", "Footsteps", "Knocking on Door", "Female Speech", "Male Speech", "Female Scream", "Male Scream", "Ringing Phone" and "Piano". NIGENS database consists of isolated sound events, meaning that there is only a single audio event class for all audio clips and no background noise. Furthermore, the database provides ground truth strongly labeled with onset and offset times, which enables detailed evaluation.

Since the purpose of the thesis is to detect human sounds in emergency use cases, five audio event classes from the NIGENS database are chosen for the target sound detection in Section 4.1.1 and Section 4.1.2. The audio event classes that are chosen for the analysis are "Female Speech", "Male Speech", "Female Scream", "Male Scream" and "Baby Cry". For Section 4.1.3, environment sound classes as well as human voice classes are included for the classification. These classes are "Alarm", "Crash", "Barking Dog", "Running Engine", "Footsteps" and "Phone". The dataset consists of audio clips of various durations. Thus, they are pre-processed so that they contain audio clips of the same length in a class. Information of the characteristics, size, sampling rate and number of audio segments is shown in Table 1.

Table 1: Sound classes obtained from NIGENS database, with their characteristics and information.

Class	Characteristics	Num files	Size(s)	Sampling Rate(Hz)
Female Speech	Females calmly speaking short sentences.	10	3	25000
Male Speech	Males calmly speaking short sentences.	10	2	25000
Baby Cry	Sequences of cries, sobs and squeals.	3	10	44100
Female Scream	High-pitched and short screams of females.	9	5	44100
Male Scream	Short single screams of males.	7	5	44100
Alarm	Sounds from old-fashioned fire bells to electronic beeps	8	5	44100
Crash	Crashing, noise-like, but sudden, bursting sounds	10	3	44100
Dog	Dogs barking, mostly several times in a row.	5	10	44100
Engine	Long continuous sounds of running engines or idling.	5	10	44100
Footsteps	Diverse sounds of (individual) people walking	6	10	44100
Phone	Mostly classic phones, sequences of long ringings.	10	5	44100

In surveillance systems, real-time detection of emergency situations is critical in both audio and visual applications. In audio-based analysis, detection of the target sound classes needs to operate real-time in potential emergency scenarios. Furthermore, in realistic environments, variety of sound events occur simultaneously. Therefore, it is important to capture the emergency sound classes within an audio stream in such environments. For this purpose, audio-based analysis is performed utilizing sliding window technique. Instead of making one prediction for an entire audio file, the system analyzes the audio data in parts. Proposed sliding window method requires finding the minimum window sizes target sounds can be correctly classified.

4.1.1 Finding Minimum Window Size of Audio Data

In real-time audio classification systems, it is important to find the minimum window size an audio data can be classified correctly. For the aim of finding the minimum window sizes, experiments on window sizes of audio files were conducted. For the five target sound classes “Female Speech”, “Male Speech”, “Female Scream”, “Male Scream” and “Baby Cry”, the experiments were performed separately. 14 different window sizes ranging from 0.08 seconds to 3 seconds were chosen for all five classes. The analysis on each class was performed on approximately 600 samples for each class.

In the process of finding the minimum window sizes for each class, raw audio files should be pre-processed to eliminate the parts of the audio files that have no sound activity. Pre-processing steps of the experiments were performed on MATLAB. In the elimination process, the ground truth text files were used to determine the parts of the audio containing sound activity. Parts of audio containing the target sound were joined together to obtain a single audio file which only consists of the target audio class. After obtaining the new audio files for all classes which consist of only the target audio class, the files were split into different window sizes. The window sizes used in the experiment are 0.08, 0.1, 0.125, 0.15, 0.2, 0.25, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0, 2.5 and 3.0 seconds. Since the YAMNet model uses 0.96 second frames, audio

files shorter than 0.96 second window sizes are zero padded before being analyzed by the model. For all experiments of a class, same audio data was used with different window sizes. The audio data shown in Table 1 were split into windows and the number of audio segments obtained from this process can be seen in Table 2. As can be seen on the table, number of audio segments of an audio class decrease as the window size increases. This means that the results with larger window sizes are based on a small number of cases, and therefore can be more uncertain.

Table 2: Sound classes, window sizes and number of audio segments of the corresponding window sizes.

Class	Window Size (s)	Num files
Female Speech	0.125	410
Female Speech	0.25	230
Female Speech	0.50	110
Female Speech	1	50
Male Speech	0.125	270
Male Speech	0.25	150
Male Speech	0.50	70
Male Speech	1	30
Baby Cry	0.20	297
Baby Cry	0.30	180
Baby Cry	0.50	117
Baby Cry	1.50	36
Female Scream	0.25	351
Female Scream	0.50	171
Female Scream	1	81
Female Scream	2	36
Female Scream	3	9
Male Scream	0.25	273
Male Scream	0.50	133
Male Scream	1	63
Male Scream	2	28
Male Scream	3	7

After the pre-processing step, all audio files were classified with the YAMNet model and the predictions, their corresponding scores and true positives were recorded. The classification steps of the experiment were performed using TensorFlow library in Python. Further statistical evaluation was performed using the SPSS platform. The statistical evaluation includes true positive analysis for all class types, the correlation between the true positives and prediction scores and the analysis of misclassified classes.

4.1.1.1 True Positives and Class Types

After the pre-processing step, audio files with the specified window sizes were given to the YAMNet model as input. All audio files were classified with the model and the highest scored predictions of the model were recorded. True Positive (TP) represents

a highest scored prediction being a correct or incorrect classification, therefore, it is 1 or 0. For all classes, average true positives with respect to class types were determined for all window sizes. Also, “Male Speech” and “Female Speech” and “Male Scream” and “Female Scream” classes were gathered as “Speech” and “Scream” classes for analyzing average true positives with respect to general class types.

4.1.1.2 True Positives and Scores

Highest scored predictions and their corresponding scores were recorded for true positives and score analysis. Highest scored predictions and their scores were used to determine the certainty of the predictions when the they are correct. The scores when the predictions are not correct were also investigated. Lastly, to determine whether to use only the first prediction of the model or the first two highest scored predictions, scores when the first predictions are incorrect and the second predictions are correct were analyzed.

4.1.1.3 Misclassified Classes

The incorrect predictions of the system are important to determine which classes the target classes were misclassified as. For this purpose, the incorrect class names from the first predictions of the model were used. This process was performed for “Speech”, “Scream” and “Baby Cry” main classes and percentages of incorrect classes were analyzed.

4.1.2 Target Human Voice Detection Using Audio Data

In this section, the pre-processed audio clips were classified using the YAMNet model. The target sound are classes “Female Speech”, “Male Speech”, “Female Scream”, “Male Scream” and “Baby Cry”. The human sound classes used in this section were obtained from the YAMNet class map, which include the classes from AudioSet ontology. According to the Google Audioset ontology, each target class have a subset of other classes. Thus, some of the Audioset classes should be accepted as correct for a particular class. Table 3 shows the classes that are a subset of our target class. These were accepted correct if the output of the model is one of them.

For the target sound detection, the audio data needs to be classified as one of the matching audio classes to be accepted correct. For example, if classification of a speech audio clip is to be accepted correct, it should be predicted by the model as ’Speech’, ’Conversation’,’Narration, monologue’,’Child speech, kid speaking’.

All target audio files that were given in Table 1 were used in this part of the experiments. Sliding window tests were done using 0.25, 0.25, 0.30, 1 and 1 s windows, all 50% overlapping for the "Female Speech", "Male Speech"" "Baby Cry", "Female

Table 3: Target sound classes and the subset audio classes from Google Audioset ontology.

Target Class	Accepted Audioset classes
Screaming	'Children shouting', 'Shout', 'Screaming', 'Whistle', 'Groan', 'Gasp', 'Wail, moan','Yell'
Speech	'Speech', 'Conversation','Narration, monologue','Child speech, kid speaking'
Crying	'Crying, sobbing', 'Baby cry, infant cry', 'Whimper'

Scream" and "Male Scream" classes respectively. These window size values are the minimum window sizes found in the experiments in Section 4.1.1.

The YAMNet model returns 3 outputs, which are the class scores, embeddings and the log mel spectrogram. The model gives 521 scores for each audio clip since there are 521 classes in the Google AudioSet. The top-scored class is the first prediction of the model. Here, both the top-scoring classes were used for the analysis.

The recall values for all classes were calculated according to this classification. Precision was not used as a performance metric in this part since there are only human voice classes in this test set. When environmental sounds were introduced to the problem as in Section 4.1.3, we could talk about false positives. Therefore, precision, recall and F-1 scores were also calculated.

For the visualization of the analysis, the audio signals were plotted. Every sliding window of an audio signal has a classification score. These scores were also plotted secondly. Then, the classification results were obtained for each window and they were plotted. These were followed by the ground truth plots for each sliding window. The score of a window is represented as the middle point of the sliding window. Since only human voice classes are included in the test dataset and no negative classes, only recall values were calculated.

4.1.3 Target Audio Class Detection in Noisy Environments

In this section, the pre-processed audio clips of environmental noise were classified along with human voice classes using the YAMNet model. For the environmental noise, "alarm", "crash", "barking dog", "running engine", "footsteps" and "ringing phone" classes from the NIGENS general sound events database [101] were used. For all of the classes, audio clips of 1 second were extracted, since it is the shortest window size that can be chosen for classifying all human voice classes. In Section 4.1.1 the shortest window sizes were chosen as 0.25 s, 0.30 s and 1 s for "Speech", "Cry" and "Scream" classes respectively. The common window size for which all of the human voice classes can be classified correctly can be chosen as 1 s. Thus, for the evaluation of all of these classes together, the window size was chosen to be 1 s. Here, for the classes other than human voice classes were included. Therefore, Google Audioset ontology was again used for determining the accepted classes for these classes. Table 4 shows the target environment sound classes and their equiva-

lents in Google AudioSet ontology. If the model gave one of these predictions as output, it was accepted correct. A confusion matrix including the human voice classes and environmental sound classes was plotted. Some predictions of the model were outside of the chosen classes, thus, they were labeled as "Other". Precision, recall and F-1 scores were calculated for all classes.

Table 4: Target environment sound classes and the output classes from Google Audioset that are accepted as correct.

Target Class	Matching AudioSet classes according to Google AudioSet ontology
Alarm	'Car alarm', 'Alarm','Alarm clock', 'Smoke detector, smoke alarm', 'Fire alarm', 'Siren', 'Civil defense siren', 'Emergency vehicle', 'Police car (siren)', 'Doorbell', 'Buzzer', 'Vehicle horn, car horn, honking', 'Air horn, truck horn', 'Bicycle bell', 'Whistle'
Engine	'Fire engine, fire truck (siren)', 'Engine', 'Light engine (high frequency)' 'Engine knocking', 'Heavy engine (low frequency)', 'Medium engine (mid frequency)' 'Car', 'Accelerating, revving, vroom', 'Idling', 'Engine starting'
Phone	'Telephone bell ringing', 'Telephone', 'Ringtone', 'Telephone dialing, DTMF'
Crash	'Smash, crash', 'Breaking', 'Crushing'
Footsteps	'Run', 'Shuffle', 'Walk, footsteps'

4.2 Performance Metrics for Audio-based Analysis

For the detection of the target audio classes in an audio file, recall values are used. Audio files are split into windows and the windows are analyzed separately. Each window consists of only one audio class. The ground truths include the start and end times of an audio event for the corresponding audio file. All audio windows were fed to the model and highest scored predictions were obtained as the outcome. Because the Google AudioSet has 521 classes, the model assigns 521 scores to each window. The model's first prediction is the top-scoring class. For this analysis, both the top-scoring class and their related scores were plotted as well as ground truth values and the audio signals.

The ground truths are split into windows as well as the audio clips for the analysis. Since both ground truth and results are binary values, recall values can be computed using the result and ground truths. Since in Section 4.1.1 and 4.1.2 there are no negative classes (environment sounds), there cannot be false positives. This means that precision and F-1 score cannot be computed for these sections. In Section 4.1.3, for the classification of human and environment sound classes, a confusion matrix was provided. Here, precision, recall and F-1 scores were computed for each class.

4.2.1 Confusion Matrix

A visual tool used in learning methods is the confusion matrix, which is primarily used to compare classification outcomes to actual data. The matrix's columns represent instances of the class, while its rows reflect the predicted category of actual examples. In the confusion matrix, every instance can be divided into one of four types which are TP (True Positive), FP (False Positive), FN (False Negative) and TN (True Negative) respectively, as shown in Figure 11.

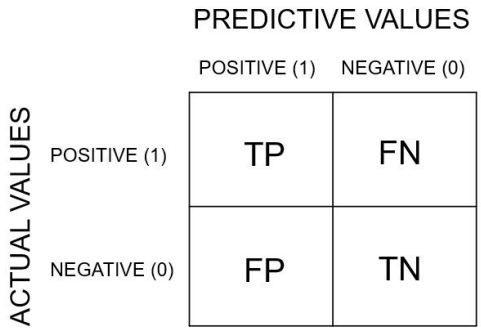


Figure 11: Confusion matrix.

4.2.2 Precision and Recall

Precision and recall are performance indicators used to evaluate the performance of a prediction model in machine learning and pattern recognition systems. The fraction of relevant samples among the samples that are predicted positive is referred to as precision. It is the ratio of true positives (TP) to the sum of true positives and false positives (FP) and defined as:

$$Precision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FP)}. \tag{1}$$

The fraction of retrieved instances among all relevant examples is referred to as recall. It is the ratio of true positives to the sum of true positives and false negatives (FN) :

$$Recall = \frac{TruePositive(TP)}{TruePositive(TP) + FalseNegative(FN)}. \tag{2}$$

4.2.3 F1-Score

F1-score is the combination of precision and recall of a classifier which takes their harmonic mean. It is commonly used for evaluating the performance of machine learning applications. It is denoted as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{3}$$

4.3 Results

The experimental results of the proposed methods on audio-based analysis are discussed in this section.

4.3.1 Finding the Minimum Window Sizes

Experiments on finding the minimum window sizes of audio clips are performed on the specified audio classes.

4.3.1.1 True Positives and Class Types

For the first experiment, "Female Speech" and "Male Speech" classes were merged into "Speech" class and "Female Scream" and "Male Scream" classes were merged into "Scream" class. "Cry" represents the "Baby Cry" class of the NIGENS dataset. Figure 12 shows the sound classes in window sizes and their mean detection. Here, detection is 1 when an audio clip of the specified window size is correctly classified and 0 when it is not. Mean detection is the average of all audio clips of that window size.

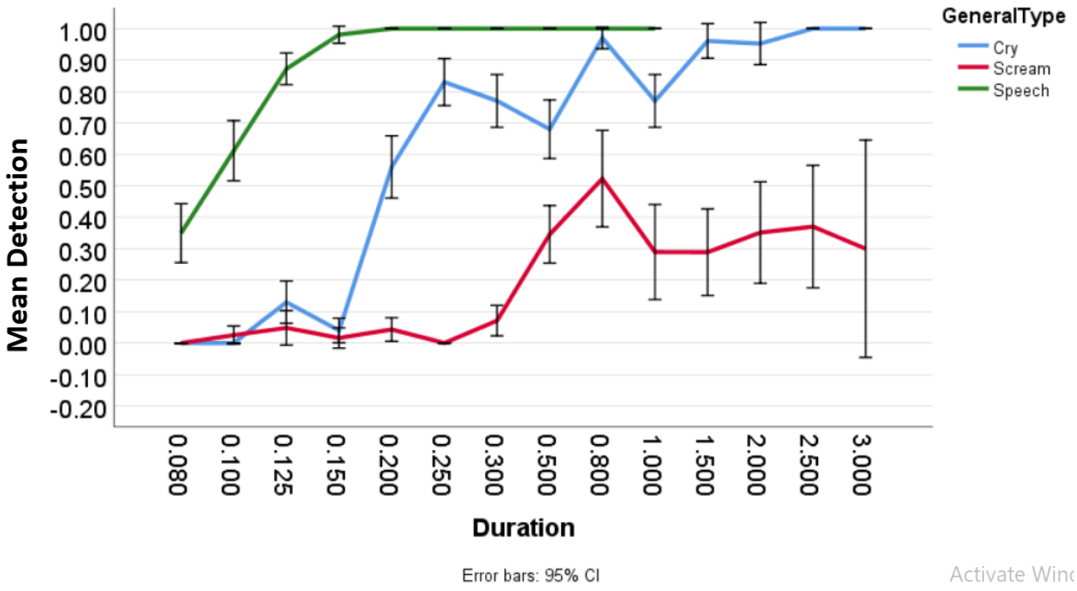


Figure 12: Mean detection of general sound classes for different window sizes.

"Speech" class gives the best true positive rate with short window sizes of audio, The model predicted "Speech" class with 1 true positive rate with audio clips longer than 0.25 seconds. "Baby Cry" class is gives the best results after 0.30 s threshold. The prediction of the "Scream" class does not change after 1 s. For more detailed view, the same analysis was performed splitting "Speech" and "Scream" classes with respect to genders.

As discussed in Section 4.1.1, number of audio segments decrease as the window sizes increase. This results in fewer audio clips for larger windows. The error bars in Figure 12 indicate the 95% confidence interval. This means that for 95% of experiments, the range of values within the confidence interval includes the true mean. A confidence interval depicts the range of values in which the true mean for the total samples is most likely to fall. Confidence intervals are used to assess whether the results are statistically significant. Here, since there are fewer samples with larger window sizes, the error bars are longer. This means that the values are more spread out and less reliable with large window sizes.

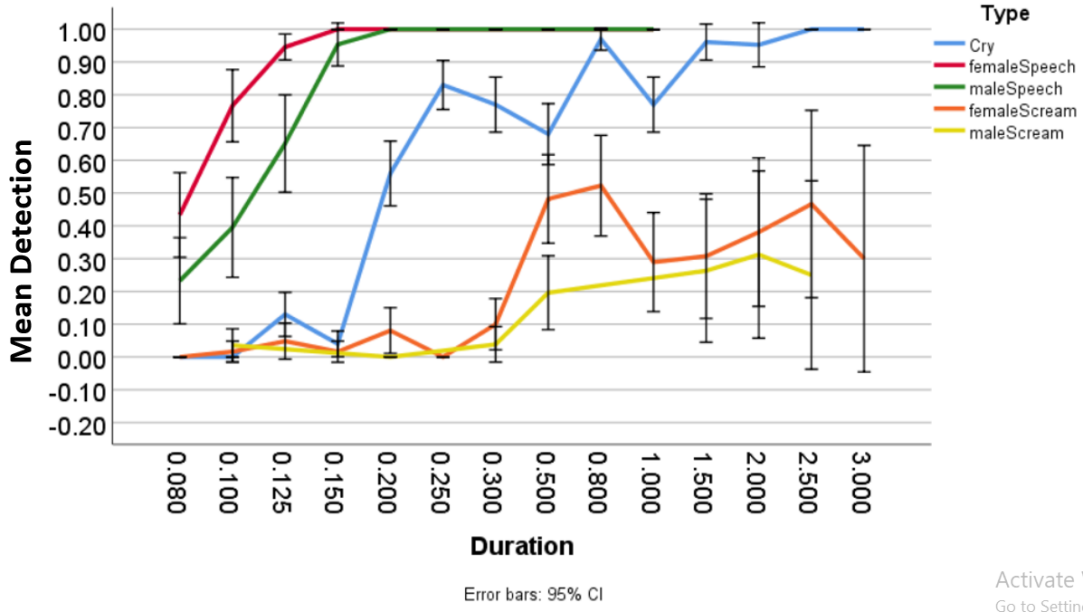


Figure 13: Mean detection of sound classes for different window sizes.

Figure 13 shows a more thorough examination of the detection with respect to audio window sizes in sound classes. It shows that "Female Speech" class gives the best detection rate with short window sizes of audio and "Male Speech" class has similar results. Both classes indicate that a window size of 0.25 s can be suitable to detect the speech classes correctly. The values reported for "Baby Cry" class increases at 0.25 - 0.30 s. The threshold for this class can be chosen at 0.30 s for achieving high performance with minimum window size.

Again error bars showing the 95% confidence interval were shown on plots. Since there are fewer samples with larger window sizes, the error bars are longer. For the "Female Scream" and "Male Scream" classes, the window sizes should be larger as

they are more difficult to classify. The minimum window sizes for these classes were chosen to be 1 s as their error bars are longer after this threshold. This means that the values are more spread out and less reliable. Furthermore, "Male Scream" could not be classified correctly for any of the audio clips with 0.25 and 0.8 s window sizes. Table 5 shows the chosen minimum window sizes for each class after this analysis.

Table 5: Minimum window sizes for target classes.

Class	Min. window size (s)
Female Speech	0.25
Male Speech	0.25
Baby Cry	0.30
Female Scream	1
Male Scream	1

For the target classes, the results depict whether they were correctly detected or not. However, if they were not classified as the target classes, they arewere classified as another class in the YAMNet class map. The classes which were confused with the target classes were discussed in 4.3.1.3.

4.3.1.2 True Positives and Scores

The YAMNet model gives 521 predictions (for each class in the YAMNet class map) and the corresponding scores for each class. The highest scored class is the YAMNet model's first prediction for the audio file. If the highest scored class is correct for the audio file, it is correctly classified. In this section, the second highest scored classes were also analyzed.

In the first analysis, the correct and incorrect predictions of the model were analyzed together without class separation. This allows seeing whether the predictions of the YAMNet model were correct or not. The results were given with respect to scores. Thus, if a prediction is correct, the mean scores were shown as the dashed line in Figure 14 for different window sizes. The straight line shows the scores the predictions that were not correctly classified. The mean scores of the correct classifications are clearly higher than the falsely classified ones, as can be observed. This means that, when the result is correct, the model gives higher scores.

When the highest scored prediction of the model is not correct, the second predictions were analyzed. In Figure 15, only the second highest scored predictions of the model were shown with their mean scores. This analysis combines all classes together in order to understand the model on a larger scale. The dashed line shows when the second prediction of the model is correct, the mean scores of the second prediction is close to 0.50 for most of the window sizes. It could be said that second predictions of the model can also be used for detection of the target sound events. However the straight line shows that when the second predictions are not correct, they still have

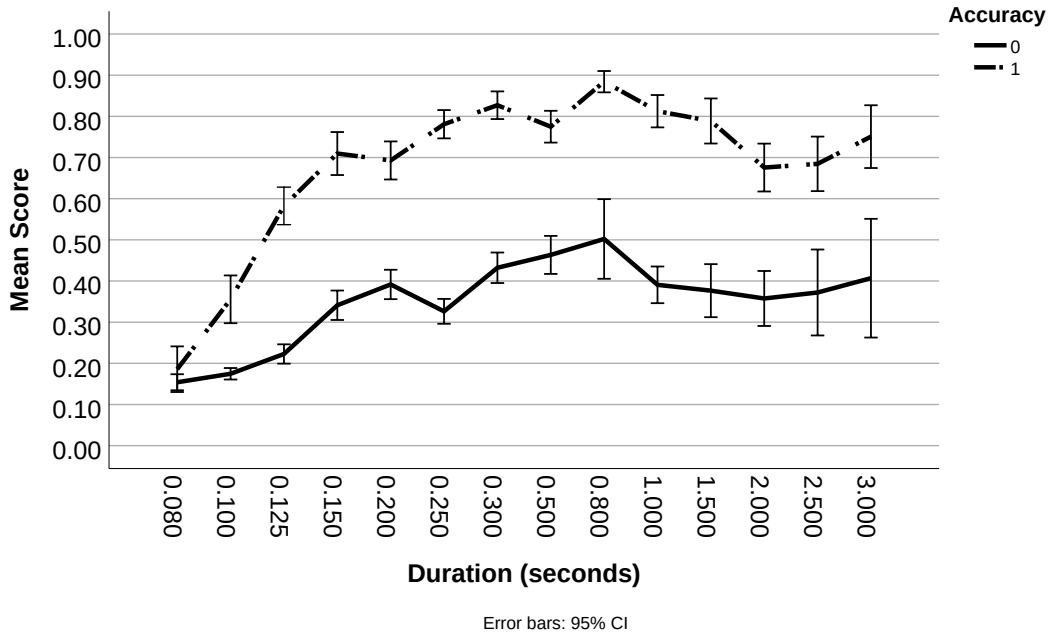


Figure 14: Mean scores of results for different window sizes.

high scores, meaning that the score itself does not provide a valuable information about the usage of the second predictions.

The class-based analysis was used to figure out how certain the model's predictions are when they are correct. For this problem, predictions with their corresponding scores are used. The score for a class shows how certain the YAMNet model is of the corresponding prediction. The results were given for the three main classes ("Speech", "Scream" and "Baby Cry"). Figure 16 shows that mean score of the "Speech" class shows an increase at early stages of the experiment. After 0.25 s, it reaches the mean score of 90 percent. In "Baby Cry" class, lower scores were reported, which shows that the certainty of the model is not as high as the "Speech" class. The "Scream" class is more difficult to classify than other two classes. Even when they were correctly classified by the model, their corresponding scores were lower than the other classes.

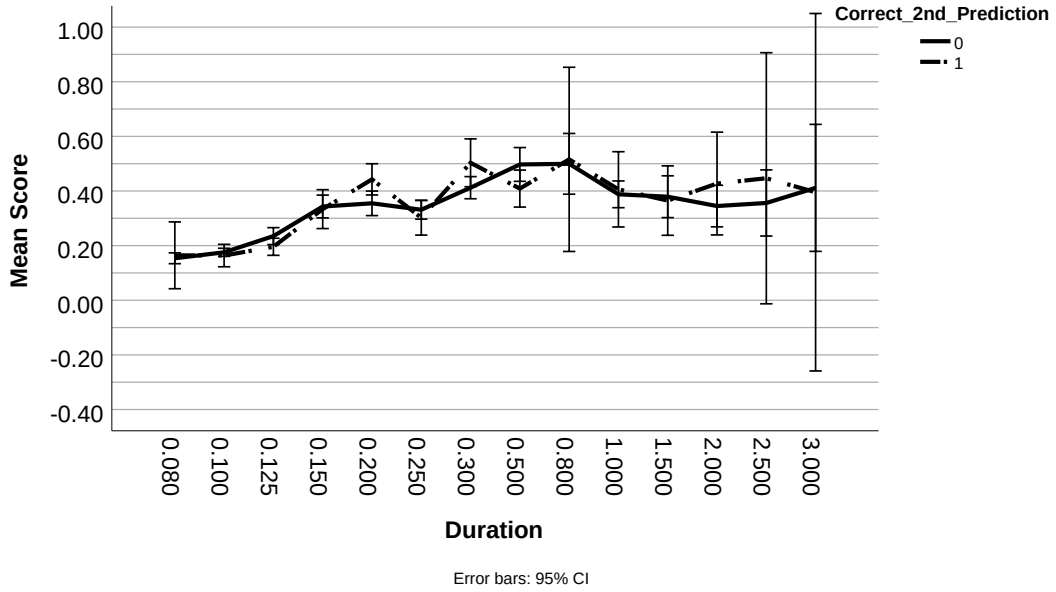


Figure 15: Mean scores of correct second predictions for different window sizes.

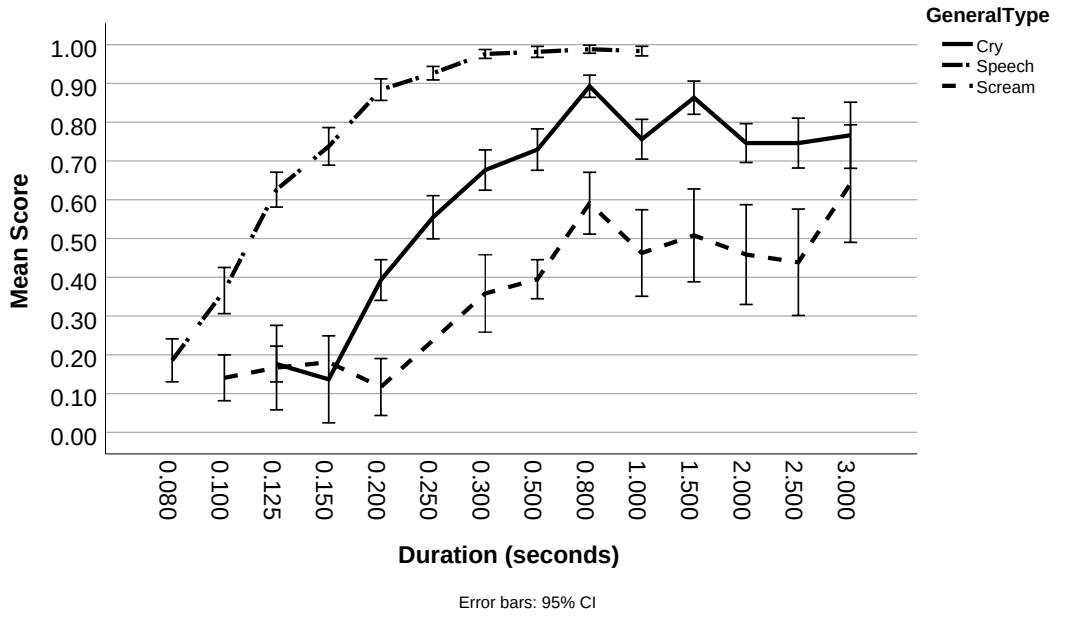


Figure 16: Mean prediction scores of predictions with audio clips of different window sizes.

4.3.1.3 Misclassified Classes

The statistical analysis of the misclasiffied audio clips was performed in this section. For the three main classes, all of the audio files of all window sizes were given to the model as input and the highest scored predictions were reported. Figure 17 shows the audio classes the "Speech" class was incorrectly classified as. According to the SPSS analysis, "Speech" was predicted as "Silence" when it is not predicted correctly. This can also be because of the gaps between the words. Although the dataset is strongly labeled, speech can contain breathing sounds and small pauses.

Figure 18 shows the audio classes the "Baby Cry" class was incorrectly classified as. This analysis does not contain a dominant class, it was mostly classified as "Sheep" and other animal classes, and human sounds like "Gasp", "Laughter". It must be noted that the sound clips of "Baby Cry" class contain intermittent crying sounds with gasps and breathing in between.

Figure 19 shows the audio classes the "Scream" class was incorrectly classified as. It was mostly classified as "Crying, Sobbing" which is close to the target class for some cases. The other classes it was wrongly classified as mostly consist of animal sounds and alarm sounds. When exposed for a short time, it can be difficult to discriminate the scream and animal sounds or alarms even for the human ear. This can also explain the performance of the classification of screaming sounds.

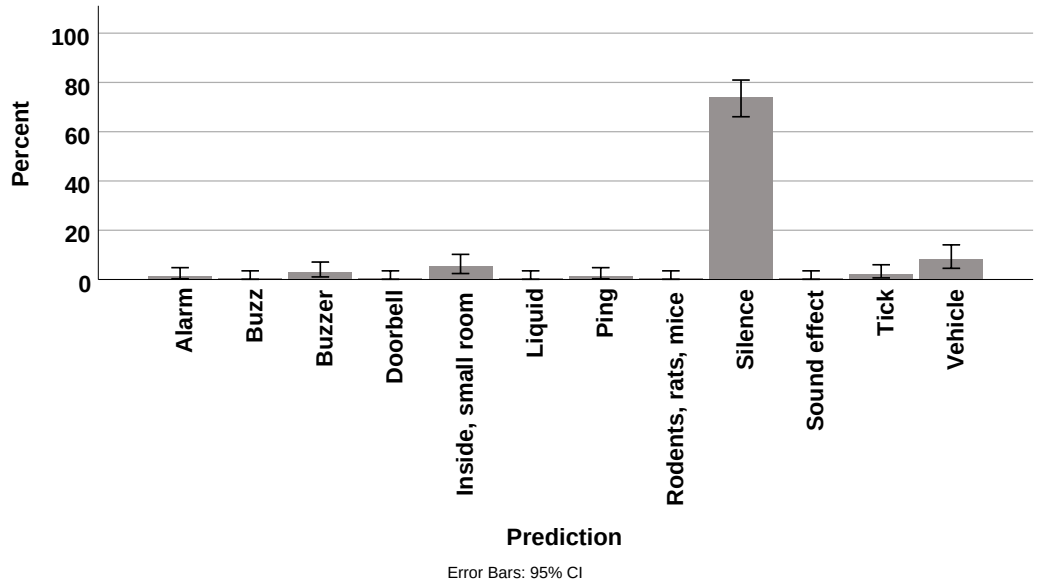


Figure 17: The audio classes which the "Speech" class is misclassified as.

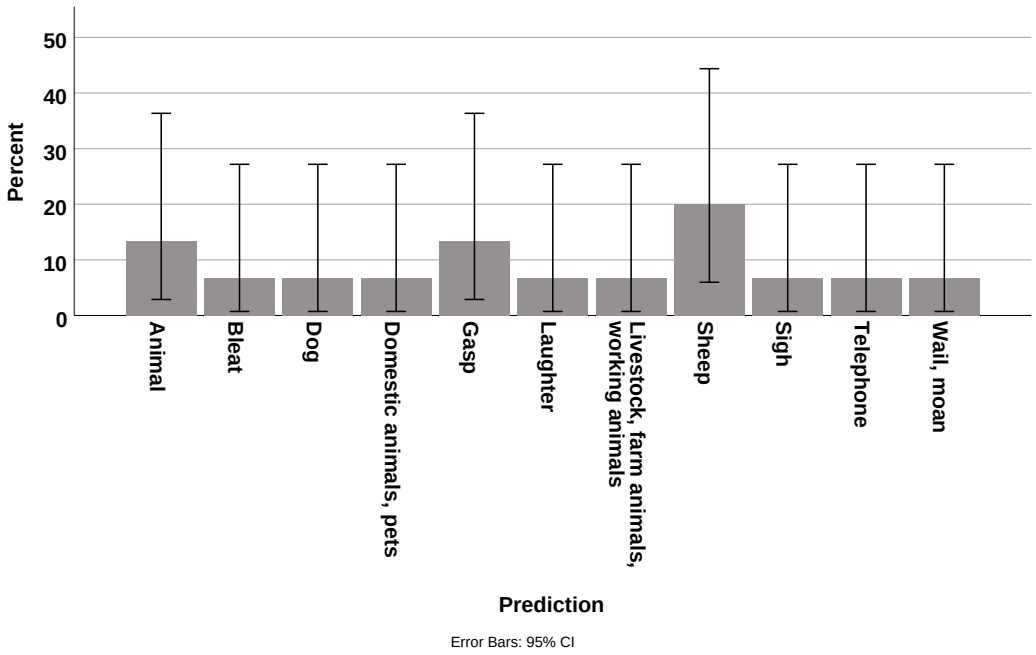


Figure 18: The audio classes which the "Baby Cry" class is misclassified as.

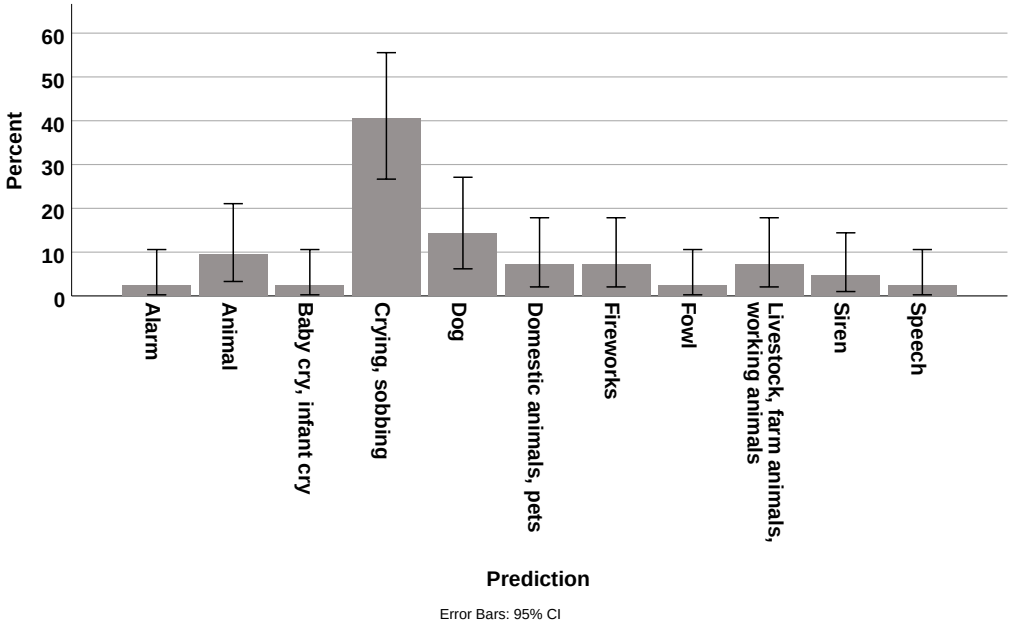


Figure 19: The audio classes which the "Scream" class is misclassified as.

4.3.2 Target Human Voice Detection Using Audio Data

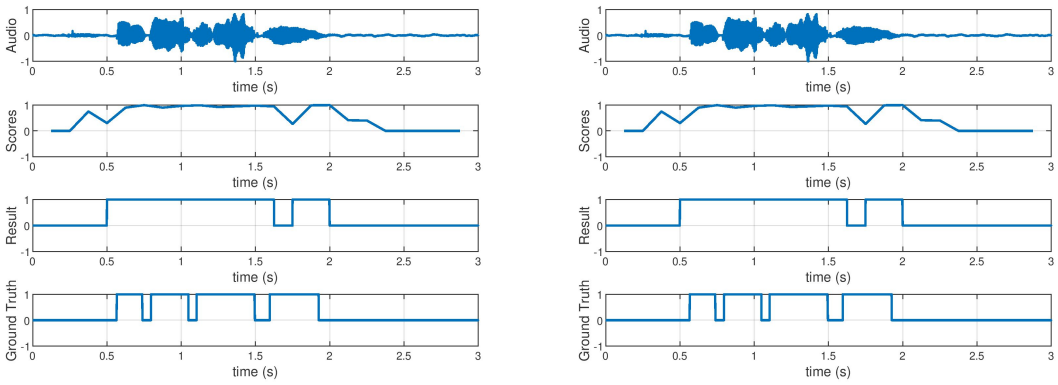
The results of the target audio class detection, the pre-processed audio clips were classified using the YAMNet model. If the audio clips of the target sound classes “Female Speech”, “Male Speech”, “Female Scream”, “Male Scream” and “Baby Cry” were classified correctly as their corresponding target classes given in AudioSet ontology as in Table 3, they were accepted as correct predictions. The window sizes shown in Table 5 were used for the classification.

Table 6 shows the recall values of all classes for target audio class detection of human sounds. According to the table, "Male Speech" performed the highest recall. This may be due to the audio clips containing clear male voices and being very similar to each other. However, speech classes both show high performance. For the "Baby Cry", "Female Scream" and "Male Scream" classes, recall values are significantly lower than the speech classes, which means the scores of the correct predictions are also lower.

Table 6: Recall values of target classes for target audio class detection.

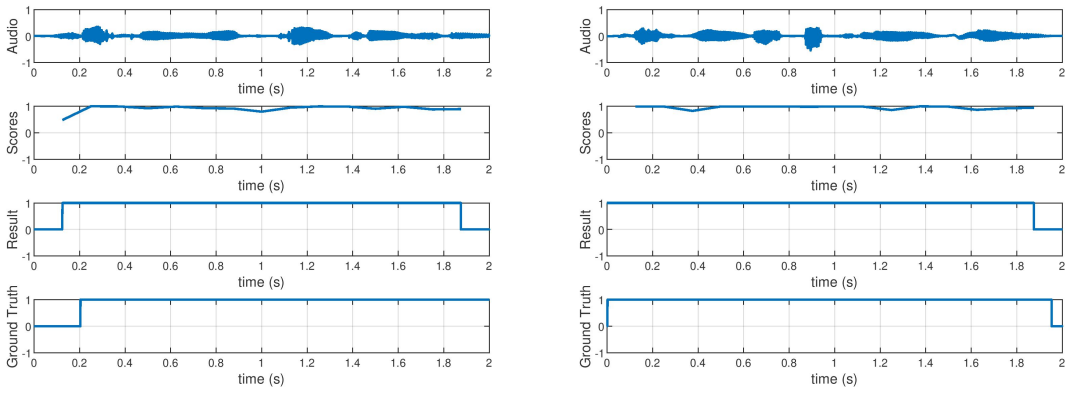
Class	Recall for target detection
Female Speech	0.9031
Male Speech	0.9139
Baby Cry	0.6499
Female Scream	0.7362
Male Scream	0.4784

In addition to the results, the plots of the audio signals, scores, results and ground truth were also given for comparison and visual analysis. The scores on each point in the plots represent the middle point of the sliding window. Figure 20, 21, 22, 23 and 24 show example two plots of the Audio signals, scores of the predictions of the model, results and ground truth for "Female Speech", "Male Speech", "Baby Cry", "Female Scream" and "Male Scream" respectively. Plots obtained from the entire human voice data mentioned in Table 1 can be found in Appendix A.1.



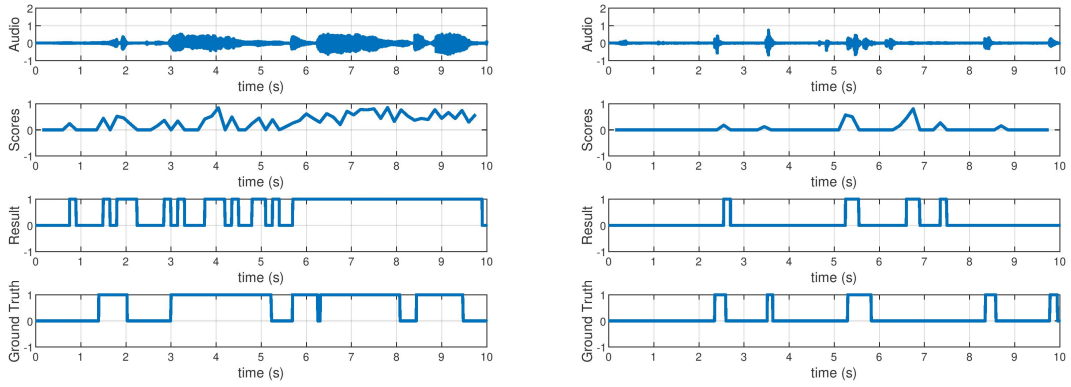
(a) (b)

Figure 20: Audio signal, scores of the predictions of the model, results and ground truth for an example "Female Speech" class for target audio class detection.



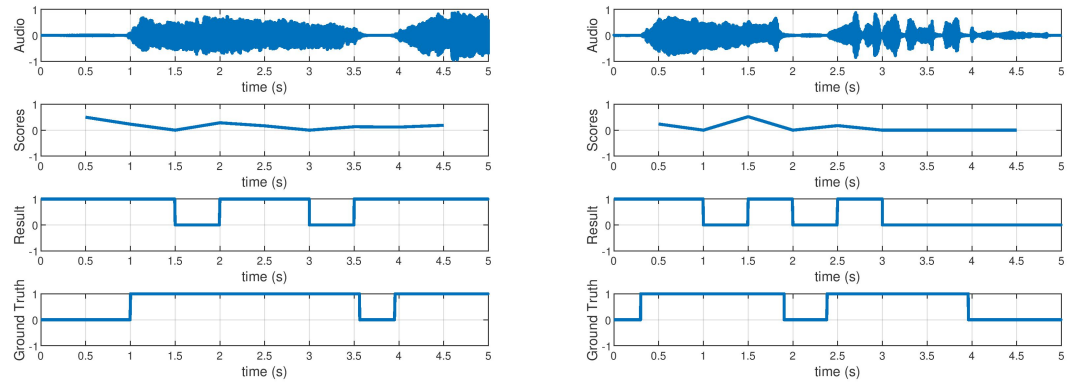
(a) (b)

Figure 21: Audio signal, scores of the predictions of the model, results and ground truth for an example "Male Speech" class for target audio class detection.



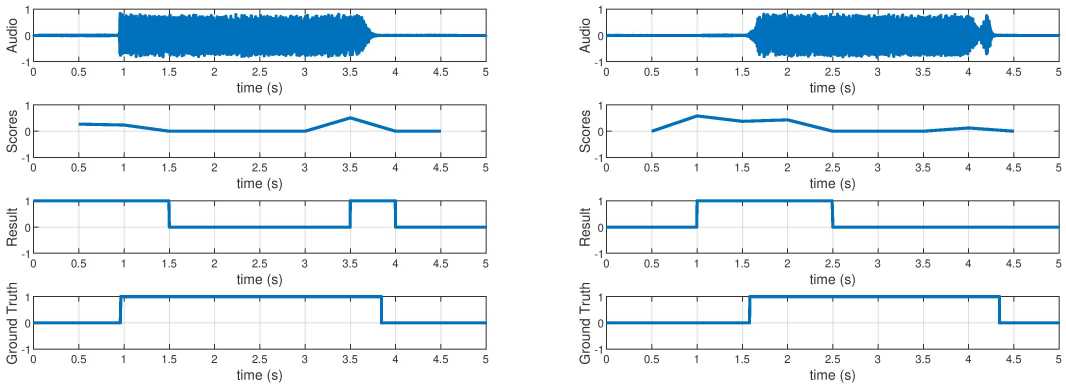
(a) (b)

Figure 22: Audio signal, scores of the predictions of the model, results and ground truth for an example "Baby Cry" class for target audio class detection.



(a) (b)

Figure 23: Audio signal, scores of the predictions of the model, results and ground truth for an example "Female Scream" class for target audio class detection.



(a) (b)

Figure 24: Audio signal, scores of the predictions of the model, results and ground truth for an example "Male Scream" class for target audio class detection.

4.3.3 Target Audio Class Detection in Noisy Environments

Using the YAMNet model, the pre-processed audio samples of environmental sound classes which can be referred as noise and human speech classes were separated. The "alarm," "crash," "barking dog," "running engine," "footsteps," and "ringing phone" classes from the NIGENS general sound events database were utilized to represent environmental noise. As the minimum window size that may be selected for categorizing all human voice classes, audio samples of 1 second were taken for each class.

Figure 25 shows the confusion matrix of the classification of 1 s audio clips of all classes. "Other" class represents the classes in Google AudioSet other than the 9 classes which we chose to investigate. According to the confusion matrix, best performing classes were "Speech", "Cry", "Alarm", "Dog", "Engine" and "Phone". The "Scream" class was mostly confused with crying, alarm and crash sounds. Table 7 shows the precision, recall and F-1 scores for the classes.

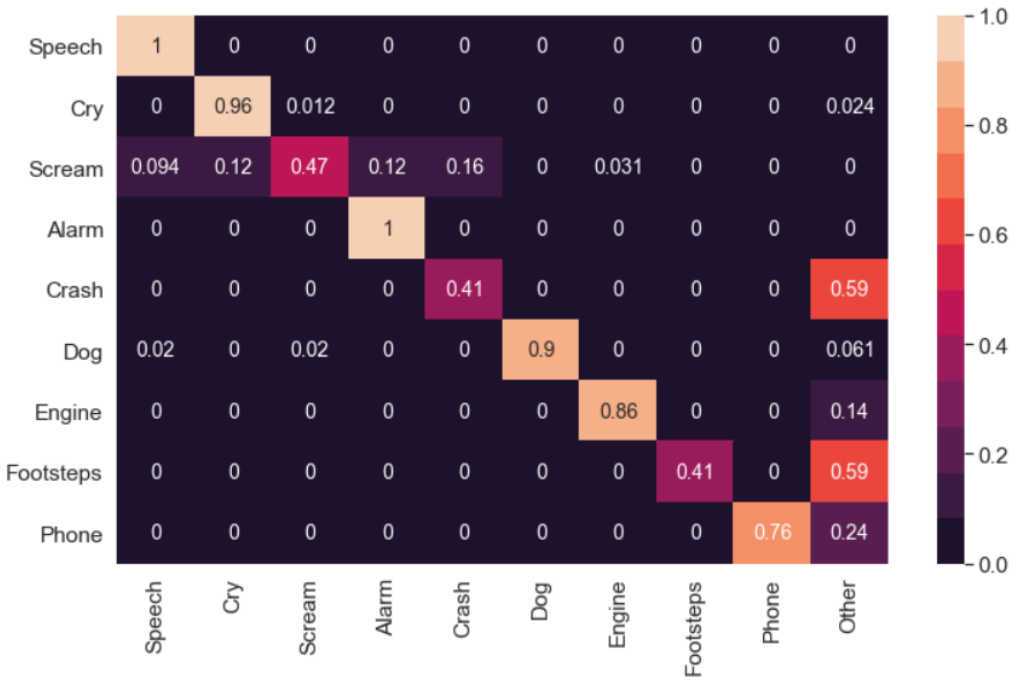


Figure 25: Confusion matrix of human voice and environmental sound classes.

The low performance of the scream class was expected since it was more difficult to detect in the previous experiments. Since even for the human ear, it can be challenging to distinguish between screaming, crying or alarm sounds when exposed for a very short period of time. "Crash" sounds were mostly confused with the "Other" class. The reason behind this was investigated through the classes it was confused with. The classes it was mostly confused were "Vehicle", "Rattle", "Glass" and "Music". Since again it is difficult to classify these sounds with human ear when listened for a short time, the low performance was understandable. For the "Footsteps" class, which also showed low performance, a similar analysis was conducted. The classes it was

Table 7: Sound classes and their corresponding precision, recall and F-1 scores.

Class	Precision	Recall	F-1 Score
Speech	0.92	1	0.96
Cry	0.91	0.96	0.93
Scream	0.94	0.47	0.63
Alarm	0.83	1	0.91
Crash	0.52	0.41	0.46
Dog	1	0.90	0.95
Engine	0.95	0.86	0.90
Footsteps	1	0.41	0.58
Phone	1	0.76	0.86

mostly confused were "Crack", "Knock" and "Crumpling, crinkling" which are very similar when listened for a brief second. A solution for this can be choosing more classes as classes that are accepted correct for a classification. The similar sounds "Crack", "Knock" and "Crumpling, crinkling" could be chosen as correct for a system to be used in some use cases. However, for this study, only the classes mentioned in AudioSet ontology were used.

CHAPTER 5

VISUAL ANALYSIS

Recent advancements in the image analysis and detection fields have enabled the development of practical video surveillance systems with built-in facial analysis functions that are accurate and useful for emergency situations [102]. Since manual surveillance appears to be inconvenient and time-consuming, it is important to utilize automatic detection and analysis methods. Surveillance systems used for security applications can be characterized in various ways depending on the scenario, such as detecting theft, detecting violence, detecting a person needing help, and so on [14].

In line with the audio-based analysis, the aim of the visual analysis section of this thesis is to detect humans and analyze their activity in cases of emergency. In this study, we attempted to detect emergency scenarios by analyzing facial expressions. A thorough investigation of how a deep learning-based visual system can be used in a real-time audio-visual system was conducted. For this purpose, face detection and a detailed examination of the faces were performed. Through facial analysis, a lot of information can be discovered such as age, gender or emotional state of a person. On a face detection database, a face detection model was developed and trained. Facial alignment was performed, which is required for a thorough facial analysis. Facial expression analysis was carried out to detect seven basic human emotions. For the analysis, a facial expression classification model was utilized and trained with a facial expression dataset and used on top of the face detection model. Experiments for the performance of the face detection and facial expression models were performed. Furthermore, the facial expression dataset was auto-labeled for age and gender and trained with facial expressions for a more detailed analysis. A demonstration of the system on videos and webcam was implemented to show the system in real-time applications.

5.1 Proposed Methods

In the visual identification of emergency circumstances, the analysis on human activities plays a significant role. By analyzing facial expressions, we aim to detect people in emergency circumstances. Image processing with deep learning applications are used widely for this purpose. In this section, the proposed methods on the visual analysis and the results of the experimental methods are discussed.

For the application of the system in real-time, firstly, faces need to be detected. Deeper analysis of the facial attributes were performed after this process. Face detection is an object detection problem, thus, an object detection architecture which is proved to be suitable for face detection was employed. WIDER Face [7] face detection database with high variability was chosen for the training and testing of the system. The face detector model was trained with facial landmarks provided by [6] as well as the face bounding boxes. Experiments with different context modules from RetinaFace [6] and SSH [5] which incorporate contextual information to the face detection problem were conducted. Alignment of the face in a face bounding box was performed by using the facial landmarks before further analysis of the faces. Following facial alignment, a multi attention based model called [8] was used to classify the seven fundamental facial expressions: "happy," "sad," "anger," "fear," "surprise," "disgusted," and "neutral." Since a person's facial expression can provide information about their mental state, it can be utilized to identify people in emergency situations. Finally, AffectNet facial expression recognition dataset [93] was automatically labeled for age and gender using pretrained ShuffleNet model [103] to collect and utilise additional information about a person. The AffectNet dataset was trained for facial expression, age and genders. A video and webcam presentation of the system was used to demonstrate the technology in action.

Figure 26 shows the block diagram of the visual analysis method. The method consists of a face detection model which also uses a ResNet-18 backbone, Feature Pyramid Network (FPN) and the RetinaNet object detection model. The facial expression classification method also uses ResNet-18 backbone and Attention Mechanism which is made up of two parts: a channel attention unit and a spatial attention unit before an attention fusion network. The Affinity and Partition loss functions as well as the model architecture were explained in detail in this section.

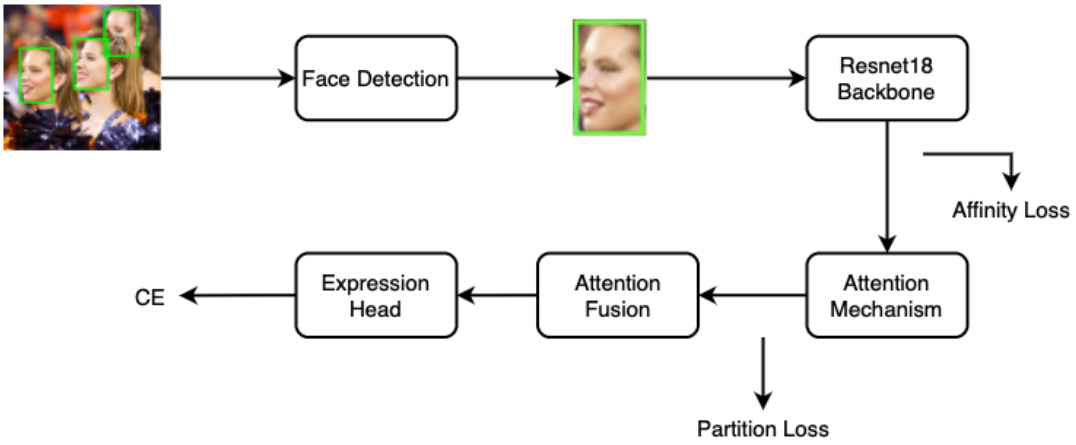


Figure 26: Block diagram of the proposed facial analysis method.

For the face detection model, facial alignment and facial expression analysis, all experiments were performed using Python language and PyTorch library.

5.1.1 Face Detection

With the use of modern CNN-based object detectors, great advances have been made in face detection. An anchor-based detection framework that use anchors to detect hard faces in an uncontrolled environment was utilized for face detection. For the detection, the model was implemented employing backbone, neck and head architectures.

An object detection system from scratch was implemented instead of employing a face detection method such as RetinaFace which focuses on hard faces. By implementing our own object detection method for face detection, we were able to control the specifics of the system and try different context modules. ResNet-18 was employed as the model's backbone. Feature Pyramid Network (FPN) [3] was used as the neck and RetinaNet [1] was used as the head of the network. The base Resnet-18 had been pre-trained on ImageNet database [52] on more than a million images. The WIDER Face dataset and five facial landmarks provided by RetinaFace [6] was used to train the model. It was trained for 80 epochs. Furthermore, experiments with different context modules were performed. In this section, the specifics of the model are discussed.

RetinaNet is a one-stage object detector based on anchors that works well with dense and small-scale objects. Since detection of small faces is an important task in face detection, RetinaNet was chosen for the task. In Figure 27 the architecture of RetinaNet model is shown.

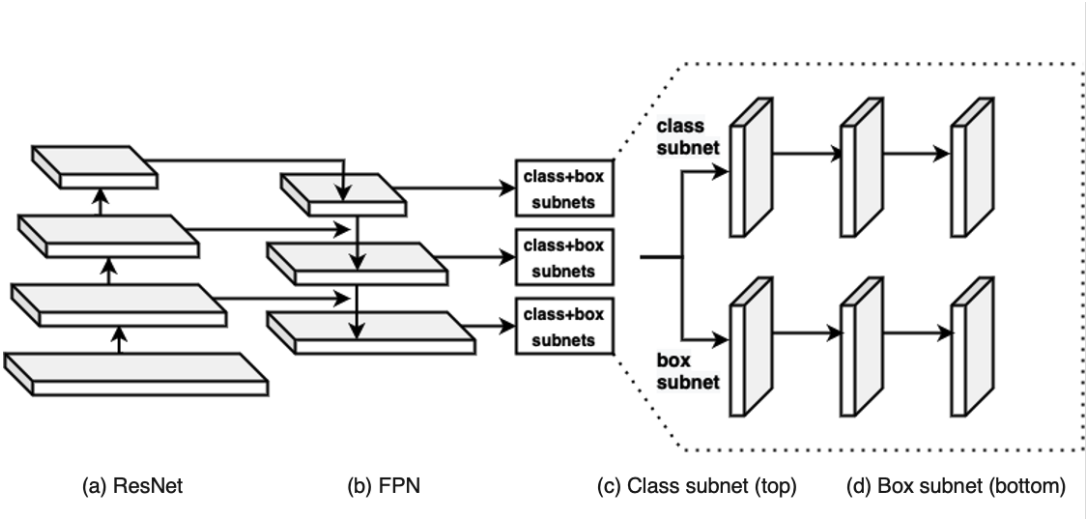


Figure 27: Architecture of our face detector built on RetinaNet [1]. (a) Backbone: Resnet-18 architecture [2]. (b) Neck: a Feature Pyramid Network [3] on top of a Resnet-18 architecture. Two sub-networks are shown after that. One is for the classification of anchor boxes (c), The first is for anchor boxes to be regressed to ground-truth object boxes. (d). For classification, Focal loss is utilized, and for regression, IoU loss [4] is employed.

The focal loss is a new loss function proposed by RetinaNet to address the problem of class imbalance. In the small face identification problem, the high class imbalance between background and foreground may overwhelm the cross entropy (CE) loss. By adding a modifying component to cross entropy loss, focal loss focuses training on hard negatives. It is defined as:

$$FL(p_t) = \ -(1-p_t)^{\gamma} log(p_t). \tag{4}$$

Focal loss uses $(1-p_t)^{\gamma}$ with the CE. When the focus parameter is $\gamma > 0$, for correctly categorized samples, the loss ($p_t > 0.5$) decreases and the model focuses on hard examples.

FPN is employed for the challenging problem of object detection at various scales, especially small objects. Detecting small object is an important topic for face detection as well. Thus, FPN was used in the face detection task in this thesis. A pyramid with levels P3 through P7 was constructed with all pyramid levels having 256 channels as in [3].

RetinaNet has two sub-networks which are classification and regression branches. For each of the anchors, the classification branch predicts the likelihood of object occurrence at each location. Regression branch consists of a fully convolutional network (FCN) [104] for each pyramid level. It calculates the groundtruth box's offset from the anchor. For each anchor box's regression to a ground truth box, again FCN is linked to each pyramid level [1].

5.1.1.1 Context Modules

It's crucial to be scale invariant while recognizing faces in uncontrolled environments. Object detection algorithms use feature maps from prior convolutional layers to recognize small objects. Enlarging the window surrounding the areas is a typical way to include context. Face detectors used different methods to include context to find small faces. In this section, some of the context modules are implemented and incorporated to the face detection model to observe its effect on the detection performance.

SSH [5] uses simple convolutional layers to replicate this method. Raising the window size around the areas by applying a larger filter is analogous to making the window size around proposals larger. It uses 3x3 convolutional layers. Figure 28 shows the architecture of SSH context module that is implemented for this section.

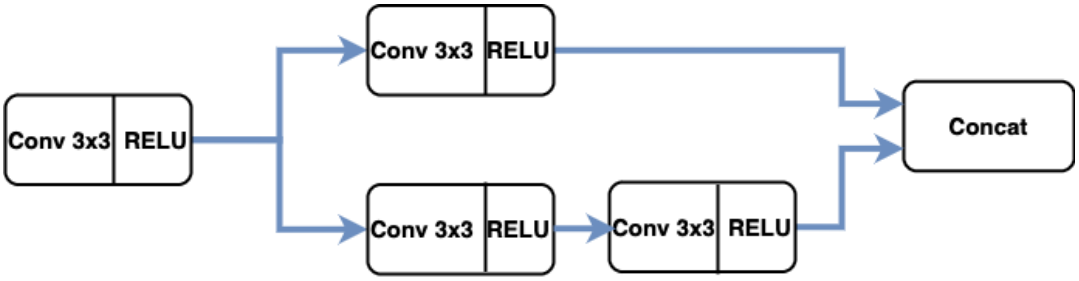


Figure 28: SSH [5] Context Module.

RetinaFace [6] uses seperate context modules for the five feature pyramid levels. The architecture of the context module is seen in Figure 29. It is also implemented to observe the effect on the face detection performance.

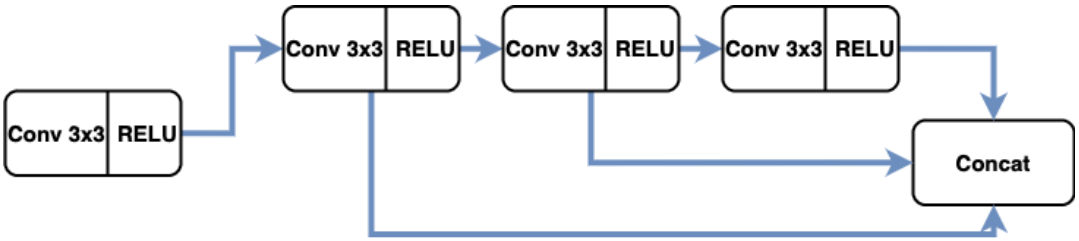


Figure 29: RetinaFace [6] Context Module.

5.1.1.2 WIDER Face Dataset

The model is trained with the WIDER Face [7] dataset. The WIDER Face dataset, which is frequently mentioned in the aforementioned publications, facilitates the understanding of image-based face detection technologies. WIDER Face is a Face Detection Benchmark developed by the Multimedia Laboratory, Department of Information Engineering, The Chinese University of Hong Kong [7]. The WIDER Face

dataset is a database created for comparison applications with images adopted from the generally publicly available WIDER dataset. The database is organized according to 61 event classes. EdgeBox [105] detected three difficulty levels of the dataset (Easy, Medium and Hard). Created by the Multimedia Laboratory of the Department of Informatics Engineering at the University of Hong Kong, the collection contains 32,203 photos and 393,703 faces, with a wide range of scale, attitude, and occlusion as demonstrated in the sample images. Training, validation, and test sets for each event class are composed random selections. Figure 30 shows example images with variations.

Figure 30: Example images from the WIDER Face database [7] which shows the variability in images.

The dataset is evaluated using the same evaluation parameters as the PASCAL VOC dataset [106]. Average Precision AP is calculated for different IoU scores.

5.1.1.3 Facial Landmarks

Significant improvements in performance have been noticed with the addition of extra supervision when WIDER Face dataset is combined with the five facial landmarks which RetinaFace [6] manually labeled. The landmarks are only added to the "annotatable" faces that have higher resolution. There are 84.6k faces on the training set of the dataset with facial landmarks and 18.5k faces in the test set that is also utilized in this study. The landmark locations are two for the center of the eyes, one for the tip of the nose and two landmarks for the two corners of the mouth. Some examples of landmarks in WIDER Face dataset can be seen in Figure 31. Our face detection system includes a face landmark regression loss with the usage of five facial landmarks.

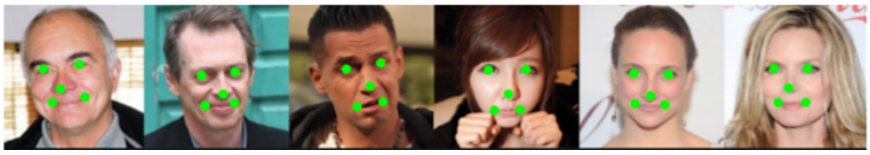


Figure 31: Example images from the WIDER Face database [7] which shows the location of the five facial landmarks.

5.1.2 Facial Alignment

Face alignment can have a substantial impact on facial analysis performance. Scaling variations, rotation angle, or curvature of the face in images captured by the camera, which vary with the distance of the face from the camera all affect the performance of the facial analysis. Thus, facial alignment plays a huge role in the facial expression analysis which is the purpose of this study. For the purpose of facial alignment, all face data should be centered. The data should then be rotated so that the eyes are aligned with a horizontal line. Then the data should be adjusted so that the faces are nearly comparable in size.

OpenCV and Python were used for the alignment. The facial landmarks provided by RetinaFace on WIDER Face dataset were used. A line between the center points of two eyes was drawn. A horizontal line was then used to calculate the angle between the two eyes. After that, the image was rotated to match the angle.

5.1.3 Facial Expression Analysis

Facial expression recognition and analysis play an important role in comprehending people's situations. In surveillance systems, it can be useful to detect people in emergency situations by giving information on the emotional state of a person. The modern facial expression classification methods focus on attention networks, thus, an attention based model was utilized in this thesis.

The Distract Your Attention (DAN) model was used to analyze facial expressions. For six epochs, AffectNet dataset with the seven basic emotions "happy", "sad," "anger," "fear," "surprise," "disgusted," and "neutral" was trained. The accuracy of the model was computed. Additionally, the accuracy of face expressions was obtained by a confusion matrix. A demo of facial expression classification was also performed. The facial expression class was displayed for various thresholds when the demo was executed. In addition, demo was performed both with webcam and videos. Also, the scores of the expressions were set to thresholds. Thus, predictions were not shown for a frame if the score was less than the threshold. For each class of facial emotion, the thresholds can be different. Experiments with different thresholds were undertaken. For each threshold in each class, precision and recall were calculated. Recall refers to how many of the ground truth targets were predicted accurately, where precision indicates how accurate the predictions are.

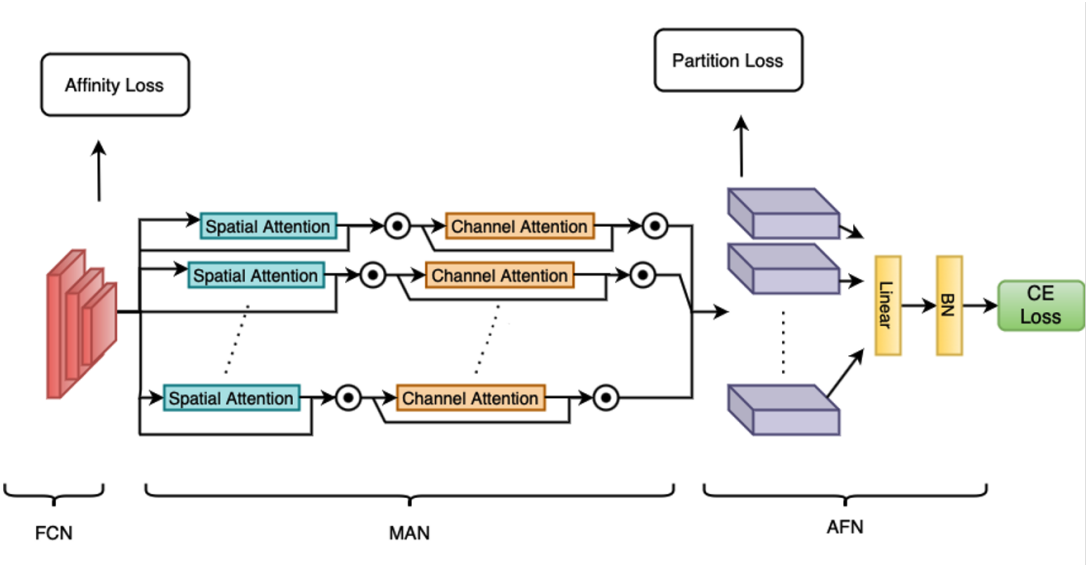


Figure 32: Architecture of DAN [8]. A MAN is created to attend diverse facial regions by a spatial attention unit and a channel attention unit after the basic features are retrieved and clustered by FCN. The attention maps are partitioned by an AFN, which assigns a confidence score.

DAN consists of A Feature Clustering Network (FCN), a Multi-head cross Attention Network (MAN) and an Attention Fusion Network (AFN), which outputs a class confidence. Figure 32 shows the architecture of DAN. The Multi-head Cross Attention Network (MAN) is made up of independent cross attention units, which are a mix of spatial and channel attention units. The spatial attention unit extracts spatial characteristics from the input features provided by the FCN. The spatial attributes are then fed into the channel attention unit, which extracts the channel features. Spatial attention unit consists of 1x1, 1x3 and 3x1 convolutions to capture local elements on a variety of scales. There are two linear layers and one activation function in channel attention unit. Attention Fusion Network (AFN) is used to improve the features MAN has learned. It applies a log-softmax function to the attention maps to find the most intriguing area. Then, to teach the attention heads to focus on various areas and also avoid overlapping attentions, a partition loss is provided. Finally, the attention maps are blended into one. Thus, a class confidence can be computed using a linear layer.

DAN uses two loss functions. Affinity loss increases the inter-class distance while decreasing the intraclass distance. Given class centers c which are sampled from d-dimensional Gaussian distribution, Y has a size of M, d is the dimension of class centers, and the standard deviation between the class centers is σ_c [8], the affinity loss is defined as follows:

$$L_{af} = \frac{\sum_{i=1}^{M} ||x'_i - c_{yi}||_2^2}{\sigma_c^2}. \tag{5}$$

The variance among attention maps is maximized with partition loss. Here, k is the number of cross attention, a parameter to adjust the descent speed of loss values. It is defined as:

$$L_{pt} = \frac{1}{NC} \sum_{i=1}^{N} \sum_{j=1}^{C} log(1 + \frac{k}{\sigma_{ij}^2}), \tag{6}$$

where *C* is the channel size of the attention maps and σ_{ij}^2 is the variance of the *j*-th channel on the *i*-th sample. The loss function of the system is as follows:

$$L = \lambda_1 L_{af} + \lambda_2 L_{pt} + L_{cls}, \tag{7}$$

where λ_1, λ_2 are the weighting hyper-parameters for L_{af}. L_{cls}. λ_1, λ_2 are set to 1.0 in the experiments, as the performance gains observed were not sensitive to their values.

In DAN (Distract Your Attention) [8], an attention module is built that pays attention to numerous face areas at the same time and builds attention maps on these regions. The accurate prediction rate is over 60% in AffectNet dataset with the 7 basic emotions, which are "happy", "sad", "surprise", "fear", "disgust", "anger" and neutral. Since the facial expressions are manifested in different parts of the face simultaneously, the work is suitable for the problem of the thesis.

After the face detection, a bounding box was generated, facial alignment was performed and DAN model was utilized to the aligned face to classify the expression of the face. The important expressions for this study are "fear" and "sadness" as they are the most similar to the "scream" and "cry" classes covered in audio-based analysis in this thesis.

5.1.3.1 AffetcNet Dataset

AffectNet is a data set compiled by Denver University Electrical and Computer Engineering Department. In this set, in which more than a million facial expressions are compiled from sources available on the Internet, a classification is made based on two emotional models of 7 emotions, which can be scanned with 1250 keywords in 6 different languages. The first of these two models attempts to estimate the intensity of valence and arousal parameters with a categorical model using deep neural networks. It states that the deep neural network baselines used give better results than conventional learning methods [93].

The AffectNet dataset, in which more than a million facial expressions are compiled from sources available on the Internet, is based on 7 emotions. AffectNet is integrated with the images in search engines such as Google, Yahoo, Bing, serving in English, Spanish, Portuguese, German, English, Spanish, Portuguese, German, Arabic and Persian languages. Twelve human experts manually annotated 450,000 of

these images in both categorical and dimensional (valence and arousal) models, and labeled images with any facial congestion. AffectNet is by far the largest database of facial effects in still images. The cropped region of face images, facial landmarks, and impact tags are available in the dataset [93]. The number of annotated images for each facial expression category can be seen in Table 8.

Table 8: Number of Annotated Images in Each category.

Expression	Category
Neutral	80,276
Happy	146,198
Sad	29,487
Surprise	16,288
Fear	8.191
Disgust	5,264
Anger	28,130
Contempt	5,135

5.1.4 Age and Gender Analysis

Facial analysis also contains age and gender analysis. Although it is not directly the main subject of the thesis, it was studied to incorporate age and gender analysis to the face analysis model since the audio analysis also includes gender-based labels (e.g. female scream and male scream) and age-based labels (e.g. baby cry). The AffectNet dataset was automatically labeled with a pretrained ShuffleNet model [103]. The DAN network was then trained with both the facial expression, age and gender labels. Then, instead of the attention network in DAN, the model was trained using a fully connected layer for age and gender training. The two results were compared. A final demo was implemented for the system for detecting the face, showing the predicted age group, gender and facial expression of the face along with a bounding box.

5.2 Performance Metrics

Object detection is a well-studied subject in the field of computer vision. The same object detection evaluation methods are used because the face detection problem is a type of object detection. Various approaches have been used for the evaluation of object detection methods. The most commonly used performance metric is AP and its variations [107]. It is used to evaluate detections in most contests such as PASCAL VOC challenge [106]. In the evaluation of object detection, True Positives, False Positives and False Negatives are determined from the predicted boxes and ground truth boxes. This concept is calculated via Intersection over Union method (IoU).

5.2.1 Intersection over Union (IoU)

In order to find out how accurate the bounding box coordinates are, once they are anticipated, Intersection over Union (IoU) metric is utilized. IoU gives the discrepancy between predictions and the ground truth bounding boxes. It is based on Jaccard index [108] which measures the similarity between the data. IoU is calculated by dividing the overlap between the predicted and ground truth bounding boxes B_p and B_{gt}, by the area of their union [107],

$$J(B_p, B_{gt}) = IoU = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}}. \tag{8}$$

A threshold is defined for attaining scores from classifications by computing the IoU score for every detection. For the IoU values greater than the threshold, it is said that the prediction is positive and for the IoU values lower than the threshold are false predictions. TP, FP and FN values are determined. False Negatives (FN) are not applicable for object detection. These numbers are used to calculate precision and recall. They are defined as:

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{AllDetections}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{AllGroundTruths}. \tag{10}$$

Precision is a model's ability to predict the relevant examples. The ability of a model to find relevant cases is known as recall. High recall and low precision means that all ground truth objects have been spotted, but there is a high FP rate. High precision systems with low recall means almost all predictions are correct, but the system has a high FN rate.

5.2.2 Average Precision (AP)

Average precision (AP) is a single numerical metric that incorporates both precision and recall. The concept of the AP evaluation metric is derived from the the PASCAL VOC competition dataset. Over a range of 11 equally spaced recall levels from 0.0 to 1.0, maximum precision values are averaged. Then the precision values are interpolated. This means that we do not use every precision value calculated at each recall, but we use the maximum precision point to the right [106].

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, 0.2, \ldots, 1\}} p_{interp}(r), \tag{11}$$

where

$$p_{interp}(r) = \max_{\widetilde{r}>r} p(\widetilde{r}). \tag{12}$$

Here, $p_{interp}(r)$ is the interpolated precision, which is calculated at each recall level, r, by taking the maximum precision measured for that r. $p(\widetilde{r})$ is the measured precision at $\widetilde{r}$.

PASCAL VOC also defined the mAP metric, which is the average AP over all classes. They defined IoU the threshold as 0.5. However, MS COCO averages the AP for both all classes and different IoU thresholds. The range of the IoU thresholds they use is between 0.5 and 0.95 and the step size is 0.05. mAP is computed over N classes as:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i. \tag{13}$$

In face detection, there is only one class which is face and the rest is the background. Thus, AP over different classes is not applicable. The results are computed for both 0.5 IoU and 0.5-0.95 IoU thresholds.

5.3 Results

The results of the proposed methods on visual analysis are given in this section.

5.3.1 Face Detection

Face detection model was implemented with Resnet-18 backbone, Feature Pyramid Network and RetinaNet architecture. It was trained with WIDER Face dataset and with landmarks provided by RetinaFace. The models were trained with 80 epochs as [6] suggests. The performances were evaluated by the Average Precision (AP) at 0.5 IoU and 0.5:0.95 IoU. The results of the face detection methods were shown in Table 9.

Different modules that incorporate context information were also implemented and added to the model. The context modules use the levels of the FPN in order to use the context information surrounding the face. We have implemented the context modules independently. Thus, there is no weight sharing, as stated in RetinaFace [6]. The model without using any context module was also trained and the results were compared. According to the table, the AP is 0.41 for 0.5:0.95 IoU and 0.77 for 0.5 IoU for RetinaNet in WIDER Face validation dataset. However, the use of context modules has not resulted in any improvements in the AP. Thus, the face detection method that was used before the the facial expression analysis, the usage of the context modules is not necessary. This method of face detection achieves lower performance than face detection specific methods. For example, RetinaFace which scores an AP = 0.91 can be used for the face detection task. We aimed to implement and have the control over the system for trying different context modules. However, a face detection method such as RetinaFace can be used for this purpose.

Table 9: Face detection results for different context modules and number of epochs.

Method	Epoch	AP (IoU = 0.5:0.95)	AP (IoU = 0.5)
RetinaNet	80	0.41072	0.77841
RetinaNet + SSH Context	80	0.40929	0.77754
RetinaNet + RetinaFace Context	80	0.40954	0.77644

5.3.2 Facial Expression Analysis

For the facial expression analysis, Distract Your Attention (DAN) model was employed. AffectNet dataset was trained with the seven basic emotions "happy", "sad", "anger", "fear", "surprise", "disgusted" and "neutral" for 6 epochs. The accuracy of the model on the validation dataset was 64.14%. Furthermore, the accuracy for the facial expressions were computed seperately. The confusion matrix for the seven expressions was shown in Figure 33.

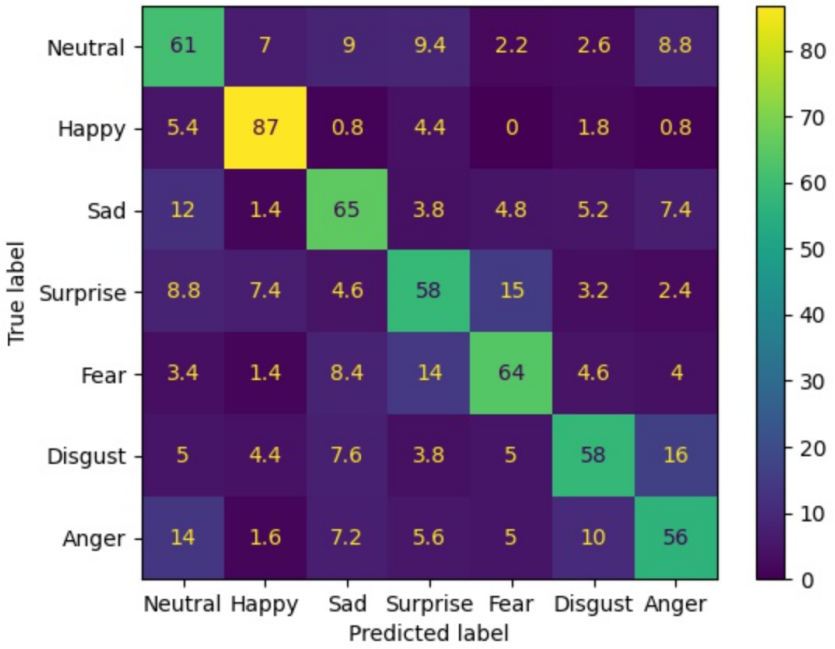


Figure 33: Confusion matrix for the facial expression classification.

According to the confusion matrix, "happy" class achieved the best performance. This can be because of the fact that happiness shows itself with more distinct features on the face and it is more difficult to confuse with the other six emotions. However, "surprise" and "fear" emotions show similarities. Thus, the most confused classes were "surprise" and "fear" classes. Other mostly confused classes are "anger" and "disgust" with accuracies being 56% and 58%, respectively.

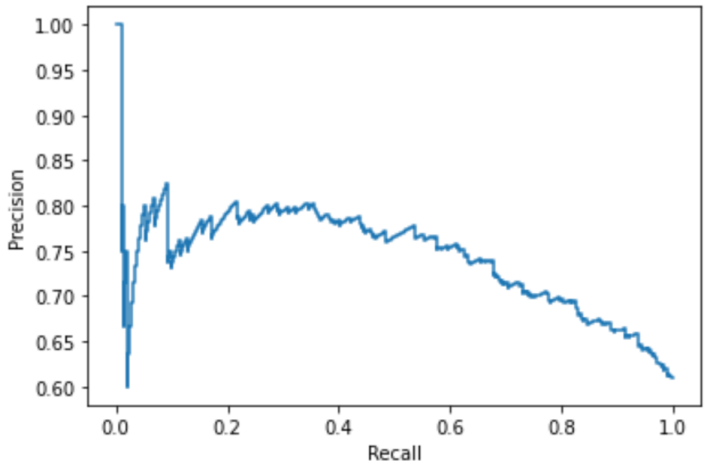


Figure 34: Precision-Recall Curve for the "Neutral" class.

An inference of facial expression classification was also performed. The predicted facial expression class was displayed if the prediction score is greater than a determined threshold when the inference demo was executed. In addition, the demo can be used with videos and webcams. For each class of facial emotion, the optimal thresholds can be different. Thus, experiments with various thresholds were carried out. Precision and recall values were computed for each threshold for every class. Here, recall denotes how many of the ground truth targets was correctly predicted. Precision denotes how many of the predictions were correct. For one class to have higher recall and lower precision, the threshold can be decreased. To make the system more sensitive to a specific expression class, the threshold for that class can be raised, resulting in more precision and reduced recall. For each facial expression class, precision-recall curves were provided.

For the emergency situations and surveillance applications, "fear" and "sad" classes play a more important role than the other facial expressions. Therefore, the system can be updated to make these two facial expressions more detectable by lowering their corresponding thresholds.

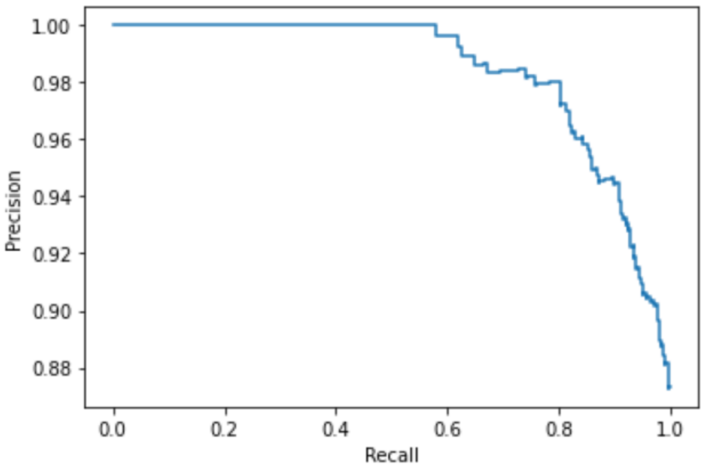


Figure 35: Precision-Recall Curve for the "Happy" class.

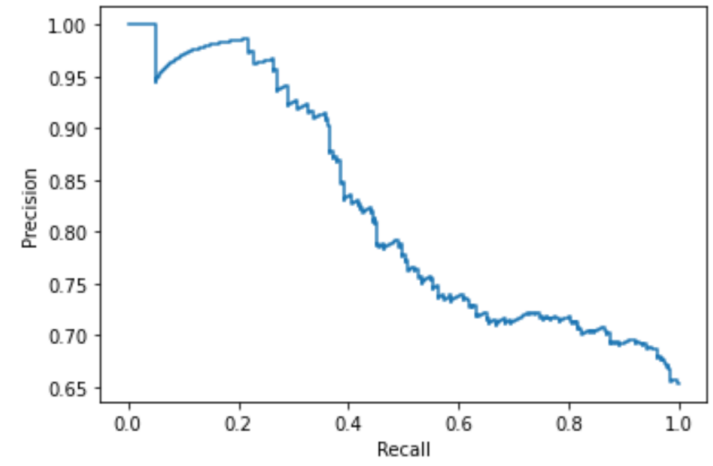


Figure 36: Precision-Recall Curve for the "Sad" class.

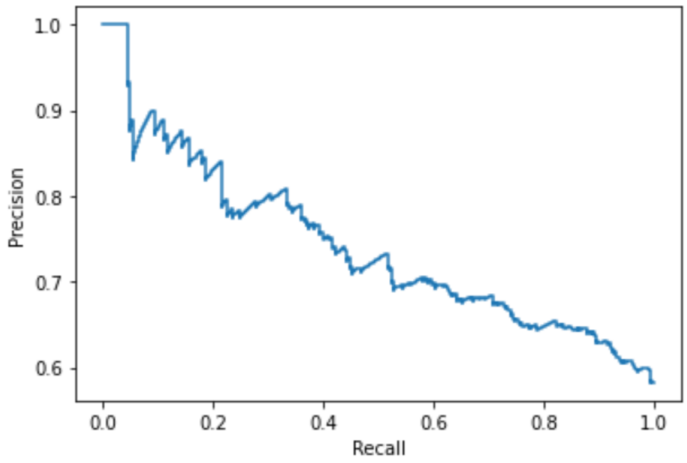


Figure 37: Precision-Recall Curve for the "Surprise" class.

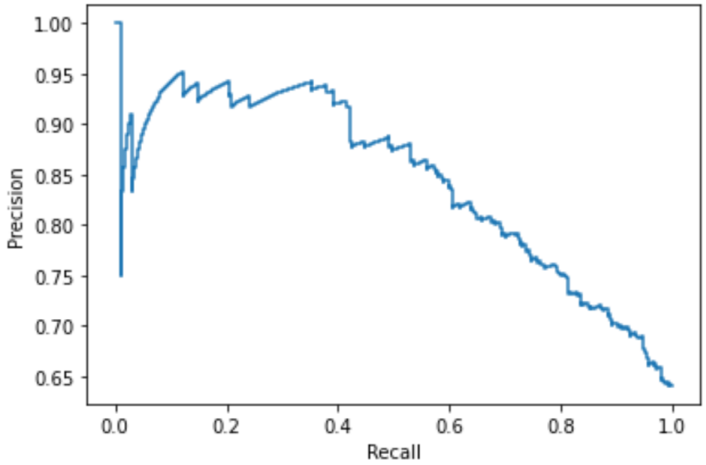


Figure 38: Precision-Recall Curve for the "Fear" class.

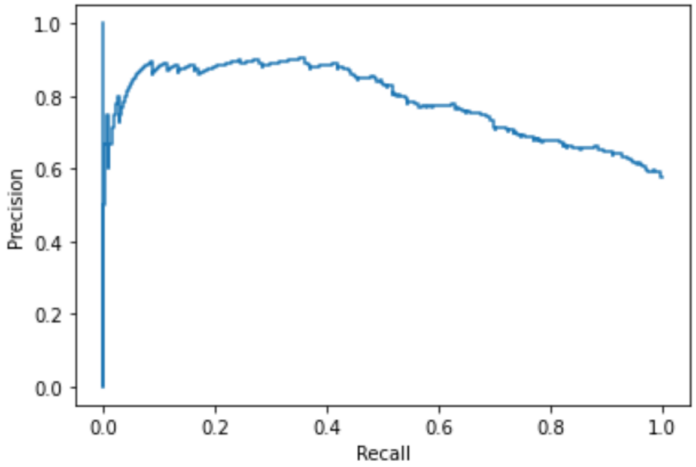


Figure 39: Precision-Recall Curve for the "Disgust" class.

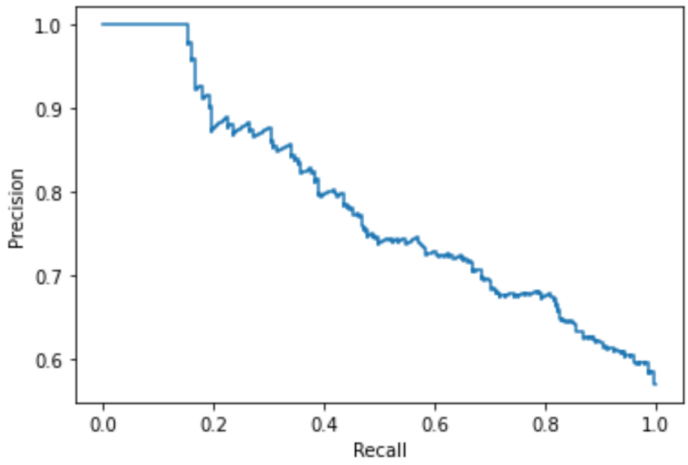


Figure 40: Precision-Recall Curve for the "Anger" class.

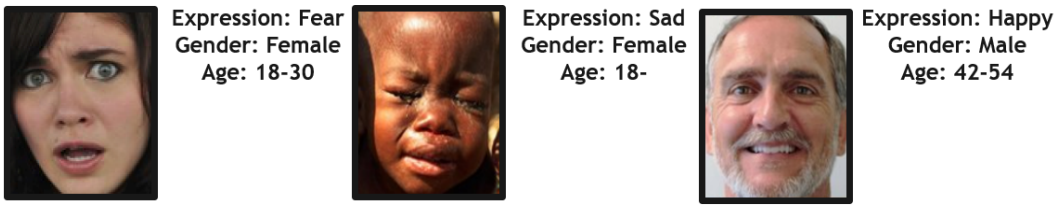


Figure 41: Outcome of visual analysis on examples images.

5.3.3 Age and Gender Analysis

For the age and gender analysis, DAN model was employed using automatically labeled AffectNet facial recognition dataset. The age and gender was trained both with using the attention modules of DAN and without them. The mean absolute error of age and accuracy of gender were compared. The results can be seen in Table 10. Here, since age estimation is a regression problem, Mean Absolute Error (MAE) was used. MAE is the average of all absolute errors. Gender is a classification problem, thus, accuracy was used as the performance metric. It is the fraction of the number of classifications the model correctly predicted to the total number of predictions made.

Table 10: Performance of age and gender training with AffectNet dataset with and without using DAN attention module.

Attention Module	MAE of Age	Best accuracy of gender
+	4.2913	0.9077
-	4.2795	0.9209

Inference with the model was demonstrated in the final demo. Here, ages were also grouped for the demo. The age ranges are 18-, 18-30, 30-42, 42-54 and 54+. Figure 41 depicts the outcome of the visual analysis.

CHAPTER 6

DISCUSSION

In audio-based analysis, the aim is to detect target emergency sound classes when variety of sounds occur simultaneously. For this purpose, not only emergency detection was performed, but also the type of the emergency was determined and analyzed. Audio was converted into samples using the sliding window approach to capture a target audio class's brief window occurrence. As a result, when the target sound momentarily occurs in a real-time operating system, emergency situations can be identified. Five human voice classes from NIGENS sound event database were chosen as "Female Speech", "Male Speech", "Female Scream", "Male Scream" and "Baby Cry". The minimum window sizes for which these classes can be accurately classified were determined. According to the experiments, the minimum window sizes were found to be 0.25 s for "Female Speech" and "Male Speech" classes; 0.30 s for "Baby Cry" and 1 s for "Female Scream" and "Male Scream" classes.

The contributions of the system includes detecting a short window of the occurrence of target audio events. Using sliding window technique with the minimum window sizes, audio data was analyzed in segments. Audio event classification was performed on these segments of audio data. The results and ground truth were converted to samples and recalls for each target audio class were computed. Performance for "Speech" classes was significantly better than the other classes as their recall rates were above 90%. "Male Scream" performance was the lowest with 47% recall, which can be due to having fewer data samples than "Female Scream" class, which performed a 73% recall. In the literature, "Scream" classes are analyzed regardless of gender and no comparison between the classification of them could be found. The training dataset does not contain gender specific classes for screaming data. Upon further inspection, the "Male Scream" data was observed to be very similar to each other and not as diverse as "Female Speech" in the test dataset. The recall for "Baby Cry" was 65% for the chosen window size, which can be seen in Section 4.3.3 increased to 96% with 1 s window size. 0.30 s segments of crying baby sounds can contain pauses and breathing sounds, which can be the reason of the lower performance in short window sizes. Also, crying baby sounds can be mistaken for various sound events like animal sounds in very short windows for human ear as well.

Experiments on the misclassified sound classes were also performed. "Speech" classes were mostly confused with "Silence", which is understandable since speech contains pause and breathing in between the words. Pause, silence or breath can cover the most of a small sliding window segment like 0.25 s in a speech data. "Scream" was

mostly confused with "Crying" and "Baby Cry" was mostly confused with animal sound classes. For these classes, in small window segments, it is also difficult for the human ear to distinguish.

Lastly, environmental sounds were introduced and classification for both human voice classes and environmental sound classes was conducted. Although the aim is to detect human sounds in an emergency, a detailed analysis on human sounds and environmental sounds was performed to also detect the type of that emergency. Here, "Speech", "Cry", "Alarm", "Dog", "Engine" and "Phone" classes performed significantly better than "Scream", "Crash" and "Footsteps" classes which can be seen with the confusion matrix and precision, recall and F-1 scores. "Scream" class was mostly confused with "Speech", "Cry", "Alarm" and "Crash" sounds.

Since the aim is to detect human sounds, "Scream", "Speech" and "Cry" sounds can be used all together in some use cases. Only detecting the type of emergency, such as classifying the emergency as crying or screaming can be more challenging. In some use cases environmental classes such as "Alarm", "Crash" can also be used for emergency detection systems. Furthermore, the system can be improved for specific tasks combining classes together in the future. For example, for search and rescue operations where detecting the type of emergency is not important, "Footsteps", "Speech", "Cry" and "Scream" classes can be combined and used together to detect a person in distress.

Five facial landmarks were employed in visual analysis, together with bounding boxes, to conduct face detection and facial alignment. Two context modules were implemented to improve the performance of the face detection. The AP (IoU=0.5) for face detection for all experiments were found to be 0.77. Context modules did not have a positive effect on detection performance. The context modules were implemented as shown in the PyramidBox [69] and SSH [5] papers. However, the model is different and the channel sizes were changed for our use case. Thus model and context modules can be incompatible to use together for improving the performance in our case. Therefore, for the face detection task, methods such as RetinaFace can also be employed.

Facial expression recognition (FER), age, and gender analysis were done utilizing the discovered faces and an attention-based approach. Seven basic emotions "happy", "sad", "anger", "fear", "surprise", "disgusted" and "neutral" were used. The accuracy of the model on the validation dataset was 64.14%. Furthermore, the accuracy for the facial expressions were computed separately. The confusion matrix for the seven expressions was shown. Highest performance was achieved for "happy" class with 87%. Happiness manifests itself with more recognizable facial traits and is more difficult to be mistaken for the other six emotions. The feelings of "surprise" and "fear," however, are comparable. Therefore, the "surprise" and "fear" classes were the most mistaken. With accuracy rates of 56% and 58%, respectively, "anger" and "disgust" are the other classifications that are frequently mistaken. "Fear" and "sad" classes are more critical than other facial emotion classes in emergency scenarios and surveillance applications. By decreasing their respective thresholds, the system may be modified to make these two facial expressions easier to spot. For the age estimation

MAE was found to be 4.28 and accuracy for gender was 92% in the final model. Ages were grouped for the demo purposes. The age ranges are 18-, 18-30, 30-42, 42-54 and 54+.

Finally, by merging the audio-based and visual elements, a more precise forecast may be created. Since the aim of the thesis is to detect emergencies, "Scream" occupies an important place in the research. As for the visual analysis "fear" and "sad" classes are important. These classes however achieved lower performances. The vision and audio-based systems therefore can be used together for supporting the emergency detection. The system can be developed to be more sensitive to the target classes as well. Moreover, in an emergency detection system, false positives are more acceptable since it is more important not to miss an emergency. In the audio-based analysis, a detection result was provided by the model for every sliding window segment. The detection results from the consecutive sliding windows can be used together for making more robust predictions. For these reasons, a fused system can be used in surveillance applications successfully.

Combining audio and vision based detection systems in various ways for different use cases is possible. To detect emergency in crowds, detection of "Scream" and "Cry" in audio data or detection of "Fear" in visual data using FER can be critical. For anomaly detection in crowds, "Alarm" and "Crash" classes can also mean emergency. For search and rescue operations, detection of one of the human sound classes or detection of the face can be critical. Both the facial expressions, age and gender information as well as the sound information can be helpful to enable robotic systems to comprehend human emotions as a part of the human computer interaction process. For baby monitoring systems, "Cry" and "Scream" classes can be used to notify the parents. Furthermore, the system can warn the parents when the face of the baby cannot be detected to protect the baby from accidents.

Extensive research was conducted about different ways the various tasks were to be employed. Additionally, a demonstration of the operational visual system in real time was put into practice. A real-time audio-visual working system may be used for surveillance applications, as shown by the sample-based audio analysis and the facial analysis.

CHAPTER 7

CONCLUSION

Since visual sensors are routinely installed in public settings, emergency event detection techniques currently used in surveillance systems mostly rely on visual data to identify any abnormal occurrences. However, in a crisis, sound data may be distinguished, assisting the monitoring system in determining whether an abnormality exists. On the other hand, when the audio is noisy and the scene is crowded, applications for visual analysis could be more beneficial. Therefore, the thesis focused on the transition from single-modality to multimodality learning utilizing deep learning techniques. In order to build an audio-visual system, the audio analysis and the visual analysis were both carried out independently.

In audio-based analysis, our contributions include determining the shortest window size that enables the detection of target audio events. Audio was analyzed in segments using the sliding window approach to capture a brief window of the target audio class. Five human sound classes were chosen as "Female Speech", "Male Speech", "Female Scream", "Male Scream" and "Baby Cry". The minimum window sizes for which these classes can be accurately classified were determined as 0.25 s for "Speech" classes, 0.30 s for "Baby Cry" and 1 s for "Scream" classes. For the detection using these window sizes, performance of "Speech" classes were significantly higher than other classes as their recall rates were above 90%. Performance for "Male Scream" class was the lowest with 47% recall rate, which can be due to having fewer data samples than "Female Scream" class in the test dataset, which performed a 73% recall. After adding the environmental sound classes, confusion matrix and precision, recall and F-1 scores were provided. "Speech", "Cry", "Alarm", "Dog", "Engine" and "Phone" classes performed significantly better with recalls 1, 0.96, 1, 0.9, 0.86 and 0.76 respectively. Recalls for "Scream", "Crash" and "Footsteps" classes were 47%, 41%, 41% respectively.

In visual analysis, five facial landmarks and bounding boxes were used to perform face detection and facial alignment. The AP (IoU=0.5) for face detection for all experiments were found to be 0.77. By using an attention-based technique, facial expression recognition (FER) and age and gender analysis were carried out on the detected faces. Seven basic emotions "happy", "sad", "anger", "fear", "surprise", "disgust" and "neutral" were used. The accuracy of the model on the validation dataset was 64.14%. The confusion matrix for the seven expressions was shown. Highest performance was achieved for "happy" class with 87%. The "surprise" and "fear" classes were the most mistaken. With accuracy rates of 56% and 58%, respectively,

"anger" and "disgust" are the other classifications that are frequently mistaken. For the age estimation MAE was found to be 4.28 and accuracy for gender was 92% in the final model.

Finally, more accurate prediction can be constructed by combining the visual factors with audio-based factors. The information fusion of both systems can be performed for specific use cases. A detailed research on how the two modalities were to be used was conducted. Furthermore, a demo showing the working visual system in real time was implemented. The sample-based audio analysis as well as the facial analysis demonstrate the viability of using a real-time audio-visual working system for surveillance applications.

For the future work, information fusion can be performed using the results of the audio-based and vision-based analysis for different use cases. In audio-based analysis, consecutive sliding windows can be aggregated for a more robust detection of the audio event. The use cases can include human detection using microphone arrays on UAVs. For this purpose, environmental noise can be introduced to the target audio classes. For the visual analysis, human body and pose detection can be employed for detecting people. Furthermore, to the seven basic facial expressions, "Contempt" can be added as it can mean a dangerous and destructive form of conflict.

REFERENCES

[1] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[3] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[4] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520, 2016.

[5] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "Ssh: Single stage headless face detector," in *Proceedings of the IEEE international conference on computer vision*, pp. 4875–4884, 2017.

[6] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020.

[7] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533, 2016.

[8] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: multi-head cross attention network for facial expression recognition," *arXiv preprint arXiv:2109.07270*, 2021.

[9] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, "Speech recognition with no speech or with noisy speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1090–1094, IEEE, 2019.

[10] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2010.

[11] S. Patil, B. Patil, and G. Tatkare, "Gender recognition and age approximation using deep learning techniques," *y (IJERT)*, vol. 9, no. 4, 2020.

[12] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 351–376, 2021.

[13] J. Gao, M. Gong, and X. Li, "Audio-visual representation learning for anomaly events detection in crowds," *arXiv preprint arXiv:2110.14862*, 2021.

[14] G. Sreenu and S. Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," *Journal of Big Data*, vol. 6, no. 1, pp. 1–27, 2019.

[15] M. Thida, H.-L. Eng, and P. Remagnino, "Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 2147–2156, 2013.

[16] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, and L. Marcenaro, "Audio-based search and rescue with a drone: highlights from the ieee signal processing cup 2019 student competition [sp competitions]," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 138–144, 2019.

[17] E. S. d. Lima and B. Feijó, "Artificial intelligence in human-robot interaction," in *Emotional Design in Human-Robot Interaction*, pp. 187–199, Springer, 2019.

[18] J. Kim, K. Min, M. Jung, and S. Chi, "Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition," *Building and Environment*, vol. 181, p. 107092, 2020.

[19] A. Nair, P. Singh, and S. Arora, "Detection of audio-based emergency situations using scale," 2019.

[20] C. Lai and L. Jiang, "An intelligent baby care system based on iot and deep learning techniques," *International Journal of Electronics and Communication Engineering*, vol. 12, no. 1, pp. 81–85, 2018.

[21] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *2005 IEEE International Conference on Multimedia and Expo*, pp. 1306–1309, IEEE, 2005.

[22] L. Vilaça, Y. Yu, and P. Viana, "Recent advances and challenges in deep audio-visual correlation learning," *arXiv preprint arXiv:2202.13673*, 2022.

[23] P. Morgado, Y. Li, and N. Nvasconcelos, "Learning representations from audio-visual spatial alignment," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4733–4744, 2020.

[24] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[25] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

[26] B. Korbar, "Co-training of audio and video representations from self-supervised temporal synchronization," 2018.

[27] M. Li, D. Li, N. Dimitrova, and I. Sethi, "Audio-visual talking face detection," in *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, vol. 2, pp. II–473, IEEE, 2003.

[28] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-based multimodal speaker detection for distributed meeting videos," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1541–1552, 2008.

[29] Y. Cao, S. Baang, S.-H. Liu, M. Li, and S. Hu, "Audio-visual event classification via spatial-temporal-audio words," in *2008 19th International Conference on Pattern Recognition*, pp. 1–5, IEEE, 2008.

[30] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance*, pp. 65–72, IEEE, 2005.

[31] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," in *Proceedings of the fourth ACM international conference on Multimedia*, pp. 21–30, 1997.

[32] T. Zhang and C.-C. J. Kuo, "Hierarchical system for content-based audio classification and retrieval," in *Multimedia Storage and Archiving Systems III*, vol. 3527, pp. 398–409, SPIE, 1998.

[33] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[34] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[35] E. C. Erkuş, V. Purutçuoğlu, and E. Purutçuoğlu, "Detection of abnormalities in heart rate using multiple fourier transforms," *International Journal of Environmental Science and Technology*, vol. 16, no. 9, pp. 5237–5242, 2019.

[36] Q. Zhou, J. Shan, W. Ding, C. Wang, S. Yuan, F. Sun, H. Li, and B. Fang, "Cough recognition based on mel-spectrogram and convolutional neural network," *Frontiers in Robotics and AI*, p. 112, 2021.

[37] S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method," *IEEE transactions on speech and audio processing*, vol. 8, no. 5, pp. 619–625, 2000.

[38] P. Wan and L. Lu, "Content-based audio retrieval: a comparative study of various features and similarity measures," in *Multimedia Systems and Applications VIII*, vol. 6015, pp. 508–515, SPIE, 2005.

[39] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *2010 18th European signal processing conference*, pp. 1267–1271, IEEE, 2010.

[40] A. Temko, C. Nadeu, and J.-I. Biel, "Acoustic event detection: Svm-based system and evaluation setup in clear'07," in *Multimodal Technologies for Perception of Humans*, pp. 354–363, Springer, 2007.

[41] S. Adavanne and T. Virtanen, "Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network," *arXiv preprint arXiv:1710.02998*, 2017.

[42] H. Aronowitz, "Segmental modeling for audio segmentation," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–393, IEEE, 2007.

[43] Z. Kons, O. Toledo-Ronen, and M. Carmel, "Audio event classification using deep neural networks.," in *Interspeech*, pp. 1482–1486, 2013.

[44] B. Elizalde, R. Revutchi, S. Das, B. Raj, I. Lane, and L. M. Heller, "Identifying actions for sound event classification," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 26–30, IEEE, 2021.

[45] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6460–6464, IEEE, 2016.

[46] F. S. Saeed, A. A. Bashit, V. Viswanathan, and D. Valles, "An initial machine learning-based victim's scream detection analysis for burning sites," *Applied Sciences*, vol. 11, no. 18, p. 8425, 2021.

[47] R. O'Donovan, E. Sezgin, S. Bambach, E. Butter, S. Lin, *et al.*, "Detecting screams from home audio recordings to identify tantrums: Exploratory study using transfer machine learning," *JMIR Formative Research*, vol. 4, no. 6, p. e18279, 2020.

[48] N. Çalik, L. D. Ata, A. Serbes, B. Bolat, and E. Yavuz, "Performance analysis of feature extraction methods in indoor sound classification," in *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pp. 2025–2028, IEEE, 2015.

[49] Q. Nguyen, S.-S. Yun, and J. Choi, "Detection of audio-based emergency situations using perception sensor network," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 763–766, IEEE, 2016.

[50] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780, IEEE, 2017.

[51] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[53] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[54] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[55] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[56] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[57] S. Liu, H. Zhou, C. Li, and S. Wang, "Analysis of anchor-based and anchor-free object detection methods based on deep learning," in *2020 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1058–1065, IEEE, 2020.

[58] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[59] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9759–9768, 2020.

[60] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 734–750, 2018.

[61] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6569–6578, 2019.

[62] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

[63] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *Proceedings of the IEEE international conference on computer vision*, pp. 192–201, 2017.

[64] J. Wang, Y. Yuan, and G. Yu, "Face attention network: An effective face detector for the occluded faces," *arXiv preprint arXiv:1711.07246*, 2017.

[65] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th international conference on information science and control engineering (ICISCE)*, pp. 424–427, IEEE, 2017.

[66] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[67] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 8231–8238, 2019.

[68] X. Li, S. Lai, and X. Qian, "Dbcface: Towards pure convolutional neural network face detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1792–1804, 2021.

[69] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 797–813, 2018.

[70] W. Wang, Y. Yan, Z. Cui, J. Feng, S. Yan, and N. Sebe, "Recurrent face aging with hierarchical autoregressive memory," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 654–668, 2019.

[71] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 397–403, 2013.

[72] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pp. 2144–2151, IEEE, 2011.

[73] H. Ouanan, M. Ouanan, and B. Aksasse, "Facial landmark localization: Past, present and future," in *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pp. 487–493, IEEE, 2016.

[74] C. Wang, "The development and challenges of face alignment algorithms," in *Journal of Physics: Conference Series*, vol. 1335, p. 012009, IOP Publishing, 2019.

[75] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou, "The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 599–624, 2019.

[76] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou, "Joint multi-view face alignment in the wild," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3636–3648, 2019.

[77] J. Deng, Y. Zhou, S. Cheng, and S. Zaferiou, "Cascade multi-view hourglass model for robust 3d face alignment," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 399–403, IEEE, 2018.

[78] V. V. Salunke and C. Patil, "A new approach for automatic face emotion recognition and classification based on deep networks," in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1–5, IEEE, 2017.

[79] A. Mehrabian, "Communication without words," in *Communication theory*, pp. 193–200, Routledge, 2017.

[80] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[81] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[82] X. Zhao and S. Zhang, "A review on facial expression recognition: feature extraction and classification," *IETE Technical Review*, vol. 33, no. 5, pp. 505–517, 2016.

[83] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[84] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[85] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[86] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[87] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 118–126, IEEE, 2017.

[88] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[89] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[90] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, and S. B. Musa, "The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi," in *2020 Fifth international conference on informatics and computing (ICIC)*, pp. 1–9, IEEE, 2020.

[91] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101, IEEE, 2010.

[92] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The japanese female facial expression (jaffe) database," in *Proceedings of third international conference on automatic face and gesture recognition*, pp. 14–16, 1998.

[93] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[94] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2852–2861, 2017.

[95] V. Singhal and A. Majumdar, "Age and gender estimation via deep dictionary learning regression," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[96] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[97] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[98] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3ae: Exploring the limits of compact model for age estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12587–12596, 2019.

[99] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[100] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[101] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The nigens general sound events database," *arXiv preprint arXiv:1902.08314*, 2019.

[102] J. Xu, "A deep learning approach to building an intelligent video surveillance system," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5495–5515, 2021.

[103] M. G. Hluchyj and M. J. Karol, "Shuffle net: An application of generalized perfect shuffles to multihop lightwave networks," *Journal of Lightwave Technology*, vol. 9, no. 10, pp. 1386–1397, 1991.

[104] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[105] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*, pp. 391–405, Springer, 2014.

[106] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

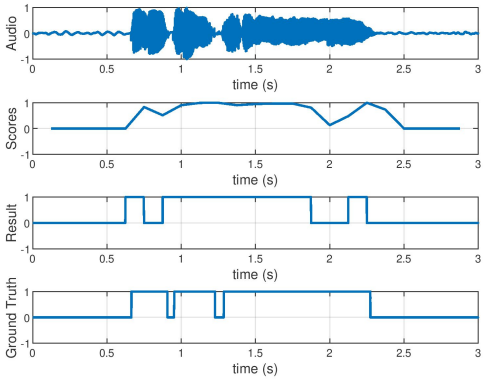
[107] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 237–242, 2020.

[108] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

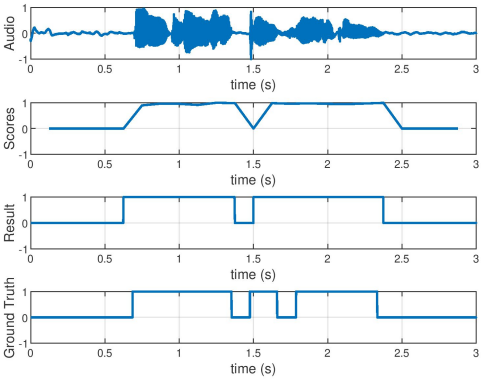
APPENDIX A

FIGURES FOR CHAPTER 4

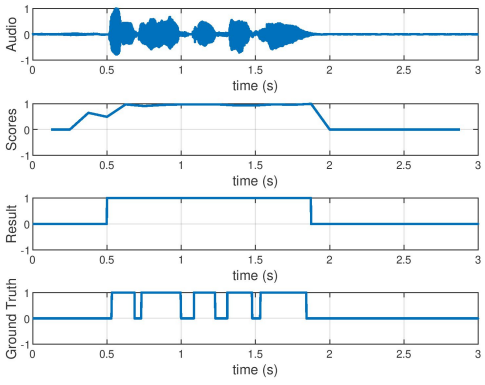
A.1 Target Audio Class Detection



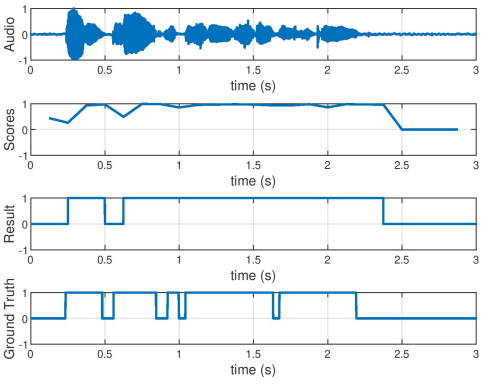
(a)



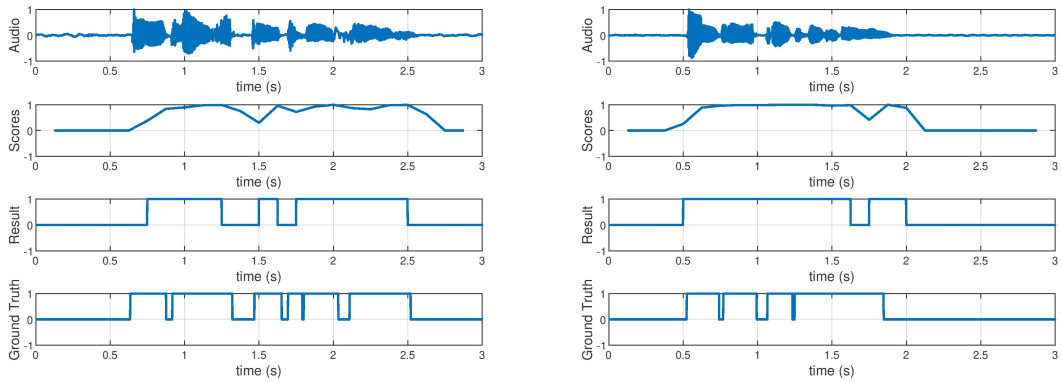
(b)



(c)

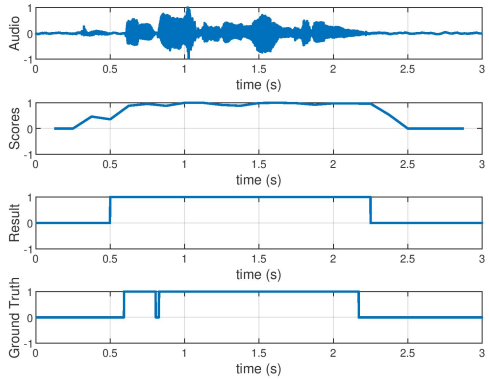


(d)

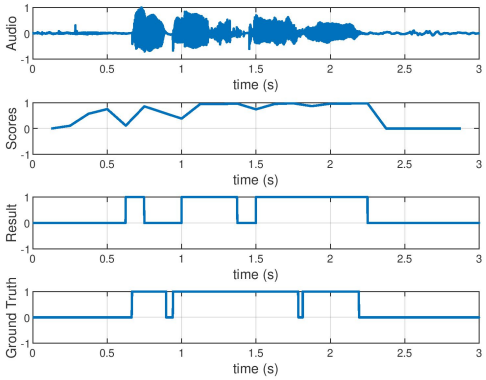


(e)

(f)

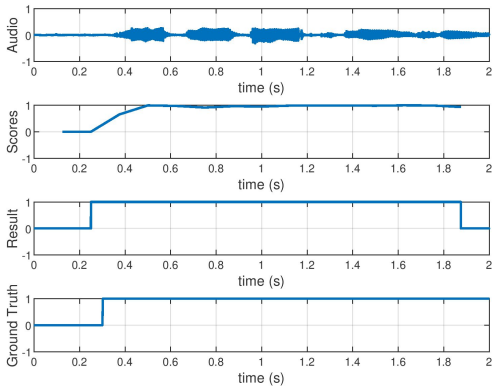


(g)

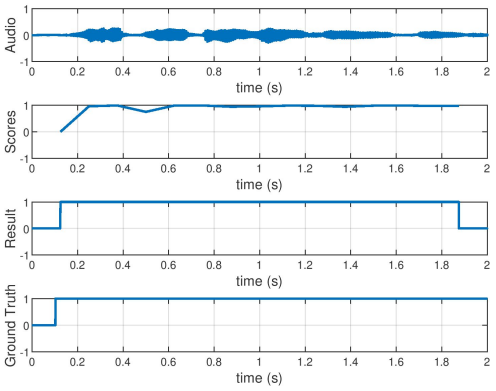


(h)

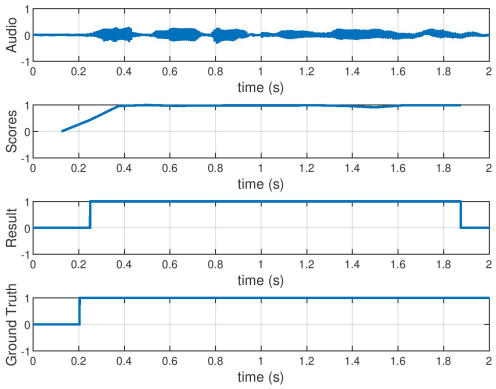
Figure 42: Audio signal, scores, results and ground truth and ground truth for "Female Speech" class for target audio class detection.


Audio
Scores
Result
Ground Truth
time (s)

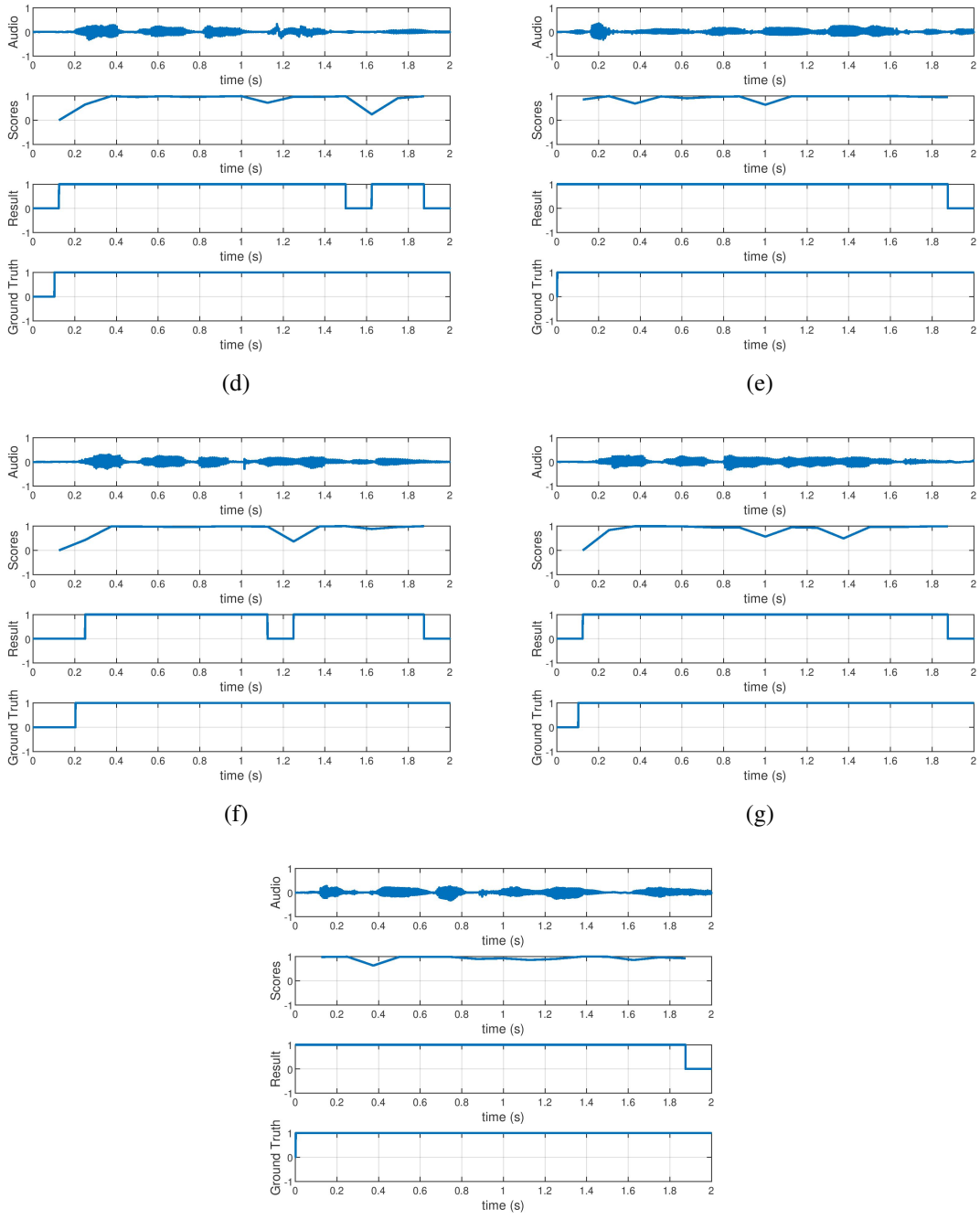
(a)


Audio
Scores
Result
Ground Truth
time (s)

(b)


Audio
Scores
Result
Ground Truth
time (s)

(c)

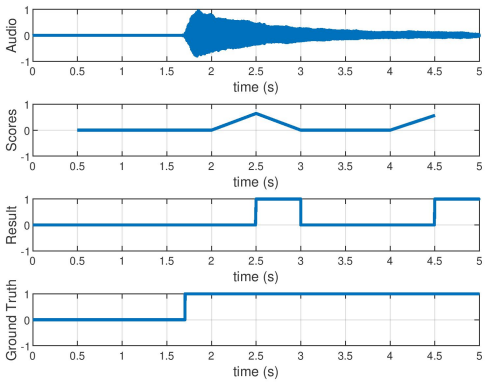


(d) (e)

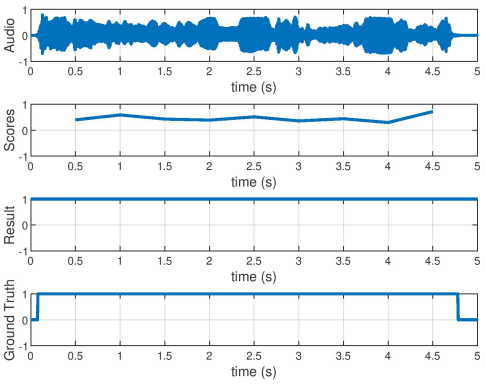
(f) (g)

(h)

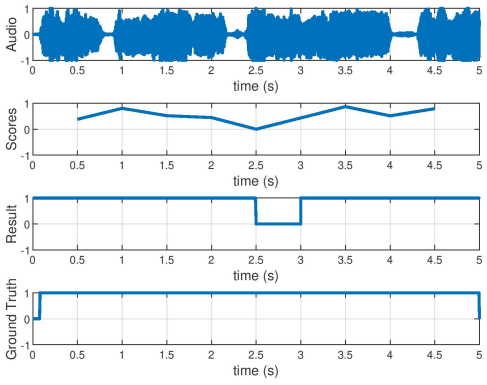
Figure 43: Audio signal, scores, results and ground truth for "Male Speech" class for target audio class detection.


Audio
Scores
Result
Ground Truth
time (s)

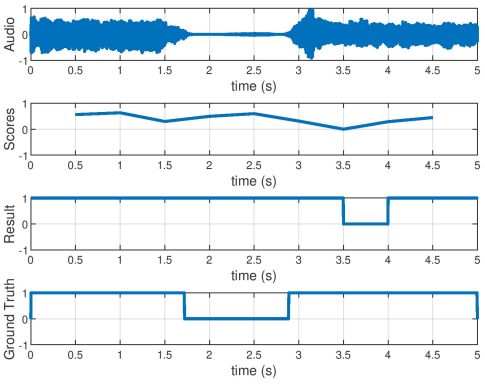
(a)


Audio
Scores
Result
Ground Truth
time (s)

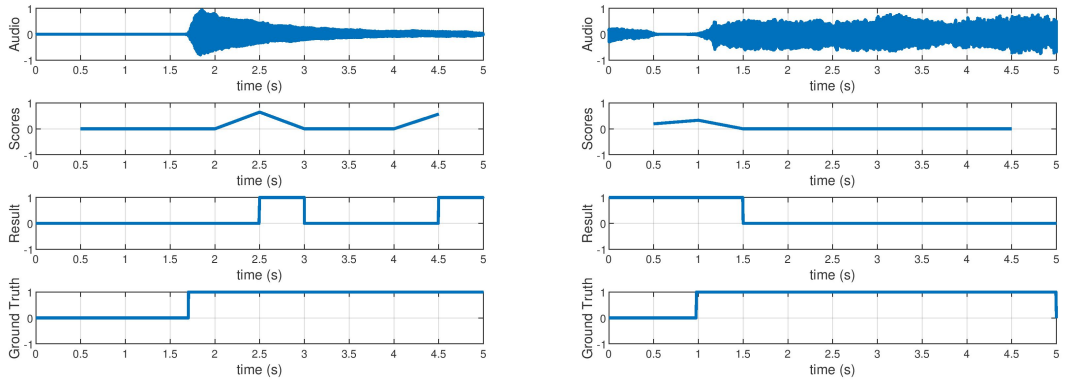
(b)


Audio
Scores
Result
Ground Truth
time (s)

(c)

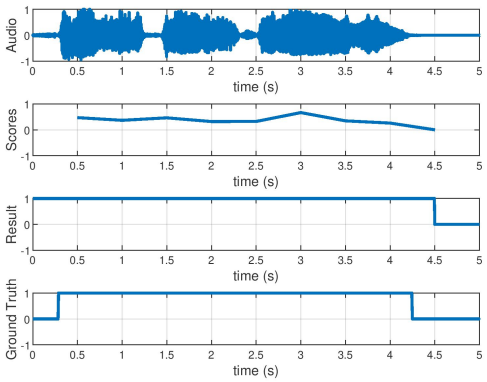

Audio
Scores
Result
Ground Truth
time (s)

(d)



(e)

(f)



(g)

Figure 44: Audio signal, scores, results and ground truth for "Female Scream" class for target audio class detection.

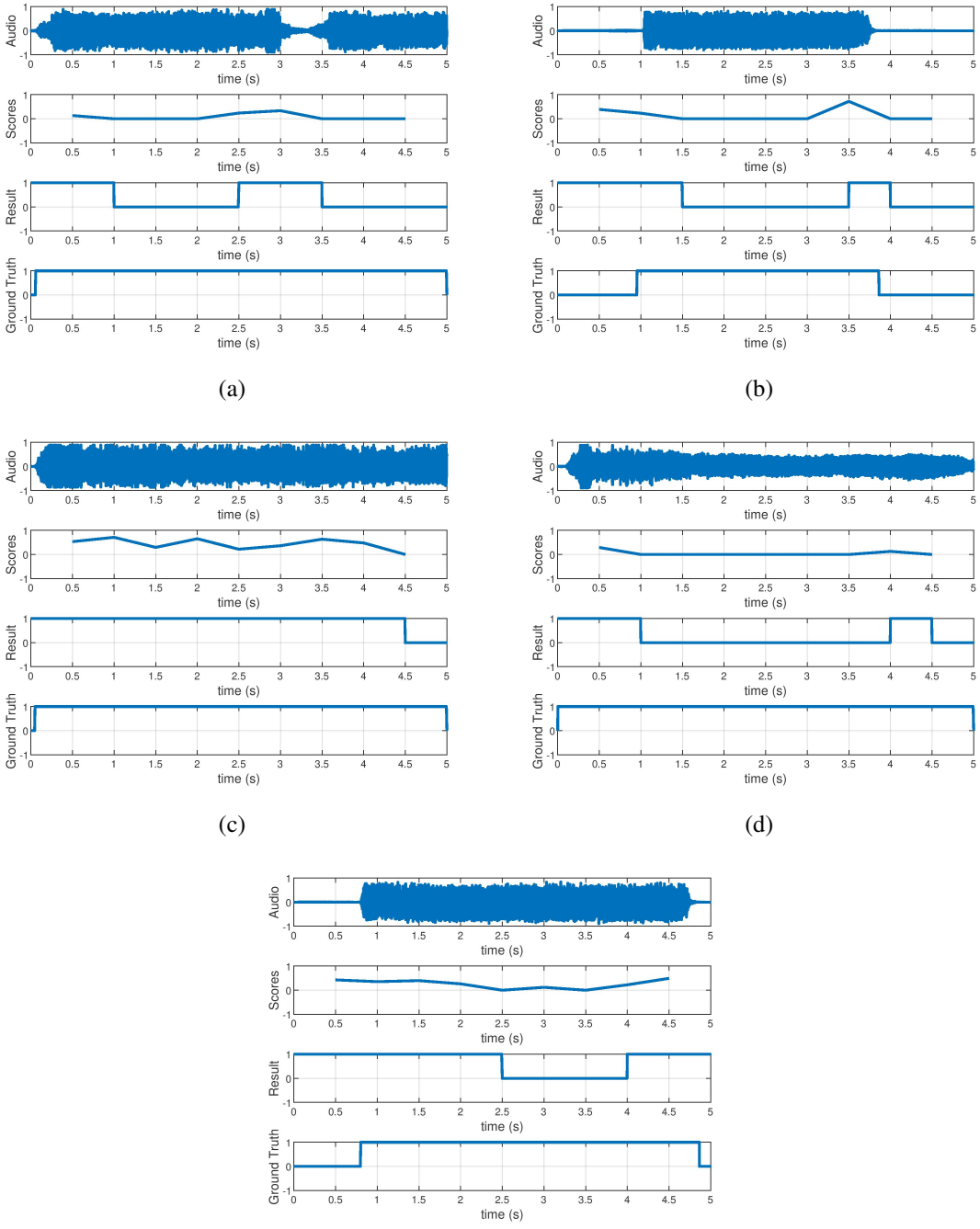


Figure 45: Audio signal, scores, results and ground truth for "Male Scream" class for target audio class detection.

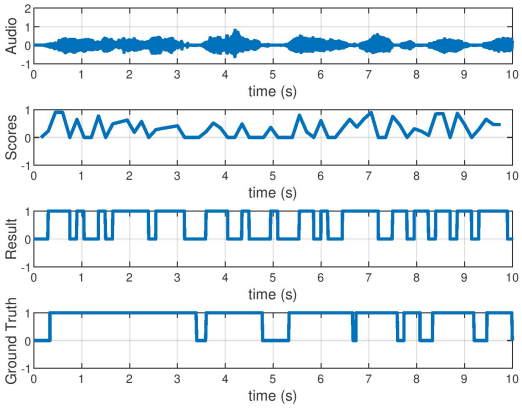


Figure 46: Result for "Baby Cry" class for target audio class detection.