

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

Word level Amharic Sign Language Recognition using Deep Learning algorithms

Biruk Mengiste Kassa

A Thesis Submitted to the Department of Electrical and Computer Engineering in Partial Fulfillments for the Degree of Master of Science in Computer Engineering.

Addis Ababa, Ethiopia

FEB 9, 2022

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

Biruk Mengiste Kassa

Advisor Dr Getachew Alemu (PhD)

This is to certify that the thesis prepared by Biruk Mengiste Kassa, titled word level Amharic sign language recognition using deep learning algorithms and submitted in partial fulfilment of the requirements for the Degree of Master of Science in Computer Engineering compiles with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by Examining Committee:

	Name	Signature	Date
Advisor:	Dr.Getachew Alemu (PhD)	_____	_____
Examiner:	Dr. Bisrat Derebssa (PhD)	_____	_____
Examiner:	Dr.Surafel Lemma (PhD)	_____	_____

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in this and in any other university, and that all source of materials used for the thesis have been fully acknowledged.

Declared by:

Name: Biruk Mengiste Kassa

Signature: _____

Date: _____

Confirmed by my advisor:

Name: Dr. Getachew Alemu (PhD)

Signature: _____

Date: _____

Acknowledgments

First of all, I would like to thank almighty God and his Mother for giving me the strength, peace of my mind, and good health to achieve whatever I have achieved so far and for guiding me all the way through.

I would like to express my sincere gratitude to my advisor Dr. Getachew Alemu (PhD)_for his consistent follow-up and his willingness to offer me his time and knowledge from the inception to the completion of this thesis.

I am extremely grateful to my wife and parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. They have been my source of strength next to God.

My sincere thanks also goes to Zenash Menken, Wendmu Debesh and Getnet Kefle who have dedicated their time for supplying different materials or equipment when they were needed. Some special thanks goes to Mrs. Shambel Belay and her some students for capturing the data set. I also wish to express my gratitude to the IT staff members for providing a training class.

My thanks extend to Eshete Damte who has spent many hours for proof, experiment, reading discussion and constructive comments on my report.

The final word of thanks goes to people who are not mentioned in name but whose support helped me complete the study successfully. Thanks to all.

Abstract

In Ethiopia, Deaf peoples are vastly increase in number. Sign language is a natural language mostly used by Deaf persons to communicate with each other. However, during communication there is a big challenge between Deaf and normal person. Deaf use sign for communication whereas normal person use speech/text for communication.

We need efficient system to exchange sign to speech/text or speech/text to sign. This thesis work focus on development of word level Amharic sign language recognition, translates Amharic word sign into their corresponding Amharic text using deep learning approach. The input for the system is video frames of Amharic sign words and the final output of the system is Amharic text.

The proposed system has three major components: preprocessing, feature extraction and classification. Two preprocessing steeps were used, cropping and RGB to Grayscale conversion. Feature extraction was done by using deep residual network (ResNet-34) and store in .csv format.

Finally, classification was done by the same deep learning algorithms ResNet-34. The system is trained and tested using a dataset prepared for this thesis purpose only for all Amharic sign words. The performance of the model measured by four different matrices (precision, recall, F1 score and accuracy).

The system classify 60 sign words and score overall accuracy of 95%. Therefore, the classification performance of ResNet-34 is very good.

Key words: Amharic sign words, deep learning algorithms, ResNet-34

Table of Contents

Chapter one: Introduction 1
1.1 Introduction 1
1.2 Statement of a problem 3
1.3 Objectives 4
1.3.1 General Objectives 4
1.3.2 Specific Objectives 4
1.4 Scope and Limitation 4
1.4.1 Scope 4
1.4.2 Limitation 4
1.5 Motivation 5
1.6 Contributions 5
1.7 Methodology 5
1.8 Thesis organization 6
Chapter Two: Background 7
2.1 World gesture based communication 7
2.1.1 American Sign Language 7
2.1.2 South African sign Language 9
2.2 Ethiopian gesture based communication 10
2.2.1 Common frequently used words 11
2.2.2 Body parts 12
2.2.3 Family member signs 14
2.2.4 Days of the week 15
2.2.5 Food and drink 17
2.2.6 Color 19
2.3 Summary 22
Chapter Three: Literature review 23
3.1 Local sign language recognition systems 23
3.2 FOREIGN GESTURE BASED COMMUNICATION ACKNOWLEDGMENT FRAMEWORKS 25
3.2.1 Convolution Neural Network (CNN) 25
3.2.2 VGG 27
3.2.3 ResNet 28
3.3 Summary 29
Chapter Four: Methodology 31
4.1 Introduction 31
4.2 Preparing frequently used sign words 32
4.3 Assign signers and record sign words 33
4.4 Video to frame conversion 34

4.5 Preprocess the frames 35
4.5.1 Cropping 35
4.5.2 Convert RGB to Grayscale 36
4.6 Pixel based image recognition algorithm 36
4.7 Vanishing gradient and Degradation problem 41
4.7.1 Vanishing Gradient 41
4.7.2 Degradation problem 42
4.7.3 Residual network 42
4.8 Feature extraction by ResNet-34 43
4.9 3D Data Classification through training by ResNet-34 44
4.10 How to test the system. 44
4.11 Summary 44
Chapter Five: Experimentation and result Discussion 46
5.1 Introduction 46
Develop a word sign recognition system and test the impact on communication with deaf. 5.2 Dataset preparation 46
5.3 The directory structure of the data set 50
5.4 Experimental Setup 51
5.5 Experimental Senarios 51
5.5 Result 53
5.6. Threats to Validity 74
5.6.1. Internal Threats to Validity 74
5.6.2. External Threats to Validity 74
5.7 Discussion 75
Chapter Six: Conclusion and Future Work 76
6.1 Conclusion 76
6.2 Future Work 77
Reference 79
Appendix A: Sample Data Used for System Design 81
Appendix B: PYTHON Code 83

List of Tables

Table 2-1: Most frequently sign words 11
Table 2-2: Body part signs 13
Table 2-3: Family member signs 14
Table 2-4: Days of the week signs 16
Table 2-5: Food and Drink signs 17
Table 2-6: colors sign 20
Table 3-1: Training and validation accuracy on RGB image 26
Table 3-2: Training and validation accuracy on grayscale image 26
Table 3-3: Local recognition systems 29
Table 3-4: foreign recognition system 30
Table 4-1: Frequently used sign words 32
Table 4-2: Three Amharic sign derived alphabet 33
Table 4-3: Some filters used for convolution 38
Table 5-1: Dataset organization 46
Table 5-2: Amharic Sign Words dataset 47
Table 5-3: Amharic derived latter's dataset 49
Table 5-4: Argentinian sing word dataset 49
Table 5-5: Experimental setup 51
Table 5-6: Evaluation result for Amharic sign word 57
Table 5-7: Classification comparison result for ResNet-34 and NN on the derived Amharic sign letters. 62
Table 5-8: Classification result comparison for Derived Amharic sign letters (ResNet-34 vs SVM) 66
Table 5-9: LSTM Vs ResNet-34 71

List of Figures

Figure 2-1: American some sign word 8
Figure 2-2: South African some sign words 10
Figure 4-1: The general overview of Amharic sign word recognition system 31
Figure 4-2: Flow chart for extract frame from video 35
Figure 4-3: Frame cropping algorithm 35
Figure 4-4: Grayscale converted frame 36
Figure 4-5: Convolution operation (CNN) 37
Figure 4-6: Non-Linearity RLU 39
Figure 4-7: Max pooling operating (CNN) 40
Figure 4-8: Vanishing Gradient 41
Figure 4-9: Identity in Residual Network 43
Figure 4-10: Feature extraction by ResNet-34 43
Figure 4-11: ResNet-34 modified layers for classification 44
Figure 5-1: Training accuracy algorithms 54
Figure 5-2: Training accuracy curve 54
Figure 5-3: Training loss algorithm 55
Figure 5-4: Training loss curve 55
Figure 5-5: Training and validation curve algorithm 56
Figure 5-6: Test accuracy vs Training accuracy curve 56
Figure 5-7: Accuracy of the proposed model 59
Figure 5-8: Accuracy, Precision and recall for some Amharic words 60
Figure 5-9: Accuracy precision and recall for Amharic derived latters 61
Figure 5-10: Comparison of NN and ResNet-34 bar graph 63
Figure 5-11: NN VS ResNet-34 64

Figure 5-12: Precision for NN Vs ResNet-34 64
Figure 5-13: Recall for NN Vs ResNet-34 65
Figure 5-14: Comparison graph for SVM Vs ResNet-34 67
Figure 5-15: Accuracy for SVM Vs ResNet-34 68
Figure 5-16: Precision for SVM Vs ResNet-34 68
Figure 5-17: Recall for SVM ResNet-34 69
Figure 5-18: Argentinian sign language result 70
Figure 5-19: Accuracy for LSTM Vs ResNet-34 74

List of Equations

Equation 5-1: Calculate the accuracy of the model 52
Equation 5-2: Calculate the precession of the model 52
Equation 5-3: Calculate the recall of the model 53

List of Acronyms

BSL- British Sign Language
ASL American Sign Language
LSA Argentine Sign Language
SASL South African Sign Language
ISL Indian Sign Language
BASL Bangladesh Sign Language
CHSL Chinese Sign Language
LSF French Sign Language
EFSCS Ethiopian Finger Spelling Characterization Framework
PCA Principal Component Analysis
NN Neural Network
SVM Support Vector Machine
AI Artificial Intelligence
CNN Convolution Neural Network
RNN Recurrent Neural network
GPU Graphical Processing Unit
SSD Solid State Drive
GRU Gated recurrent unit
LSTM Long short term memory
SVLM Space Variation Luminance Map based picture improvement technique
ReLu Non-Linearity
ETHMA Ethiopian Manual Alphabet **ETHSL**

CHAPTER ONE: INTRODUCTION

✓ 1.1 Introduction

A particular country or region uses a communication system that have set of sounds and written symbols for talking or writing. Language is a system that consists of the development acquisition, maintenance and use of complex system of communication, particularly the human ability to do so. Human language has the properties of productivity and displacement, and relies entirely on social conversion and learning.

All languages have underlying structural rules that make meaningful communication possible. The five main components of language are phonemes, morphemes, lexemes, syntax, and context. Along with grammar, semantics, and pragmatics, these components work together to create meaningful communication among individuals.

Now a day's one of the appropriate and developing language all over the world is sign language. Sign Language is a complete, natural language that has the same linguistic properties as spoken languages, with grammar that differs from spoken languages. Sign language is expressed by movements of the hands and face. It is the primary language of many Deaf societies, and is used by many hearing people as well.

There are many sign language exist in the world. Different sign languages used in different countries or regions. For examples, British Sign Language (BSL), American Sign Language (ASL), South African Sign Language (SASL), Argentine Sign Language (LSA), Brazilian Sign Language and so on. Americans who know American Sign Language may not understand British Sign Language. Some countries adopt features of American Sign Language in their sign languages. Movements of the hands and face express American Sign Language. It is the primary language of many North Americans.

In broad, there are born Deaf and late Deaf. People who are born hearing and become hard of hearing late in life, are physically Deaf, but culturally hearing [1]. They grew up speaking

a spoken language, using the telephone, the TV, the radio, and so on. They speak, read, write and base their opinions on the world they knew before they became Deaf and can describe their idea in a better way than born Deaf can. People who are born the Deaf community, first native language are a sign language, not a spoken one, are culturally Deaf. This people view the world from their perspectives. They are physically hearing but culturally Deaf [1].

Deaf in Ethiopia certainly brings a special set of challenges, and nationwide, there are very few services available to the hearing impaired community. Being Deaf in Ethiopia has its own troubles to access basic information or services, receive an education, communicate with the rest of the world, hold a meaningful job or trade, participate in basic community activities, and so on.

In families where parents are learning a new language, such as Amharic Sign language, with which to communicate with their child, children have a tendency to acquire inconsistent or incorrect linguistic input [2]. People who are born Deaf can hear nothing at all. In order to communicate with people, they are very reliant on lip-reading and/or sign language. People who are born Deaf find lip- reading much harder to learn compared to those who became hearing impaired after they had learnt to communicate orally or with sounds.

Different research conducted to come up with a system that converts sign to text or vice versa for different sign languages all over the world. This action allows communicating Deaf persons with each other and with other hearing societies easily. However, in Ethiopia still there is a language barrier. It is difficult to communicate with Deaf peoples. So need more research on the area to eliminate the communication gap between hearing impaired peoples with others.

Technically speaking, the main challenge of sign language recognition lies in developing descriptors to express hand-shapes and motion trajectory. In particular, hand-shape description involves tracking hand regions in video stream, segmenting hand-shape images from complex background in each frame and gestures recognition problems. Motion trajectory also related to tracking of the key points and curve matching. Although many

research works have been conducted on these two issues for now, it is still hard to obtain satisfying result for sign language recognition due to the variation and occlusion of hands and body joints. Besides, it is a nontrivial issue to integrate the hand-shape features and trajectory features together.

✓ 1.2 Statement of a problem

Language is a system of communication used by a particular country or community. It is the method of human communication and most of the language represented by spoken. Unfortunately, some of human beings do not possess this skill due to much reason. This impaired people need some special communication language called sign language.

In Ethiopia, Legesse Zerubabel [3] has attempted to develop a recognition system for Amharic alphabet signs which translates a given alphabet sign into text. His work only focuses on the recognition of selected ten basic alphabet signs from static images and has certain limitation not detect and classify the motion of hand signers.

In addition, Nigus Kefyalew [4] has attempted to develop a recognition system for Amharic alphabet sign, which translate a given alphabet, sign to text. The developed system recognize all the first Amharic alphabet sign and the derived Amharic alphabet signs of "ሀ" ,"ለ", "ሐ" only. The scope of Nigus Kefyalew also limited. In [4], use for segmentation by adaptive threshold algorithm and extract feature manually. He considers three major feature descriptors: shape feature, motion feature and color feature and classification through SVM and NN. His work is limited on character level too.

In general, there is recognition system in Ethiopia that recognizes all Amharic sign characters and some of the derived sign characters only. However, no system in Ethiopia that recognize sign words and sentence. The existed sign recognition system extract feature with manual based that may lids to error. Video classification is done by old machine learning algorithms rather it is better to use deep learning algorithms to achieve a high classification result.

In this thesis work, try to develop a new system by using deep neural network that recognize Amharic words. The new system train and test by collected video dataset for different signers.

The proposed word level sign recognition system want to recognize frequently used 60 sign words from sign videos. In this approach, to fulfill the above limitations we use deep neural network for feature extraction and classification.

RQ1: Develop a word sign recognition system and test the impact on communication with deaf.

1.3 Objectives

1.3.1 General Objectives

The general objective of this study is to design a new recognition system and test Amharic sign words using deep learning algorithm.

1.3.2 Specific Objectives

The specific objectives are:

- Collect the data set from different signers for this thesis and future research.
- Design a new recognition system by using deep learning algorithm ResNet-34
- Detect and classify Amharic signs words.
- Recognize common 60 sign words.
- 3D feature extraction and classification on Amharic words by using deep neural network.

1.4 Scope and Limitation

1.4.1 Scope

Word level acknowledgment is the exploration scope of this research.

1.4.2 Limitation

This thesis work has the following limitations:

- This application not recognizes all Amharic words.
- Works on word level recognition only. Phrase and Sentence level recognition is not included in this thesis.
- This thesis is tested on ResNet deep learning algorithm only
- Abbreviations are not included in this research.

✓ 1.5 Motivation

- Hearing-impaired peoples are dramatically increased in number.
- There is a big communication gap between hearing and hearing-impaired people
- Software developing tools like python are free
- A lot of research works are done in this area.

✓ 1.6 Contributions

The main contributions of this research can be summarized as:

- Recognition of Amharic sign words.
- Implementation of 3D deep neural network on Amharic sign words.
- Preparation of Amharic sign words video dataset.

✓ 1.7 Methodology

Research methodology follows the following steps:

- ✓ Preparing tools for further processing
  - Python built in function will do the extraction and classification process.
  - Microsoft Visio 2010, Microsoft word 2019 and latex used for designing.
- ✓ prototype development and Evaluation
  - Signer generally use a prototype to evaluate a new design to enhance precision.
  - Prototyping serves to provide specifications for a real, working system rather than a theoretical one.

1. Data collection
   - Data collection starts with determining what kind of data required.

- Selection of a sample from a certain signers.
- Word level Amharic sign video which used as a data set to train as well as to test the recognition system.

2. Extract the frames from multiple video sequence of each gesture.
3. Pre-processing (RGB to Grayscale conversion and resize the frame) are implemented.
4. Preprocessed frames are given to deep neural network ResNet-34 for feature extraction.
5. Feature extracted frames are given to ResNet-34 for classification through training.
6. Finally, test the model.

✓ 1.8 Thesis organization

The rest of this document is organized as follows.

Chapter two cover the general over view of world sign language. Brief description of signing words and show the backgrounds in word level sign language are described. This chapter explains basic theoretical foundations used for signing words in the deaf community and address signing of family members, singing of colors, dates and some others.

State-of-the-art and challenges in Sign Language are discussed in Chapter Three. This chapter contains review of literatures on Ethiopian Sign Languages recognition and foreign Sign languages recognition researches. From foreign Sign Languages reviewed, state-of-the-art approaches are briefly discussed.

Our proposed approach is explained in Chapter Four (Methodology). After a review of researches done on Amharic Sing Language recognition, the problem identified was design and selection of best features. The best algorithm for feature extraction and classification is selected based on reviewed literature. The proposed approach discussed in this chapter.

The experimental procedures and results are discussed in Chapter Five. In this chapter, experimental procedures, datasets, tools and the experimental scenarios used for evaluating our hypothesis are discussed.

Finally Chapter Six presents conclusions from experimental observations and future works to show further areas of improvement on Sign Language recognition systems.

CHAPTER TWO: BACKGROUND

This chapter cover the general over view of world sign language. Such as, American Sign Language, South African sign language and Ethiopian sign language. In particular, Ethiopian sign language, we focused on how to sign Amharic words related to hand shape, gesture, face expression and orientation [5] are discussed briefly.

✓ 2.1 World gesture based communication

Speech based communication is inappropriate for deaf community. Most of the deaf communities do not have the capability of producing speech. Therefore, they need a special communication language. Using sign language is preferable means of communication between deaf and deaf people with the rest of hearing community. Different sign language system exists all over the world. For example French Sign Language [6], American Sign Language (ASL) [7], Indian sign language (ISL) [8], South African Sign Language (SASL) [9], Chinese Sign Language (CHSL) [10], Bangladesh Sign Language (BASL) [11] and so on.

2.1.1 American Sign Language

American Sign Language (ASL) is a natural language [7], that serves as the predominant sign language of Deaf communities in the United States and most of Anglophone Canada. Besides North America, dialects of ASL and ASL-based creoles are used in many countries around the world, including much of West Africa and parts of Southeast Asia. ASL is also widely learned as a second language, serving as a lingua franca. ASL is most closely related to French Sign Language (LSF). It has been proposed that ASL is a Creole language of LSF, although ASL shows features atypical of Creole languages, such as agglutinative morphology.

Each sign in ASL is composed of a number of distinctive components, generally referred to as parameters. American signs may use one hand or both. All signs can be described using

the five parameters involved in signed languages, which are hand shape, movement, palm orientation, location and non-manual markers.

ASL possesses a set of 26 signs known as the American manual alphabet, which can be used to spell out words from the English language. Such signs make use of the 19 hand shapes of ASL. For example, the signs for 'p' and 'k' use the same hand shape but different orientations. A common misconception is that ASL consists only of finger spelling; although such a method (Rochester Method) has been used, it is not ASL [12].

Finger spelling is a form of borrowing, a linguistic process wherein words from one language are incorporated into another. In ASL, finger spelling is used for proper nouns and for technical terms with no native ASL equivalent. There are also some other loan words which are finger spelled, either very short English words or abbreviations of longer English words, e.g. *O-N* from English 'on', and *A-P-T* from English 'apartment'. Finger spelling may also be used to emphasize a word that would normally be signed otherwise.

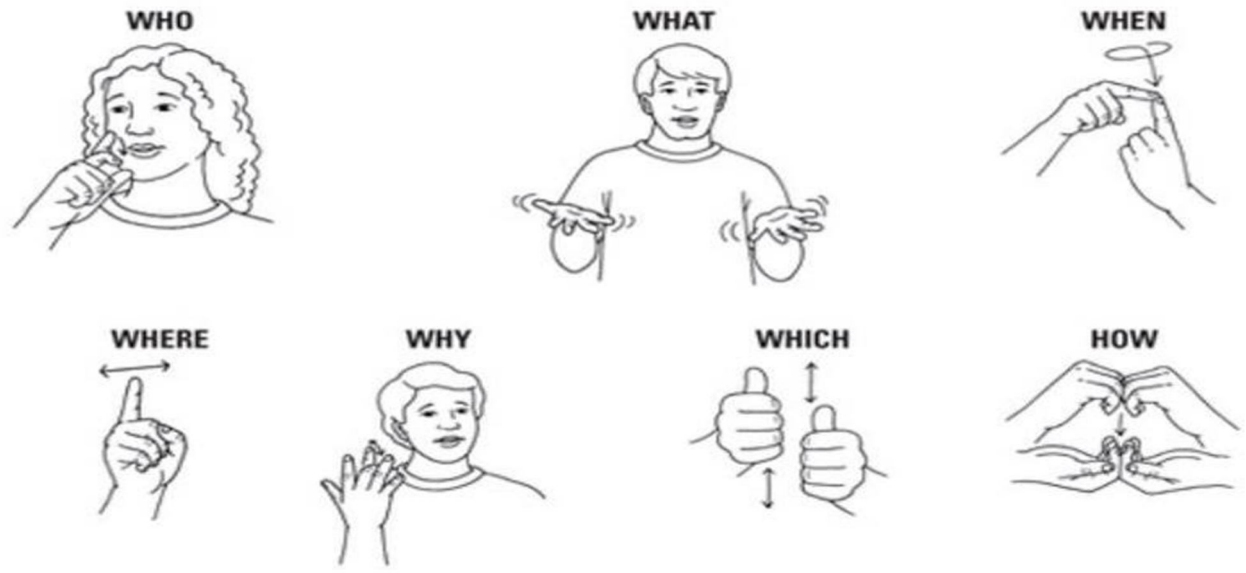


Figure 2-1: American some sign word

As shown from the above American sign words some of sign word are represent by one hand only, some others represent by using two hands and the rest need hand, gesture and orientation to represent sign words. Words expressed above are the same sign expiration to Amharic sign words ማን ፣ ምን ፣ መቼ ፣ የ ት ፣ ለ ምን ፣ የ ትኛ ው፣ እ ን ዴት .

2.1.2 South African sign Language

South African Sign Language (SASL) is the primary sign language used by Deaf people in South Africa. The South African government added a National Language Unit for South African Sign Language in 2001 [9]. SASL is not the only manual language used in South Africa, but it is the language that is being promoted as the language to be used by the Deaf in South Africa, although Deaf peoples in South Africa historically do not form a single group.

Finger spelling is a manual technique of signing used to spell letters and numbers (numerals, cardinals). Therefore, finger spelling is a sign language technique for *borrowing* words from spoken languages, as well as for spelling names of people, places and objects. It is a practical tool to refer to the written word.

Some words which are often finger spelled tend to become signs in their own right (becoming "frozen"), following linguistic transformation processes such as alphanumeric incorporation and abbreviation. For instance, one of the sign-names for Cape Town uses incorporated finger spelled letters *C.T.* (transition from hand shape for letter 'C' to letter 'T' of both wrists with rotation on a horizontal axis). The month of July is often abbreviated as 'J-L-Y'.

Finger spelling words is not a substitute for using existing signs: it takes longer to sign and it is harder to perceive. If the finger spelled word is a borrowing, finger spelling depends on both users having knowledge of the oral language (English, Sotho, Afrikaans etc.). Although proper names (such as a person's name, a company name) are often finger spelled, it is often a temporary measure until the Deaf community agrees on a Sign name replacement.



Figure 2-2: South African some sign words

- ✓ **2.2 Ethiopian gesture based communication**

Based on [13], Ethiopian sign language is derived from American Sign Language. Even if it derived from American Sign Language it differs by Alphabets. Americans use A to Z Alphabets and 26 in numbers whereas, Ethiopian use ሀ to ፐ Alphabets which is called Amharic Alphabets and they are 33 in numbers vertically without the derived letters. Amharic is the second-most commonly spoken Semitic language in the world (after Arabic) [14]. Ethiopian sign Language was developed from Amharic language since Amharic is the official language of Ethiopia.

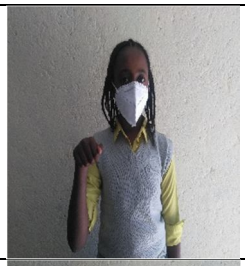
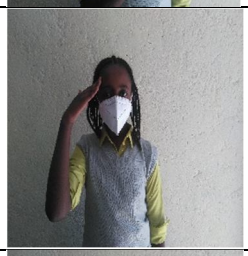
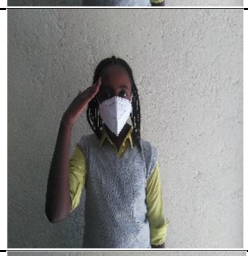
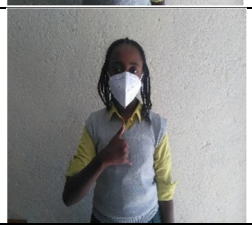
In Ethiopia, there is a limited communication between the hearing community and the deaf community. The deaf community use sign language to communicate among themselves and the hearing community use spoken language for communication. Translator person is a solution for filling the communication gap between the deaf people with non-deaf people. Even though the translators play a great role in the translation process, sometimes it's difficult to exchange confidential information between these two parties, for example in closed court, medical area, and different social issues. In addition, using translator person is not economical cost, time, and effort wise. The Deaf of Ethiopia live everywhere with in the country, but they exclude from meaningful interaction with others. They looked down upon as mentally deficient and evil because of their lack of spoken communication.

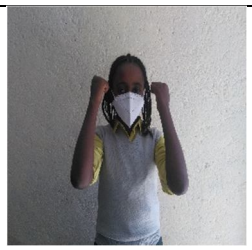
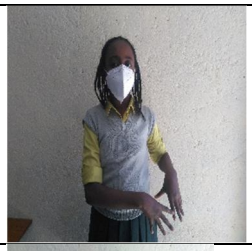
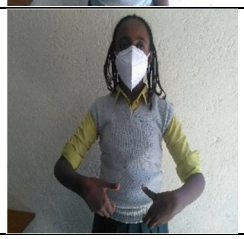
Now a day in towns, there is more awareness generated regarding the Deaf. The government schools are inclusive. Many parents are interested to send their children to schools. Missionaries establish school for the deaf and train the Deaf. In Ethiopia, the Deaf have their

own means of communication. This communication holds manual and lip reading. Manual communication express by hand and super body part for represent gestures.

2.2.1 Common frequently used words

Table 2-1: Most frequently sign words

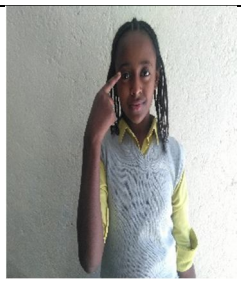
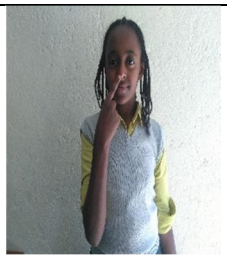
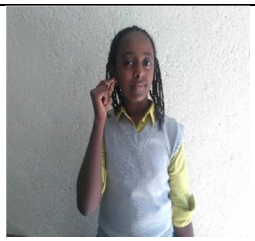
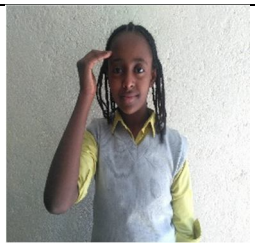
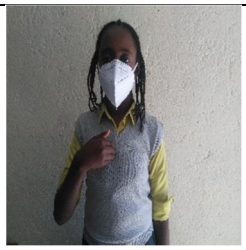
Name of signs	Sign	Descriptions
Yes		‘s’ hand shape making a nodding motion
No		‘1’ hand shape shaking left to right, head moving the same way
Know		Flat hand tapping head
Don’t know		Flat hand tapping head then moving away while shaking head ‘no’
Same		‘y’ hand shape with palm facing down moving back and forth

Forget		Both hands start in ’s’ Hand shapes on sides of head, palms facing backward. Both hands move backward while opening into ‘5’ hand shapes
Student		Both hands start in ‘5’ Hand shapes and close into flattened ‘o’ hand shapes twice, palms facing down
Angry		‘e’ hand shape at the mouth, opening and closing slightly, palm facing in
Tired		Both hands bent, fingertips touching near ribs, then slouching at wrists

2.2.2 Body parts

Human body parts are endorsed by the blend of hand shape and directions. The majority of the body parts are addresses by right hand shape pointed in to the ideal body part. For models, eye, nose, ear, and tongue. Then again, body parts like chest, hand and heart are demonstrated utilizing two hands. Chest express utilizing lift and right pointer finger and drawing heart like picture on the zone of chest. The alternate method of sign of the body part is utilizing bend or circle direction to demonstrate face, leg and others.

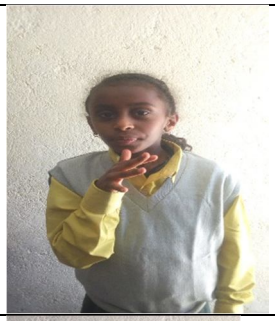
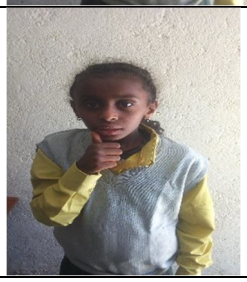
Table 2-2: Body part signs

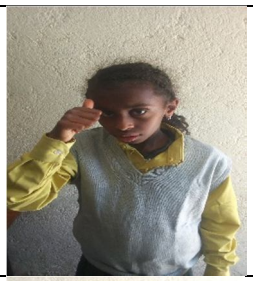
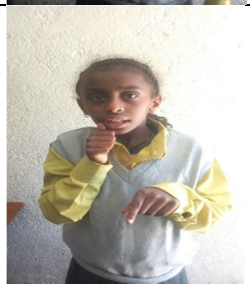
Name of body parts	Sign	Descriptions
Eye		'1' hand shape, pointer finger touching eye
Nose		'1' hand shape, pointer finger touching Nose
Ear		'1' hand shape, pointer finger touching ear
Head		'1' hand shape, pointer finger touching head
Chest		
Leg		'1' hand shape, two hand pointer fingers Indicating down to the leg

2.2.3 Family member signs

Male and female are signs diversely utilized in Amharic gesture-based communication. Male is communicated by putting the right-hand pound associating into the right-side head and showing move away to one side by spread the four fingers. Female is communicated by putting the right-hand pound associating into the right-side jawline and showing move away to one side by spread the four fingers. Father is addressed by putting the bang at the focal point of the brow and vibrating four fingers. Mother is sign by putting the bang at the focal point of the jawline and vibrating four fingers. By the right-hand pound start from the right-side head and goes down to one side hand make association side by utilizing pointer figure which gives the indication of sibling. By the right-hand pound start from the right-side jawline and goes down to one side hand make association side by utilizing pointer figure is sign for sister. The remainder of the relatives are marked dependent on the above male and female sign portrayals.

Table 2-3: Family member signs

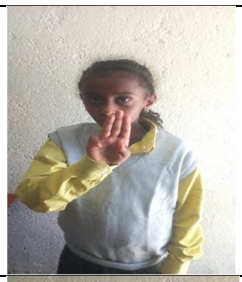
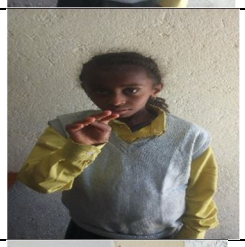
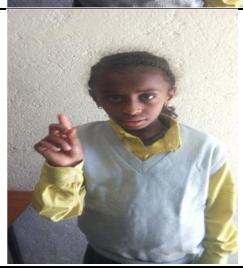
Family name	Sign	Descriptions
Father		'5' hand shape, thumb touching forehead, palm facing down, with fingers wiggling
Mother		'5' hand shape, thumb touching chin, palm facing down, with fingers wiggling
Girl		'a' hand shape making a small line on the chin, repeat once

Boy		‘a’ hand shape making a small line on the temple, repeat once
Sister		Right hand in ‘a’ Hand shape touches side of chin with thumb, then both hands in ‘1’ hand shapes next to each other in front of the chest, palms facing down
Brother		Right hand in ‘a’ Hand shape touches side of head with thumb, then both hands in ‘1’ hand shapes next to each other in front of the chest, palms facing down

2.2.4 Days of the week

Days in sign language are represented by right hand by showing different shape and orientation. The only day that is expressed using two hand signs is Sunday.

Table 2-4: Days of the week signs

day name	Sign	Descriptions
Monday		Ethiopian ‘S2’ Hand shape shaking left and right
Tuesday		Ethiopian ‘m’ Hand shape shaking left and right
Wednesday		Ethiopian ‘r’ Hand shape shaking left and right
Thursday		Ethiopian ‘H2’ Hand shape shaking left and right
Friday		‘f’ hand shape shaking left and right
Saturday		Ethiopian ‘q’ Hand shape shaking left and right

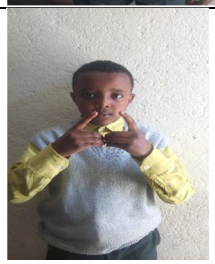
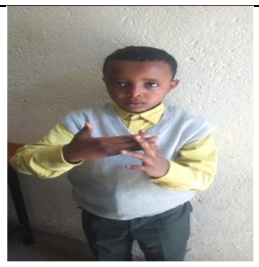
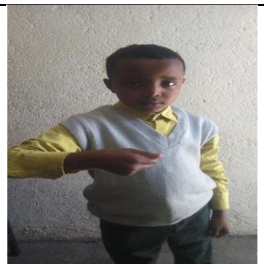
Sunday		Both hands in Ethiopian ‘g’ hand shapes, moving in circles

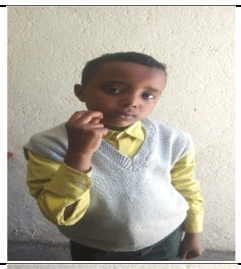
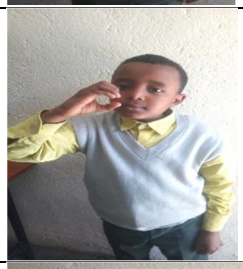
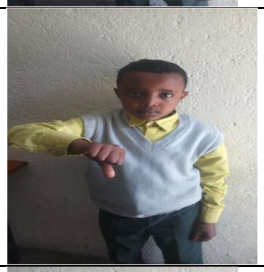
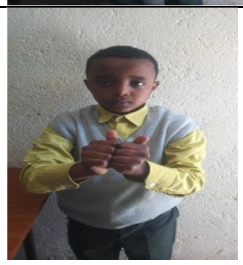
2.2.5 Food and drink

Food and drink in Ethiopian communication via gestures are endorsed by the uncommon conduct of the food and the beverage. For instance, bread (ዳቦ) is endorsed by telling the best way to eat utilizing two hands. Injera (እንጀራ) is communicated by telling the best way to heat እንጀራ, by twisting four fingers and turning the right-hand pound descending and make a circle. Blemish (ጣር) is effortlessly communicated by as though an individual is trying the ጣር by holding the ጣር in left hand Palm and contact it by right hand pointer figure at that point test by tongue. Holding fork and spoon by two hands and show eating activity is getting paperwork done for Pasta (ፓስታ). Macaroni (ሞኮረኒ) is communicated in equivalent to ፓስታ; the thing that matters is that eat sign is communicated by utilizing fork as it were. Milk (ወተት) is communicated by telling the best way to drain. As a rule, the portrayal of food and drink in Ethiopian gesture-based communication is connected with the activities one does in burning-through the feast/drink.

Table 2-5: Food and Drink signs

Name of food	Sign	Descriptions
Food		Both hands in flattened ‘o’ Hand shapes, right hand above left hand, shaking slightly, right hand in front of lips

Eat		One hand in flattened ‘o’ Hand shape, moves toward the lips like putting food in your mouth
Fruit		Both hands in Ethiopian ‘f’ Hand shapes, start with pointer fingers touching and palms facing each other, move apart and turn so, both palms are facing down
Meat		Left hand in ‘5’ Hand shape, palm facing body, right hand flat, palm facing up. Right hand sits between middle and ring fingers and moves back and forth- towards and away from the body
Vegetables		Hand begins in ‘s’ Signing “many” upside Down hand shape, palm facing the body with fingers on the bottom. ‘s’ drops into ‘4’ hand shape, fingers still pointing down with palm facing in

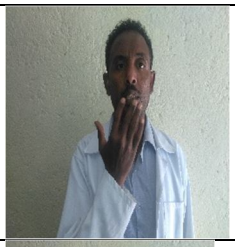
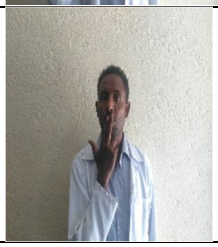
Water		Ethiopian ‘w’ hand shape at chin, palm facing in
Drink		‘C’ hand shape, moves like you are holding a cup and drinking
Injera		‘a’ hand shape with thumb down, moves in a circle
Bread		Both hands in ‘a’ Hand shapes, start next to each other, palms facing out and moves apart while turning so palms are facing each other
Tea		Left hand in ‘o’ Hand shape, palm facing right. Right hand in ‘f’ hand shape with pointer finger/thumb moving like they are stirring a spoon in a cup over the left hand

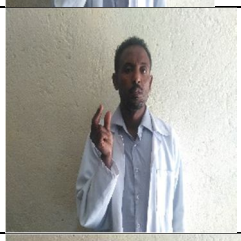
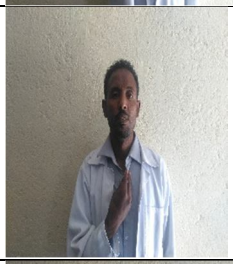
2.2.6 Color

Ethiopian gesture-based communication is communicated by utilizing right hand direction. Notwithstanding the direction, a portion of the tones are addressed by the start of the name

of that tone. For instance, አረንጋዴ is communicated by showing the state of Amharic Alphabet አ and turning the hand double the correct direction. ቢጫ is communicated as አረንጋዴ. The solitary contrast is in ቢጫ beginning by showing Amharic Alphabet በ. ሰማያዊ፣ ግራጫ፣ ወይንጠጅ additionally follow the above method of articulation. A portion of the tones are communicated in a one of a kind way. For models, ቀይ is communicated by putting right hand pointer on the lips and dropping the give over. ነጭ is communicated by putting the correct hand on the neck at that point holding the entire fingers together. ጥቁር is communicated by putting right hand pointer on the temple at that point moving to the correct ways. As recorded above, colors are communicated by various ways.

Table 2-6: colors sign

Color name	Sign	Descriptions
Color		Hand flat, palm facing chin, fingers wiggling-touching lips.
Red		Hand shape touching lips and moving down.
Purple		Ethiopian ‘w’ Hand shape turning back and forth slightly
Pink		Ethiopian ‘r’ Hand shape touching lips and moving down

Blue		Ethiopian ‘s’ Hand shape turning back and forth slightly
Brown		Hand shape touching the side of the nose and moving down
Yellow		Ethiopian ‘b’ Hand shape turning back and forth slightly
Green		‘a’ hand shape turning back and forth slightly
White		Opened ‘5’ Hand shape facing chest brought outward to flattened ‘o’ hand shape
Black		Hand shape, starting above left eye and drawing line across eyebrows, palm facing down

✓ 2.3 Summary

In this chapter, general overview regarding to word representation of Sign language have been presented. World sign language representation especially American sign word representation and South African sign word representation is covered. Singing word in Amharic sign language consists of hands, gesture, orientation and facial expirations. Signing word in Ethiopia is highly related to the real action that the word represented. For example, enjera is represent by showing how to bake enjera. Frequently used sign words, signing family, signing color, signing days and other are explained briefly.

CHAPTER THREE: LITERATURE REVIEW

This chapter contains review of literatures on Ethiopian Sign languages recognition, foreign sign languages and deep learning image recognition researches. From foreign languages reviewed, state-of-the-art approaches are briefly discussed.

✓ 3.1 Local sign language recognition systems

Neguse Kefiyalew [4], propose Amharic Sign Language Recognition based on Amharic Alphabet Signs. Deals with Amharic sign language translation, translate Amharic alphabet sign into their corresponding text. The system has three major components: preprocessing with segmentation, feature extraction and classification. The preprocessing starts with the cropping and extracting of frames. Segmentation is done to segment hand gestures. Thirty-four features extracted from shape, motion and color of hand gestures to represent both the base and derived class of Amharic sign characters. Finally, classification models built using Neural Network and Multi-Class Support Vector Machine.

Frames are extracted automatically from the video using MATLAB built in function. The number of frame is determined by the function depending on the playtime of the video. In this research work, the frame is not less than 50 frames. Image preprocessing is use by different image preprocessing techniques. Based on [4], apply four preprocessing techniques, which are cropping, converting RGB frame to grayscale, contrast adjustment and sharpening.

Segmentation use adaptive threshold algorithm to segment hand sign from the background. The author [4] says, adaptive threshold algorithm helpful for noise image affected by shadow, shading and lighting effect. Beside, apply different morphological operators such as dilation, erosion and refill missed objects of the segment.

Extracted feature used as an input performing classification through training. Based on [4], use three major feature descriptors: such as, shape feature descriptors, motion feature

descriptors and color feature descriptors. Fourier Descriptor (FD) used for describe shapes of Amharic alphabets which extract 31 set of combined shape feature descriptors (fd1, fd2, fd3… and fd31) to represent all the 34 Amharic alphabets

Two classifications is used, neural network (NN) and support vector machine (SVM). The recognition system is capable of recognizing these Amharic alphabet signs with 57.82% and 74.06% by NN and SVM classifiers respectively. Therefore, the classification performance of Multi-Class SVM classifier found to be better than NN classifier.

The second researcher is Legesse Zerubabel [3], by using Ethiopian finger spelling classification system (EFSCS), classify the hand sign of Ethiopia finger spelling into a class that represent each Amharic Alphabet. The system receive sign image, preprocess the sign image and goes to feature extraction, hand detection, segmentation, sign classification and finally associate with corresponding Amharic Alphabet. He recognize only ten basic Amharic Alphabet (ሀ ፡ መ፡ ረ ፡ ሰ ፡ ሸ ፡ በ ፡ ነ ፡ ኘ ፡ አ).

The author [3], principal component analysis (PCA) and Harr-like feature applied for hand detection and sign classification.

The performance measure of [3] begin by classifying the image data based on the result of detector into three grope of true positive, false negative, false positive. True positive belongs to classification into correct target class, false positive belongs to the system classify into hand object but does not have hand object. False negative belongs to the system does not classify the image but the image have hand information. A total of 438 images were collected for 10 Amharic alphabet signs.

For hand detection, two experiments on neural network based hand detector with harr-like and PCA-driven features conducted. In addition, another experiment conducted on boosted classifier based hand detector with harr-like feature. The overall result of experiment is respectively 98.86%, 96.59%, 77.27%.

For sign classification, the first two experiments conducted on neural network based sign classifier by combining with the harr-like and PCA-driven feature. The third experiment conducted on template matching based sign classifier. The overall result of the experiment is respectively 88.08%, 96.22%, 51.44%.

✓ 3.2 FOREIGN GESTURE BASED COMMUNICATION ACKNOWLEDGMENT FRAMEWORKS

In foreign sign language recognition system we try to see the state of art and cover different deep neural network perform under the area of sign language.

3.2.1 Convolution Neural Network (CNN)

Under convolution neural network CNN the review cover two papers [15, 16], deals with 2D and 3D classification. The first paper compare the three channel image (RGB) to Grayscale image for image classification and the second paper focus on 3D classification by using CNN-RNN.

Based on [15], proposed deep learning based sign language recognition system for static signs. The three-channel image frames (RGB) are retrieved from the camera. The dataset hold 35,000 images which include 350 images for each of the static signs. There are 100 distinct sign classes that include 23 alphabets of English, 0–10 digits and 67 commonly used. The dataset consists of static sign images with various sizes, colors and taken under different environmental conditions to assist in the better generalization of the classifier.

The author [15], the model feature extraction and training is based upon convolutional neural networks. The proposed model is trained using the Tesla K80 Graphical Processing Unit (GPU), 12 GB memory, 64 GB Random Access Memory (RAM) and 100 GB Solid State Drive (SSD).

The system [15], results in the highest training and validation accuracy show as follows:

a) Training and validation accuracy on RGB image by different optimizer.

Table 3-1: Training and validation accuracy on RGB image

Model	Training accuracy (%)	Training loss	Validation accuracy (%)	Validation loss	Optimizer
I	99.17	0.0280	98.80	0.0684	Adam
II	99.59	0.0378	98.27	0.1940	RMSProp
III	99.72	0.0126	98.56	0.0759	SGD

a) Training and validation accuracy on grayscale image by different optimizer.

Table 3-2: Training and validation accuracy on grayscale image

Model	Training accuracy (%)	Training loss	Validation accuracy (%)	Validation loss	Optimizer
I	99.24	0.0280	98.85	0.0684	Adam
II	99.76	0.0378	98.35	0.1940	RMSProp
III	99.90	0.0126	98.70	0.0759	SGD

Real time sign language gesture recognition from video sequences is proposed by [16]. Video sequences contain both the temporal as well as the spatial features.

In [16], used two different models to train both the temporal as well as the spatial features. Used model on the spatial features of the video sequences is Inception model which is a deep CNN (convolutional neural net). CNN was trained on the frames obtained from the video sequences of train data. Use RNN (recurrent neural network) to train the model on the temporal features. Trained CNN model was used to make predictions for individual frames to obtain a sequence of predictions or pool layer outputs for each video. Now this sequence of prediction or pool layer outputs was given to RNN to train on the temporal features. The data set used consists of Argentinian Sign Language (LSA) Gestures, with around 2300 videos belonging to 46 gestures categories. Using the predictions by CNN as input for RNN 93.3% accuracy was obtained and by using pool layer output as input for RNN an accuracy of 95.217% was obtained.

3.2.2 VGG

Under convolution neural network VGG we look two papers [17, 18] that deal with 3D sign image classification by using different models and state of art regarding with 3D handling and classification.

Based on the new large-scale dataset [77], are able to experiment several deep learning methods for word-level sign recognition and evaluate their performances in large scale scenarios. Specifically in this paper, implement and compare two different models. Holistic visual appearance based approach, and 2D human pose based approach.

The models used, VGG-GRU, Pose-GRU, Pose-TGCN and I3D are implemented in PyTorch. It is important to notice that [17] use the I3D pre-train weights. [17], train all the models with Adam optimizer. However, I3D does not converge when using SGD to fine-tune it in the experiments. Thus, Adam is employed to fine-tune I3D. All the models are trained with 200 epochs on each subset. [17], terminate the training process when the validation accuracy stop increasing.

Preprocessing is started by resize the resolution of all original video frames such that the diagonal size of the person bounding-box is 256 pixels. Randomly crop a 224×224 patch from an input frame and apply a horizontal flipping with a probability of 0.5.

Our results [17], show that pose-based and appearance-based models achieve comparable performances up to 62.63% at top-10 accuracy on 2,000 words/glosses, demonstrating the validity and challenges of the dataset.

The second paper [18], presents a fusion-based ensemble of VGG networks for the Multimodal Emotion Recognition Challenge 2017. Image fusion is used to aggregate consecutive frames from video sequences for the representation of temporal information.

The author [18], an ensemble of four VGG Face models which have been fine-tuned on the MEC dataset, is utilized to extract facial expression features from the fused images. The VGG Face-Bi-LSTM and VGG Face-Bi-GRU are also implemented for comparison.

For data preprocess and fusion, face detector of OpenCV is applied for initialization of face tracking. Interface and MTCNN is utilized to obtain face landmarks.

The accuracies of the fine-tuned VGG Face-ensemble, VGG Face-Bi-LSTM and VGG Face-Bi-GRU on validation data are 51.06%, 43.95% and 44.92% respectively, indicating the effectiveness of our method.

3.2.3 ResNet

Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D Residual Networks was proposed by [19]. The dynamic sign language recognition could be achieved by intelligent algorithm with analyzing video sequence features and classifying hand gestures.

Video sequence features extraction module, which performs the task of long-term spatiotemporal features extraction with inputting segmented video frames. In this step, the networks are trained on chunks of full-time video, so the spatial context of the action performed can be preserved favorably. The feature vectors could be obtained by training full-time segmented videos on B3D ResNet model. Each video feature vector will be provided to the third part for analyzing dynamic information of sign language. Eventually, to create a joint representation for these independent streams.

In the author [19], the third part is dynamic sign language recognition module, which can analyze long-term temporal dynamics and predict the hand gesture label. Through analyzing each video feature vectors, the frames label could be predicted. Thus, the video sequence label could be predicted. According to the top label prediction scores, this label will be regarded as the label of video sequence and be outputted as the recognition result. Therefore, the dynamic sign language could be recognized effectively.

The proposed model [19], could effectively recognize different hand gestures with extracting video spatiotemporal features and analyzing features sequence. With our model, it could get a good performance on complex or similar sign language recognition. The results show that the proposed method can obtain the state-of-the-art recognition accuracy 89.8%.

✓ 3.3 Summary

In this unit, the specialist attempted to cover writing survey on communication through signing acknowledgment arrangement of local just as foreign languages. Nearby gesture-based communication examined by two analysts are summered as follows.

Table 3-3: Local recognition systems

Name	Preprocessing	Feature extraction	Classification	Result
✓ **Legesse Zerubbabel [2008]**	✓ cropping, ✓ RGB to Grayscale, ✓ contrast adjustment, ✓ sharpness	✓ Harr-like ✓ PCA-driven ✓ Template matching based.	✓ NN	✓ 88.08% (Harr-like) ✓ 96.22% (PCA-driven) ✓ 51.44% (Template matching)
✓ **Neguse Kefiyalew [2018]**	✓ Cropping. ✓ RGB to Grayscale. ✓ contrast adjustment ✓ sharpness ✓ Segmentation	✓ Select from the shape of the frames and direction of	✓ NN ✓ SVM	✓ 57.82% (by NN) ✓ 74.06% (by SVM)

		motion of the frames.		

Foreign video base profound learning recognition are talked about by various specialists' outlines as follows

Table 3-4: foreign recognition system

Name	Preprocessing	Feature extraction	Classification	Result
The author [15]	Cropping and Normalization.	CNN	CNN	By using SGD optimizer for RGB and Grayscale training accuracy 99.72% and 99.90 %respectively.
The author [16]	Extract frame and background removal	Inception CNN	LSTM (RNN)	Accuracy of 95.217%
The author [18]	OpenCV, MTCNN	VGG	Ensemble Bi-LSTM GRU	51.06%, 43.95% and 44.92%
The author [17]	Cropping Flipping	VGG	VGG-GRU Pose-GRU Pose-TGCN I3D	62.63%
The author [19]	Cropping Normalization	3D ResNet	BLSTM	89.8%

CHAPTER FOUR: METHODOLOGY

This chapters explains the methodology followed in this research.

- ✓ **4.1 Introduction**

In chapter 2, we observe how to sign words and notice the important features of signing words. This chapter targeted to answer handling those important features and show how to recognize sign words. We cover, the general architecture of the study, preparation of dataset, preprocessing steps, feature extraction algorithms and classification mechanisms in detail.

This chapter organize as follows. Section 4.1 Introduction, section 4.2. preparing frequently used sign words, section 4.3 Assign signers and record the sign words, section 4.4 Video to frame conversion, section 4.5 Preprocessing steps of the frames, section 4.6 Pixel based image recognition algorithm called convolution neural network (CNN), section 4.7 Vanishing gradient and Degradation problem, section 4.8 Residual network, section 4.9 Feature extraction by ResNet-34, section 4.10 3D Data Classification through training by ResNet-34 and finally section 4.11 How to testing the system.

The General Architectural View of a new system

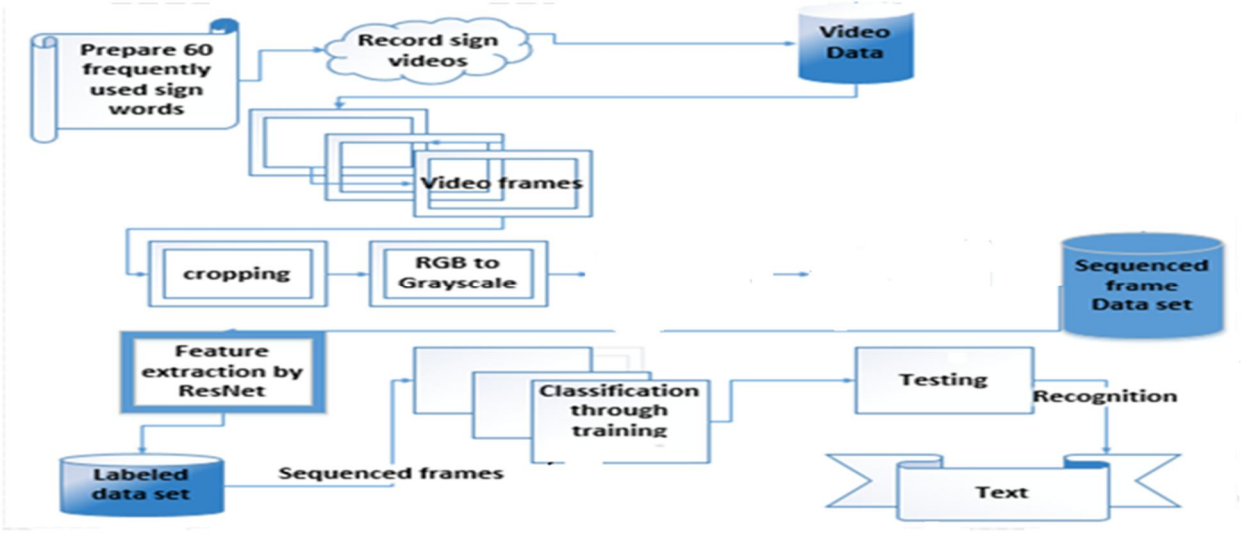


Figure 4-1: The general overview of Amharic sign word recognition system

Before going through the detail, the general directional flow diagram of word level 3D sign to Amharic text recognition system is represented by the above general architecture. As

shown from the above figure, the input is video and the target is text. The initial task is observing the signers how to spell signs carefully. Based on the observation we recognize that, most of sign words are expressed by motion and orientation. The above architecture is sequentially viewed as, choose frequently used 60 sign words, record the selected sign words, video to frame conversion, preprocessing frames, feature extraction by using deep neural network ResNet-34, Classification through training by using ResNet-34. Finally, validating the system by giving input video and expecting the desire text output.

✓ 4.2 Preparing frequently used sign words

For easily communicate with signers, learning sign words are very essential. Most of Amharic sign words are expressed by hand shape, orientation, gesture, and facial expression. However, some of sign words like work (ስራ) express by using two hands. Based on [20], learning sign language follow different steps, each steps hold categories of sign words.

Almost all Amharic words are signed by Deaf community. It is difficult to know the exact number of sign words. For this thesis work, we prepare 60 frequently used sign words. These words are categorized into six classes. **ሰላምታአስጣጥ** , **ጥያቄምልክት, ቤተሠብ, ፆታ, ስሜትገላጭቃላቶች**, and **የተለመዱ ቃላት.** This words are presented in the below table.

Table 4-1: Frequently used sign words

ሰላምታአስጣጥ		**ጥያቄምልክት**		**ቤተሠብ**		**ለፆታ**		**ስሜትገላጭቃላት**		**የተለመዱ ቃላት**	
ተ/ቁ	ቃላት	ተ/ቁ	ቃላት	ተ/ቁ	ቃላት	ተ/ቁ	ቃላት	ተ/ቁ	ቃላት	ተ/ቁ	ቃላት
1	ሰላም	16	ምን	25	አባት	38	እኔ	45	መደሰት	50	ከነገወድያ
2	ስም	17	ለምን	26	እናት	39	አንተ	46	ማዘን	51	ነገ
3	ማን	18	መቼ	27	ወንድም	40	አንቺ	47	ሞኖደድ	52	ትናንት
4	ነው	19	የት	28	እህት	41	እስዋ	48	መሳቅ	53	እሺ
5	እግዚአብ	20	የትኛው	29	አጎት	42	እሱ	49	መሳም	54	ቀን

	ሄ ር										
6	መል ካ ም	21	ስ ን ት	30	ሚስ ት	43	እ ኛ			55	ሳ ምን ት
7	ምሽ ት	22	እ ያ ን ዳ ን ዱ	31	አ ክ ስ ት	44	እ ነ ሱ			56	ወ ር
8	ጉ ዞ	23	ምን ም	32	ዘ መድ					57	ዓ መት
9	ብ ል ል ኝ	24	ሌ ላ	33	ል ጅ					58	ብ ር
10	አ መሰ ና ለ ሁ			34	ወ ን ድ አ ያ ት					59	ጎ በ ዝ
11	ን ጋ ት			35	ሴ ት አ ያ ት					60	ሰ ነ ፍ
12	አ ዎ			36	ባ ል						
13	አ ይ ደ ለ ም			37	እ ጮ						
14	ቤ ተ ሰ ብ										
15	ይ ቅ ር ታ										

Amharic derived alphabet Dataset

On the other hand, work on the derived Amharic alphabet sign for the comparison of the proposed model with the previous research done by [5].

Table 4-2: Three Amharic sign derived alphabet

No	Sign Words	No	Sign Words
1	ሁ	10	ሌ
2	ሂ	11	ል
3	ሃ	12	ሎ
4	ሄ	13	ሑ
5	ህ	14	ሒ
6	ሆ	15	ሓ
7	ሉ	16	ሔ
8	ሊ	17	ሕ
9	ላ	18	ሖ

- ✓ **4.3 Assign signers and record sign words**

After preparing the target sign words in section 4.2, the next step is observing how to sign the selected sign words by different signers. This action help to understand the basic feature of signs like motion feature, gesture representation, facial expression and direction feature of each sign words. Then prepare tool for recording sign words. Samsung mobile devise is used for recoding sign words. Each sign words are recording one by one. And set play time not more than 4s. Individual signer sign all 6o sign words. Totally, we collect 2400 video for 40 different signers.

The procedures used to collect the sign video for this work are listed below:

- Finding people that are willing to participate in the data collection.
- Giving 2:00 hr. training of signing Ethiopian Amharic words for the participants before data collection, if they don't have experience of signing before.
- By considering the participant, select signs that are frequently used and easy to sign.
- Collecting the data with Samsung mobile and the camera fixed in front of the signer.
- And record the sign that has not more than 4 seconds.

✓ 4.4 Video to frame conversion

A video stream is composed of many frames at a frame rate of at least 25 frames per second (fps) that a human cannot perceive any discontinuity in the video. The key frame extraction in video summarization is intended to eliminating replication and extraction of key frames from the video. Resent key frame extraction techniques like clustering, shot, visual content-based key frame extraction are available [27]. For this thesis work, we simply use the first 50^{th} consecutive frames. The mechanism is described as follows by flow chart.

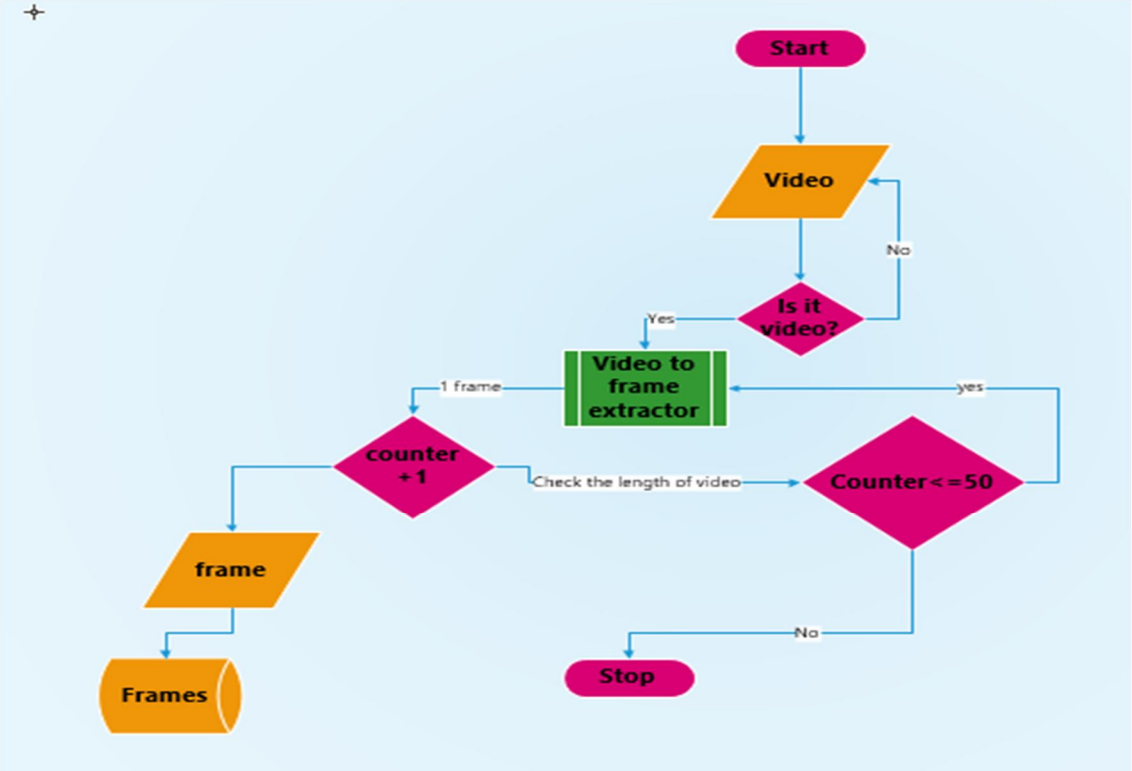


Figure 4-2: Flow chart for extract frame from video

4.5 Preprocess the frames

4.5.1 Cropping

The frame, which are extracted from the original video, have huge size and has unnecessary frame components. The proposed model can take the input image having high, width as 224 x224. The frame is cropped at the end of video to frame conversion and check it in before giving to the proposed model.

```
for imagePath in batchPaths:
    # load the input image using the Keras helper utility
    # while ensuring the image is resized to 224x224 pixels
    image = load_img(imagePath, target_size=(224, 224))
    image = img_to_array(image)
```

Figure 4-3: Frame cropping algorithm

4.5.2 Convert RGB to Grayscale

On the other hand, grayscale image represent by 8 bits, whereas, RGB image represent by 24 bit. Because of taking high computation and more process time by RGB, RGB to grayscale conversion is necessary.

The space-variant luminance map-based image enhancement method (SVLM) described in [21], is used for RGB to grayscale conversion, which uses the intensity component as an input to the method. Here the intensity is calculated with different weights.

Grayscale= ((0.299 * R) + (0.587 *G) + (0.114 *B))

According to the above equation, Red (R) has contribute 30%, Green (G) has contributed 59%, which is greater in all three colors, and Blue (B) has contributed 11%.

Grayscale image



Figure 4-4: Grayscale converted frame

4.6 Pixel based image recognition algorithm

Object-oriented classification uses both spectral and spatial information for classification. While pixel based classification is based solely on the spectral information in each pixel, object-based classification is based on information from a set of similar pixels called objects or image objects.

Spectral imaging refers to a group of analytical techniques that collect spectroscopic information and imaging information at the same time. The spectroscopic information tells

us about the chemical makeup at the individual points of the image (pixels) allowing a chemical map of the imaged area to be produced.

Convolution Neural Network (CNN) is one of a deep neural network, which is powerful in the field of image recognition and classification. CNN's tend to perform better than other image and video recognition algorithms in fields of image classification, medical image analysis and natural language processing.

There are four key operations in the CNN [28]. These operations are the fundamental blocks of every CNN.

1. Convolution
2. Non-Linearity (ReLU)
3. Pooling or Sub Sampling
4. Classification (Fully Connected Layer)

Convolution

Convolution viewed as, extract features from the image by [22], video or frame sequential data is the main concern of convolution. Convolution works by finding out the spatial relationships between pixels by learning image features using small region of interest. A convolution operation is an element wise matrix multiplication operation [23]. Where one of the matrices is the image, and the other is the filter or kernel that turns the image into something else. The output of this is the final convoluted image.

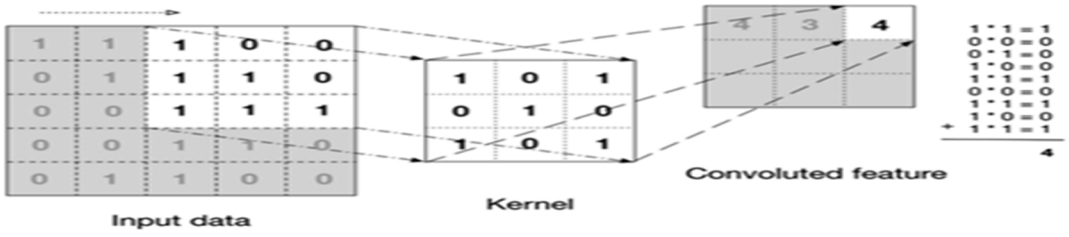


Figure 4-5: Convolution operation (CNN)

There are multiple convolutional filters available for us to use in Convolutional Neural Networks (CNNs) to extract features from images. Filter is a 3x3 matrix in CNN and the matrix formed by sliding the filter over the image is called convolved feature or feature map.

In image processing, there are a set few filters that are used to perform a few tasks. Such as blur, sharpen and edge detection.

Table 4-3: Some filters used for convolution

Operation	Filter
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$
Gaussian blur (approximation)	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$

All these are achieved just by changing the numeric values of the filter matrix before the convolution operation [22], this means that different filters achieve different results depending on what the end goal of the model is.

ReLU

ReLU stands for Rectified Linear Unit and is a non-linear operation [22]. Non-Linearity operation is perform after competition of convolution operation. It is applied to each element individually and it replaces all negative pixel values in the feature map to zero. The purpose of the ReLU is to introduce non-linearity [23] since real world training is non-linear and the CNN should model to that.

The Non-Linearity activity is utilized after the convolution activity referenced previously.

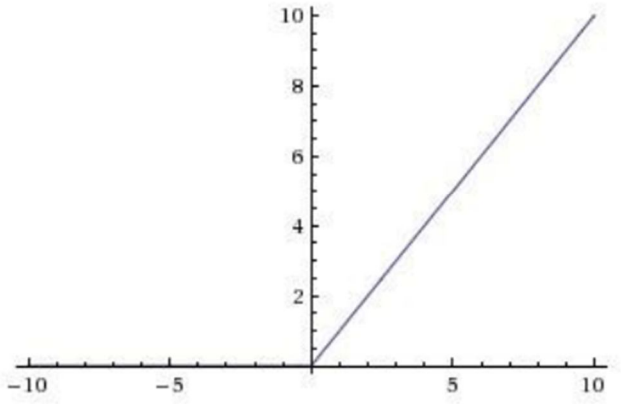


Figure 4-6: Non-Linearity RLU

ReLU function works by giving a max between an input number and zero.

$$g(x) = \max(0, x)$$

ReLU is likewise a computationally modest initiation work dissimilar to other actuation capacities like sigmoid and tanh in light of the fact that it requires less difficult numerical activities.

The Pooling Step

Pooling or subsampling is a layer, which reduces dimensionality of feature maps generated from the convolutional layer while retaining the most important information [22]. Pooling can be of several different types such as: Max pooling, Average pooling, Sum pooling etc.

Max Pooling is when we define a window of a certain size and take the largest element from it. Instead of taking the largest element, we could also take an average (Average Pooling) or sum all the elements in it (Sum Pooling). We continue to move the filter over the entire

image like the stride we took in convolution till we have a pooled layer of the specified type in the architecture.

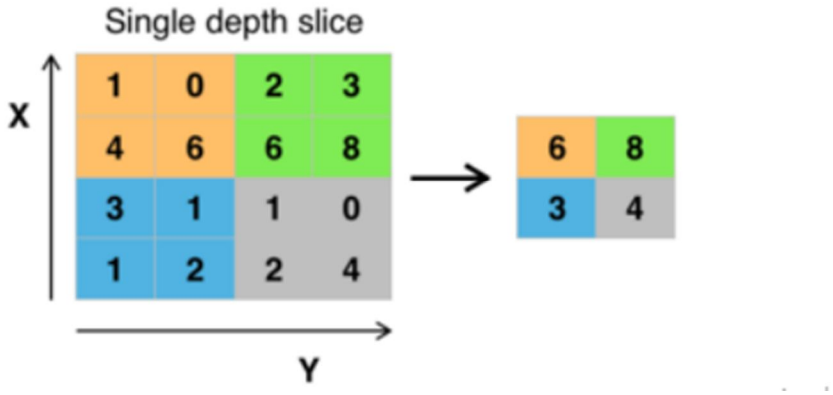


Figure 4-7: Max pooling operating (CNN)

The pooling layer further lessens the dimensionality [24] of the information picture and thusly diminishes the quantity of boundaries and calculations in the organization. It gives us a portrayal of the information picture in a cleaner more compact structure.

Fully Connected Layer

The fully connected layer is a multilayer perceptron, which uses activation functions such as SoftMax in the output layer. There are several activation functions like SoftMax, but we shall only discuss SoftMax for the purposes of this thesis. The term fully connected layer implies that every neuron in the layer is connected to every neuron in the previous layer. The convolutional layer along with the pulling layer generates a summarization of the original input image, which is fed into the fully connected layer. The fully connected layer send gives an output, which can be either classification or regression.

The fully connected layer allows for operation such as backpropagation, which are key features, which enable a neural network to perform classification with a high accuracy as it does. The SoftMax layer uses a SoftMax function to squash a vector between zero and one and it is the most used activation function in classification.

Based on [23], the most common deep learning architectures for CNN today are: VGG, ResNet, Inception, and Xception. For this thesis work we use ResNet-34.

4.7 Vanishing gradient and Degradation problem

In theory, Recurrent Neural Networks (RNN) are absolutely capable of handling long term dependencies [25]. The gradient expresses the change in all weights with regard to the change in error. Since the layers and time steps of deep neural networks relate to each other through multiplication, gradient is susceptible to vanishing or exploding.

4.7.1 Vanishing Gradient

The gradients of the network's output with respect to the parameters in the early layers become extremely small [25]. In other words even a large change in the value of parameters for the early layers doesn't have a big effect on the output. Hence the network can’t learn the parameter effectively.

This happens because the activation functions (sigmoid or tanh) squash their input into a very small output range in a very nonlinear fashion. For example, sigmoid maps the real number line onto a "small" range of [0, 1]. As a result, there are large regions of the input space which are mapped to an extremely small range. In these regions of the input space, even a large change in the input will produce a small change in the output - hence the gradient is small.

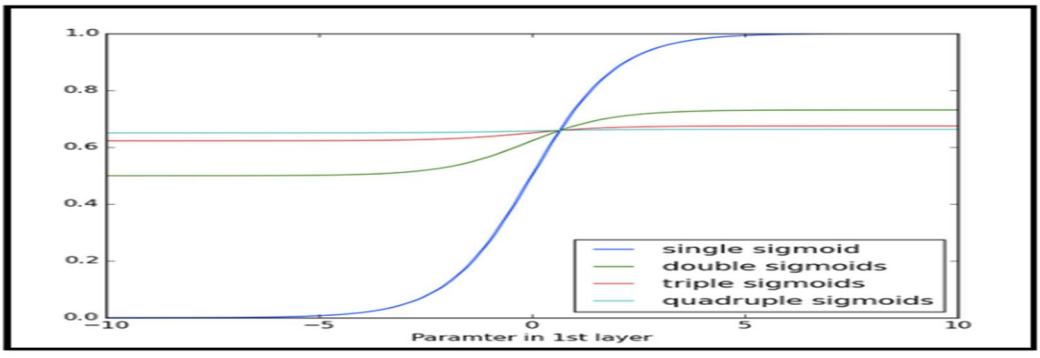


Figure 4-8: Vanishing Gradient

This becomes much worse when we stack multiple layers of such non-linearities on top of each other. For instance, first layer will map a large input region to a smaller output region,

which will be mapped to an even smaller region by the second layer, which will be mapped to an even smaller region by the third layer and so on. As a result, even a large change in the parameters of the first layer doesn't change the output much.

4.7.2 Degradation problem

The degradation problem has been observed while training deep neural networks. As we increase network depth, accuracy gets saturated.

It is evident that all the CNN architectures are susceptible to image degradations. It is interesting to observe that some shallower models like VGGs that achieve less accuracy in many classification tasks are more resilient to degradations [25].

4.7.3 Residual network

Deep residual network is almost similar to the CNN networks, which have convolution, pooling, activation, and fully connected layers stacked one over the other. The only connection to the simple network to make it a residual network is the identity connection between the layers [26]. The below figure show the residual block used in the network. You can see the identity connection as the curved arrow originated from the input and sinking to the end of the residual block.

During training stage the residual network alters the weights until the output is equivalent to the identity function. In turn the identity function helps in building a deeper network. The residual function then maps the identity, weights and biases to fit the actual value.

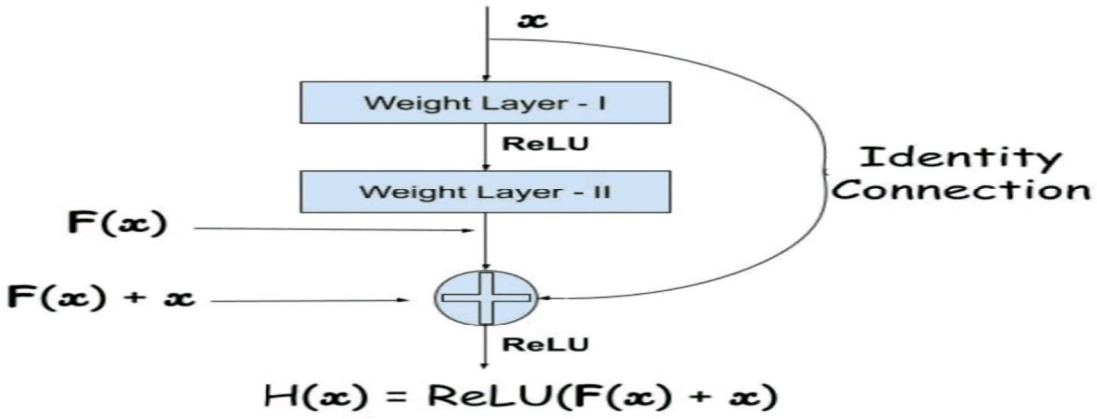


Figure 4-9: Identity in Residual Network

4.8 Feature extraction by ResNet-34

For feature extraction we use ResNet-34 model. ResNet-34 model is a family of Residual connection network (ResNet). ResNet models are defer one another by their number of layers. ResNet-34 has 34 layers. This model receive one video at a time. Like original CNN it works on pixel based convolution operation with skip connection.

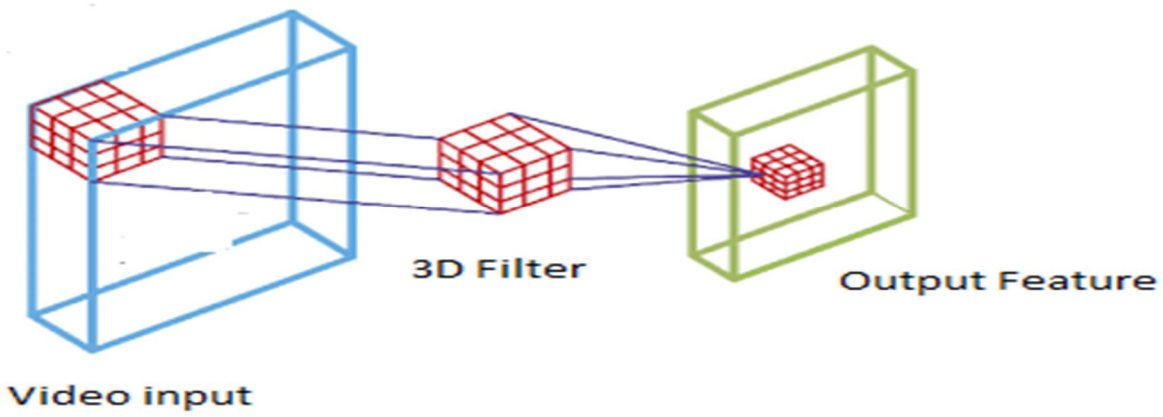


Figure 4-10: Feature extraction by ResNet-34

The input for model is video. Perform 3D filter operation and max pooling (filter operation and max pooling was discussed in the above section), by ResNet-34 finally, generate the output feature stored in the form of .csv file. This is the input for the classification model.

4.9 3D Data Classification through training by ResNet-34

For classification purpose we use the modified form of ResNet-34. Our approach takes as input a 3D. 3D ResNet-34 is unique by the number of dimensions of convolutional kernels and pooling. Our 3D ResNet-34 perform 3D convolution and 3D pooling. The sizes of convolutional kernels are 3 × 3 × 3.

3D ResNet-34 receive feature extracted output sequential frames at a time and mapped in to one sign word. This process perform again and again for the rest of 59 sign words. Train the network by supporting each process by 40 videos. The proposed architecture with shallow network is shown as follows.

Layer Name	Architecture	
	18-layer	34-layer
conv1	$7 \times 7 \times 7, 64$, stride 1 (T), 2 (XY)	
conv2_x	$3 \times 3 \times 3$ max pool, stride 2 $\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 4$
conv4_x	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 6$
conv5_x	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 3$
	average pool, 400-d fc, softmax	

Figure 4-11: ResNet-34 modified layers for classification

4.10 How to test the system.

The model evaluation is done by the combination of feature extracted (.csv) evaluation dataset and model.predict () method. A unique video dataset is prepared as the same as training and validation dataset. Extract the feature of training, validation and evaluation dataset before the model is train. Then after, this feature extracted evaluation dataset is fit into model.predict () method. Model.predict () method predict the class of each input.

✓ 4.11 Summary

This part logically experienced the planning of word level Amharic communication via gestures recognition system. The proposed framework has various parts for cooperating deliberately. From the outset, plan regularly utilized sign words and see how to sign them.

At that point after, video acquiring happens. The main sequential 50 frames were changed over from every video. Feature extraction was followed. This feature extraction was finished by utilizing incredible profound neural network ResNet-34 and saved it the form of .csv format. At last, ResNet-34 was trained using these features from which 60 output classes are generated to classify or recognize the selected Amharic Sign language words.

CHAPTER FIVE: EXPERIMENTATION AND RESULT DISCUSSION

- ✓ **5.1 Introduction**

In the past part, we portrayed the general plan of the framework. This part present about execution detail of the proposed acknowledgment framework. Segment 5.2 presents about association of datasets and directory structure for the information. Segment 5.3 clarifies about the experimental arrangement. Segment 5.4 present about test situation. Segment 5.5 clarifies about outcome in detail. Segment 5.6 disclose about threat to validity. At long last, the outline of this section is introduced.

In particular, it is aimed to answer the following research questions.

- ✓ **Develop a word sign recognition system and test the impact on communication with deaf. 5.2 Dataset preparation**

The video dataset utilized for this postulation work comprises of Amharic Sign words, Amharic sign inferred letters, and Argentinian sign word. Around 2400 Amharic sign words are utilized from 60 distinctive classes. We utilize the inferred letters of ሀ ለ ሐ around 450 recordings was utilized. Around 3000 sign videos were use for 64 diverse classes of Argentinian communications via gestures.

Table 5-1: Dataset organization

With manual	**Automatic**
✓ **Video acquisition by Samsung mobile** ✓ **Move the data to the specified directory.**	✓ **Key frame extraction.** ✓ **Cropping** ✓ **RGB to Grayscale conversion** ✓ **Split the frames for training, validation and for testing.**

During the time spent information assortment, great deals of difficulties have been happened. The principal challenge was discovering signers that can sign Amharic word appropriately. It is elusive adequate number of effective endorsers who signs Amharic

words. To address this test, we take a proportion of gathering willing individuals to take a preparation for marking the chose sign words and inferred Amharic letters. After brief timeframe preparing make a camera shot for them.

The second challenge was happened during recording, the prepared signer signing 78 words (60 sign words and 18 inferred letters) is a repetitive task. This tops to set aside a loge effort to finish video procurement.

The third challenge was discovering materials and books about Ethiopian gesture-based communication. We discovered some broad rule for Ethiopian gesture-based communication yet it is elusive distributed clarification about Ethiopian sign words. We tackle this issue by finding a decent signer/Deaf individual who give as a preparation.

Amharic word datasets are available here:

https://drive.google.com/drive/folders/1vF5wSRWg5cT0iiFOG3jF6HItFGEz08mO

Table 5-2: Amharic Sign Words dataset

No	Sign Words	Number of Sample	No	Sign Words	Number of Sample
1	ሠላም	40	31	ይቅርታ	40
2	ስም	40	32	ዘመድ	40
3	ማን	40	33	ልጅ	40
4	ነው	40	34	ወንድአያት	40
5	እግዚአብሄር	40	35	ሴትአያት	40
6	መልካም	40	36	ባል	40
7	ምሽት	40	37	እጮኛ	40
8	ጉዞ	40	38	እኔ	40
9	በልልኝ	40	39	አንተ	40
10	አመስግናለሁ	40	40	አንቺ	40
11	ጎጋት	40	41	እስዋ	40

12	አዎ	40	42	እሱ	40
13	አይደለም	40	43	እኛ	40
14	ቤተሰብ	40	44	እነሱ	40
15	ምን	40	45	መደስት	40
16	ለምን	40	46	ማዘን	40
17	መቼ	40	47	መናደድ	40
18	የት	40	48	መሳቅ	40
19	የትኛው	40	49	መሳም	40
20	ስንት	40	50	ከነገ ውዲያ	40
21	እእያንደንዱ	40	51	ነግ	40
22	ምንም	40	52	ትናንት	40
23	ሌላ	40	53	እሺ	40
24	አባት	40	54	ቀን	40
25	እናት	40	55	ሰምንት	40
26	ወንደም	40	56	ወር	40
27	እህት	40	57	ዓመት	40
28	አጎት	40	58	ከትናንትበስቲያ	40
29	ሰነፍ	40	59	ብር	40
30	አክስት	40	60	ጎበዝ	40
				Total	2400

For the comparison of the machine learning algorithm with deep learning algorithm, we utilize the Amharic inferred letters of ሀ [ሁ፤ሂ፤ሃ፤ሄ፤ህ፤ሆ], the inferred of ለ [ሉ፤ሊ፤ላ፤ሌ፤ል፤ሎ] and the inferred of ሐ [ሑ, ሒ, ሓ, ሔ, ሕ and ሖ] which is utilized by the past analyst [4] work on machine learning algorithms SVM and NN.

Table 5-3: Amharic derived latter's dataset

No	Sign Words	Number of Sample
1	ሁ	40
2	ሂ	40
3	ሃ	40
4	ሄ	40
5	ህ	40
6	ሆ	40
7	ሉ	40
8	ሊ	40
9	ላ	40
10	ሌ	40
11	ል	40
12	ሎ	40
13	ሑ	40
14	ሒ	40
15	ሓ	40
16	ሔ	40
17	ሕ	40
18	ሖ	40
	Total	720

Table 5-4: Argentinian sing word dataset

No	Sign Words	No	Sign Words	No	Sign Words
1	Son	19	Catch	37	To-land
2	Food	20	Name	38	Yellow

3	Trap	21	Yogurt	39	Give
4	Accept	22	Man	40	Away
5	Opaque	23	Drawer	41	Copy
6	Water	24	Bathe	42	Skimmer
7	Colors	25	Country	43	Sweet-Milk
8	Perfume	26	Red	44	Chewing gum
9	Born	27	Call	45	Photo
10	Help	28	Run	46	Thanks
11	None	29	Bitter	41	
12	Deaf	30	Map	42	
13	Enemy	31	Milk	43	
14	Dance	32	Uruguay	44	
15	Green	33	Barbeque	45	
16	Coin	34	Spaghetti		
17	Where	35	Patience		
18	Breakfast	36	Rice		

5.3 The directory structure of the data set

Create two folders with **train videos** and **test_videos** in the project root directory. It should contain folders corresponding to each category, each folder containing corresponding videos. Examples of training data structure as shown below.

train_videos

├── አመሰናለሁ

│ ├── 0001.mp4

│ ├── 0002.mp4

│ ├── 0003.mp4

│ └── 0004.mp4

├── ሰላም

│ ├── 0001.mp4

│ ├── 0002.mp4

│ ├── 0003.mp4

```
│   └── 0004.mp4
├── ሌላ
│   ├── 0001.mp4
│   ├── 0002.mp4
│   ├── 0003.mp4
│   └── 0004.mp4
└── ጉዞ
    ├── 0001.mp4
    ├── 0002.mp4
    ├── 0003.mp4
    └── 0004.mp4
```

✓ 5.4 Experimental Setup

Table 5-5: Experimental setup

Manufacturer	Hp, core i5
Memory	8.0GB RAM, 700 GB hard disk
Processor	speed of 2.7 GHz
Operating system	Windows 10 64 bit

✓ 5.5 Experimental Senarios

To analyses the hypothesis of this examination, four exploratory situations were finished. The first is, to check if the framework can effectively perceive Amharic signs word and make an interpretation of it to the text form. For this situation, two preprocessing steps are finished. RGB to Grayscale transformation and cropping. RGB to Grayscale transformation is required in view of reducing high calculation of Red, Green and Blue. Crop the frame 224 x 224 on the grounds that the picked feature extraction algorithm receives the input shape of 224 X 224. ResNet is an incredible algorithm for handling picture. So, the favored feature extractor and classifier for perceiving word level Amharic sign is ResNet-34 (talked about in the past section). At long last, video base sign words are tried subsequent to preparing.

The second experimental scenario is, to test the impacts of automatic feature extractor. For this, we utilize three Amharic derived letters of ሀ፡ ለ፤ ሐ, an aggregate of 18 character. We

utilize the equivalent preprocessing, feature extraction and classification ventures as the main situation. What's more, contrast the performance with the past researcher [4] done on them by manual feature extraction.

The third test situation is, centered on the language test for this examination. For this we use Argentinean Sign language motion video dataset. First set up the dataset for extricating frames, and afterward by utilizing the proposed feature extractor and classification algorithm to arrange Argentinean sign words. At last contrast the outcome with experiment one result.

The fourth experimental scenario is, work on other CNN-RNN model for comparison. By using Amharic word datasets, the feature of each frame is extracted by convolution neural network (CNN) and classification is done by (Long and short term memory) LSTM. Finally, compare the classification result LSTM with the proposed algorithm ResNet-34.

In our work, we have utilized four execution matrixes to examine the characterization execution of ResNet-34. These exhibition matrixes are: Accuracy, Precision, Recall and F-score.

Accuracy is the absolute number of accurately anticipated signs to all test tests.

Equation 5-1: Calculate the accuracy of the model

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Precession is the proportion of accurately recognized occasions of a class to the all-out number of positive perceptions distinguished a sign as that class.

Equation 5-2: Calculate the precession of the model

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall is the proportion of accurately recognized occasions of a class to all the examples where the sign was genuinely that class.

Equation 5-3: Calculate the recall of the model

$$\text{Recall} = \frac{tp}{tp + fn}$$

F-score is the consonant mean of exactness and review [19].

Equation 5-4: Calculate F1 score of the model

$$F_{Score} = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

In these conditions, TP, TN, FN, FP are individually true positive, true negative, false negative, false positive. A genuine positive happens when the anticipated class is equivalent to the real class. A false positive happens when a classifier arranges sign character as the off-base class. A true negative is the point at which the classifier accurately predicts sign character not piece of a wrong class and a false negative happens when the classifier doesn't order sign character into the right class [19].

- ✓ **5.5 Result**

Experiment 1: was done to answer the above research question. This experiment aims to see the desired system well recognize Amharic sign word.

The dataset is divided into three sections, training, validation and evaluation. ResNet-34 is utilized as feature extractor and classifier. For assessment purposes, the models acquired from preparing stage are tested by utilizing a unique Amharic sign words dataset. In this manner we figure the precision, accuracy, recall and F-score.

ResNet-34 model is a powerful and preferable model for this research work. The training accuracy is presented by using the following algorithms

```
import matplotlib.pyplot as plt
plt.plot(History.history['val_accuracy'])
plt.title('ResNet-34 Model training accuracy')
plt.ylabel('accuracy')
plt.xlabel('epochs')
plt.legend(['train'],loc='upper left')
plt.show()
```

Figure 5-1: Training accuracy algorithms

After compilation of the training, we try to saw the training accuracy by giving two parameters which are val_accuracy and epoch. The below graph presents the training accuracy of each and every epoch.

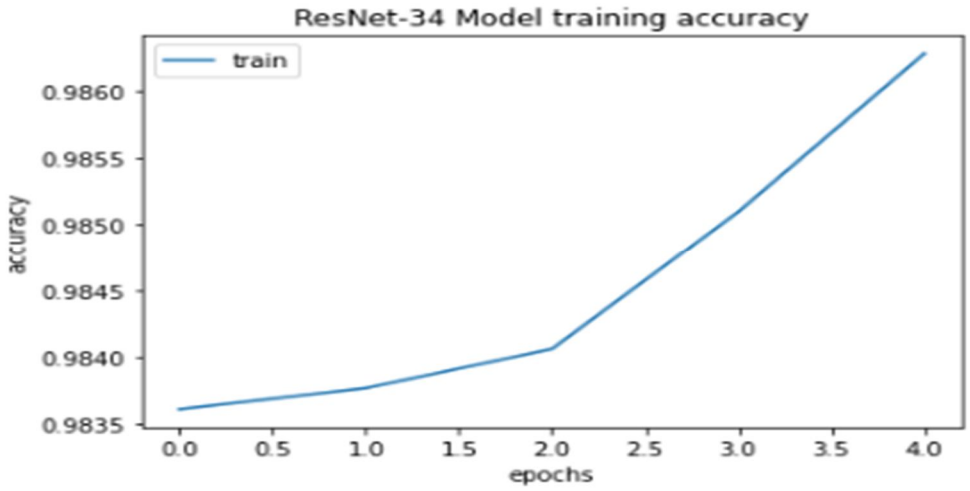


Figure 5-2: Training accuracy curve

The training accuracy curve as you can see in the above figure 5.2 is upward and score above 98% for epoch 4. The proposed model trains the sequence of frames very well. When the epoch 25 the training accuracy is 100%. The proposed model loss is presented as follows:

```
import matplotlib.pyplot as plt
plt.plot(History.history['loss'])
plt.title('ResNet-34 Model training loss')
plt.ylabel('accuracy')
plt.xlabel('epochs')
plt.legend(['train'],loc='upper left')
plt.show()
```

Figure 5-3: Training loss algorithm

A little bit modified algorithm for extracting training loos curve by using the above algorithms. The algorithm receives two parameters those are loss and epoch. The output of the above algorithm is presented by using matplot library as shown below.

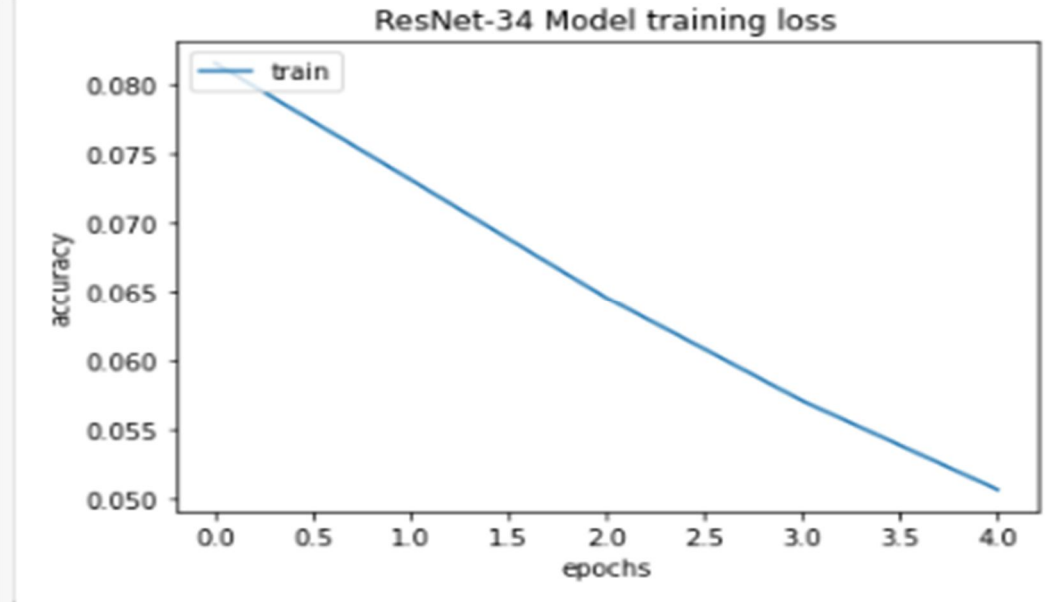


Figure 5-4: Training loss curve

The proposed model training loss curve is clearly downward. Whenever the training increase, the loss is going decrease. The training loss is very low under 0.06 within 25 epochs. This research work implements with 25 epochs.

For the better understanding of the model performance, we observe the training and test accuracy. We must saw the two curves (training and test curve). These two curves are presented by using the following algorithm:

```
import matplotlib.pyplot as plt
plt.plot(H.history['accuracy'])
plt.plot(H.history['val_accuracy'])
plt.title('ResNet-34 Model training accuracy')
plt.ylabel('accuracy')
plt.xlabel('epochs')
plt.legend(['train','test'],loc='upper left')
plt.show()
```

Figure 5-5: Training and validation curve algorithm

The above algorithm receives three parameters (test accuracy, val_accurecy and epoch). The output for the above algorithm is shown below.

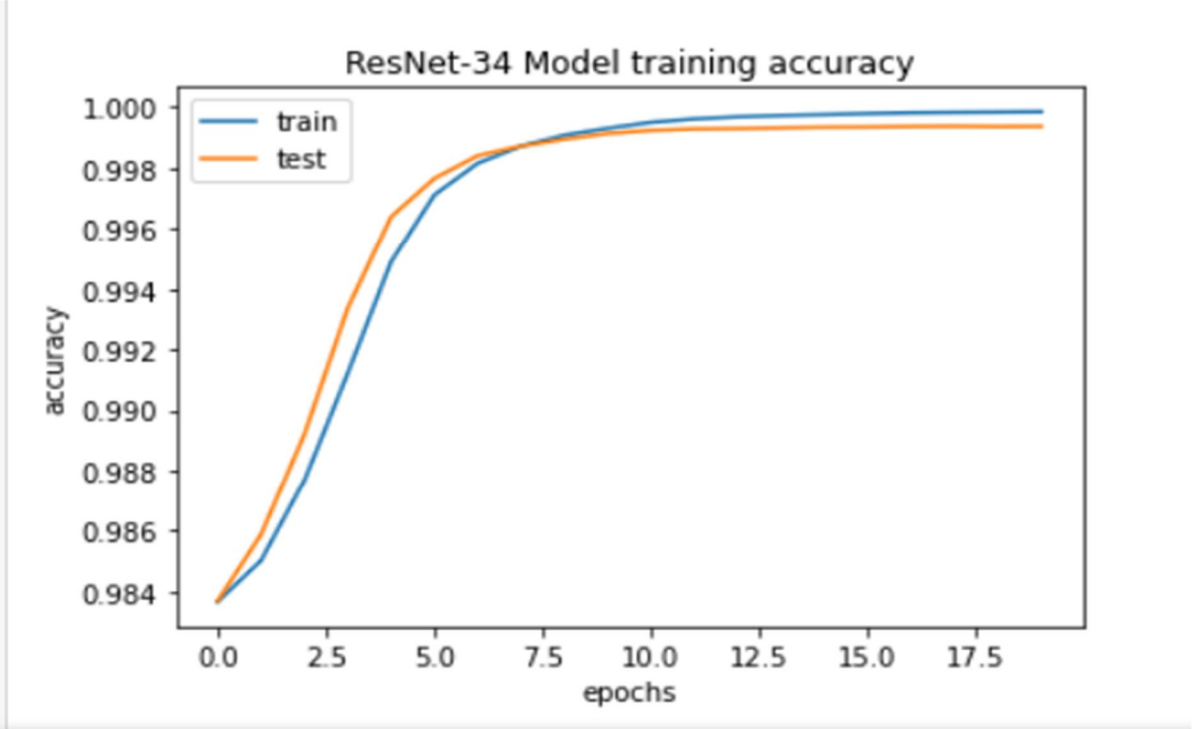


Figure 5-6: Test accuracy vs Training accuracy curve

The test and the training accuracy are clearly presented by figure 5.6. The above plot shows test curve is a little bit better than a validation curve. This incident is happening because of t he data used for validation is somewhat noise than the data used for testing. The top perfor mance of the system is above 98%. The model is well training the data and test result also ve ry good. The evaluative result of each word is presented as follows.

Table 5-6: Evaluation result for Amharic sign word

	Precession (%)	Recall (%)	F1 score (%)
ለምን	94	99	97
ሌላ	100	99	100
ልጅ	100	100	1.00
መልካም	100	100	100
መሰቅ	84	100	91
መቼ	76	100	86
መናደድ	100	100	100
መደሰት	100	100	100
ሚስት	60	15	24
ማን	94	100	0.97
ማዘን	100	100	100
ምሽት	100	100	100
ምን	92	100	96
ምንም	93	1.00	96
ሰላም	100	100	100
ሰነፍ	96	100	98
ሳምንት	97	100	99
ሴትአያት	99	98	99
ስም	100	100	100

ስንት	99	100	100
ቀን	84	100	91
በልልኝ	100	100	100
ባል	85	097	0.90
ቤተሰብ	99	100	1.00
ብር	94	1.00	0.97
ትላንት	100	00	00
ነው	82	1.00	90
ነገ	95	1.00	98
ንጋት	100	100	100
አመሰግናለሁ	93	100	96
ዓመት	100	100	100
አባት	94	100	97
አንተ	96	100	98
አንቺ	100	100	100
አክስት	0.98	1.00	0.99
አዎ	100	100	100
አይደለም	100	100	100
እህት	82	96	88
እሱ	100	100	100
እስዋ	96	100	98
እሺ	100	100	100
እነሱ	100	100	100
እናት	0.99	0.98	0.99
እኔ	100	100	100

እኛ	94	100	097
እያንዳንዱ	98	100	99
እግዚአብሄር	100	100	100
እጮኛ	100	100	100
ከትናንትበስቲያ	100	100	100
ከነገወድያ	100	100	100
ወር	100	100	100
ወንድአያት	61	1.00	0.76
ዐጎት	99	100	100
ዘመድ	100	100	100
የት	100	100	100
የትኛው	95	1.00	98
ይቅርታ	99	100	100
ጉዞ	100	100	100
ጎበዝ	100	100	100

The above table shows the test result for 60 sign words. The top twenty score of precession, recall and F1 score is respectively 99%, 100%, 99%. The total average accuracy of the system is presented below.

accuracy			0.95
macro avg	0.92	0.95	0.93
weighted avg	0.92	0.95	0.93

Figure 5-7: Accuracy of the proposed model

The total accuracy of the system out of 60 word is 95%. The macro average and weighted average of precession, recall and F1 score is respectively 92%, 95% and 93%.

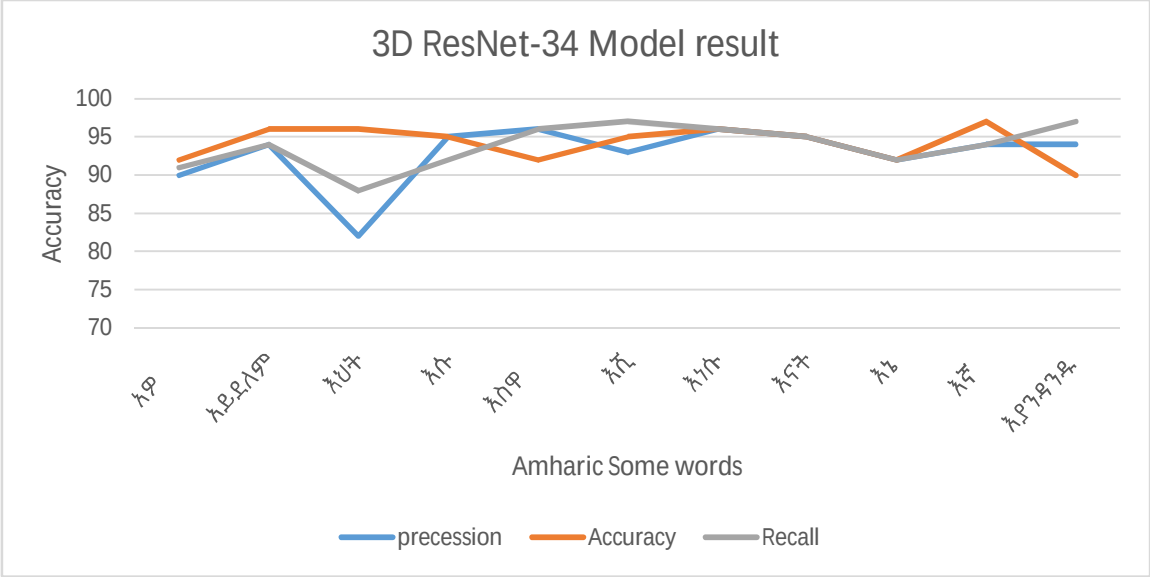


Figure 5-8: Accuracy, Precision and recall for some Amharic words

The above graph represents the accuracy, precision and recall for certain sign words. The overall result is nearly 99%, which indicate the model is consistent. This result comes from two things. The data used and the preferred model. The dataset is noise free; the length is not more than 4s, record by 16px Samsung camera and use 2400 sign video for training. The proposed model ResNet-34 is highly powerful model for image recognition.

Experiment 2: was done to compare Machine learning classifier SVM and NN done by [4] with the proposed deep learning classifier ResNet-34. Based on [4], the classification is done on character level, for the better comparison we work again on character level too. So that we collect a dataset of the first three derived letters of Amharic sign videos from signers (discussed in session 5.2) and feature extraction and classification done by deep learning algorithm of ResNet-34. The classification result of ResNet-34 is present as follows.

	precision	recall	f1-score	
ሁ	0.93	0.73	0.81	
ሂ	0.98	1.00	0.99	
ሃ	0.75	1.00	0.86	
ሄ	0.89	1.00	0.94	
ህ	1.00	1.00	1.00	
ሆ	0.93	1.00	0.96	
ሉ	0.57	1.00	0.72	
ሊ	0.98	1.00	0.99	
ላ	1.00	0.06	0.11	
ሌ	0.75	1.00	0.86	
ል	1.00	0.84	0.91	
ሎ	0.82	1.00	0.90	
ሐ	0.89	1.00	0.94	
ሒ	0.98	1.00	0.99	
ሓ	0.76	0.80	0.78	
ሔ	0.96	1.00	0.98	
ሕ	0.84	1.00	0.91	
accuracy				0.86
macro avg	0.84	0.86	0.82	
weighted avg	0.84	0.86	0.82	

Figure 5-9: Accuracy precision and recall for Amharic derived latters

The above result shows the precision, recall and F1 score on Amharic derived letters. The top three accuracy of ResNet-34 classification is 100%. The average accuracy on 18 Amharic derived letters classification result is 86%. In particular the classification result of "ሉ" is 57% this is because of the dataset used for "ሉ" is a little bet noise than others derived letters. This problem will solve by using more video data related to the specific class. The proposed classification algorithm ResNet-34 recognize all derived Amharic letters effectively. The comparison between ResNet-34 and NN is presented as follows.

Table 5-7: Classification comparison result for ResNet-34 and NN on the derived Amharic sign letters.

	Accuracy (%)		Precession (%)		Recall (%)		F1 score (%)	
	Based on [4], NN	The proposed system (ResNet-34)	Based on [4], NN	The proposed system (ResNet-34)	Based on [4], NN	The proposed system (ResNet-34)	Based on [4], NN	The proposed system (ResNet-34)
ሁ	57	86	27	93	30	73	28	81
ሂ	59	85	26	98	25	100	25	99
ሃ	57	87	23	75	22	100	22	86
ሄ	57	86	25	89	23	100	24	94
ህ	59	89	23	100	23	100	23	100
ሆ	59	87	25	93	22	100	23	96
ሉ	53	80	20	57	19	100	20	72
ሊ	49	84	21	98	21	100	21	99
ላ	36	89	21	100	21	60	21`	66
ሌ	50	88	28	75	28	100	28	86
ል	40	89	17	100	20	84	18	91
ሎ	31	89	18	82	18	100	18	90
ሐ	37	99	25	89	24	100	25	94
ሒ	51	99	25	98	25	100	25	99
ሓ	57	99	18	76	20	80	19	78
ሔ	62	99	26	96	24	100	25	99
ሕ	62	99	29	84	27	100	28	91
ሖ	65	93	25	96	28	98	26	99

The basic idea behind neural network (NN) is to simulate lots of densely interconnected brain cells inside a computer so recognize patterns and make decision in a humanlike way. NN is old machine learning algorithms. ResNet model was proposed to solve the issue of diminishing gradient. The idea is to skip the connection and pass the residual to the next connection. ResNet is deeper residual network. The comparison result accuracy, Precision, Recall and F-Score for NN Model vs ResNet-34 shows there is a clear big difference. This is why deep learning through layers that enable a computer to develop a hierarchy of complicated concepts from simpler concepts. One of the biggest advantages of using **deep learning** approach is its ability to execute feature engineering by itself. In this approach, an algorithm scans the data to identify features which correlate and then combine them to promote faster learning without being told to do so explicitly. This ability helps data scientists to save a significant amount of work.

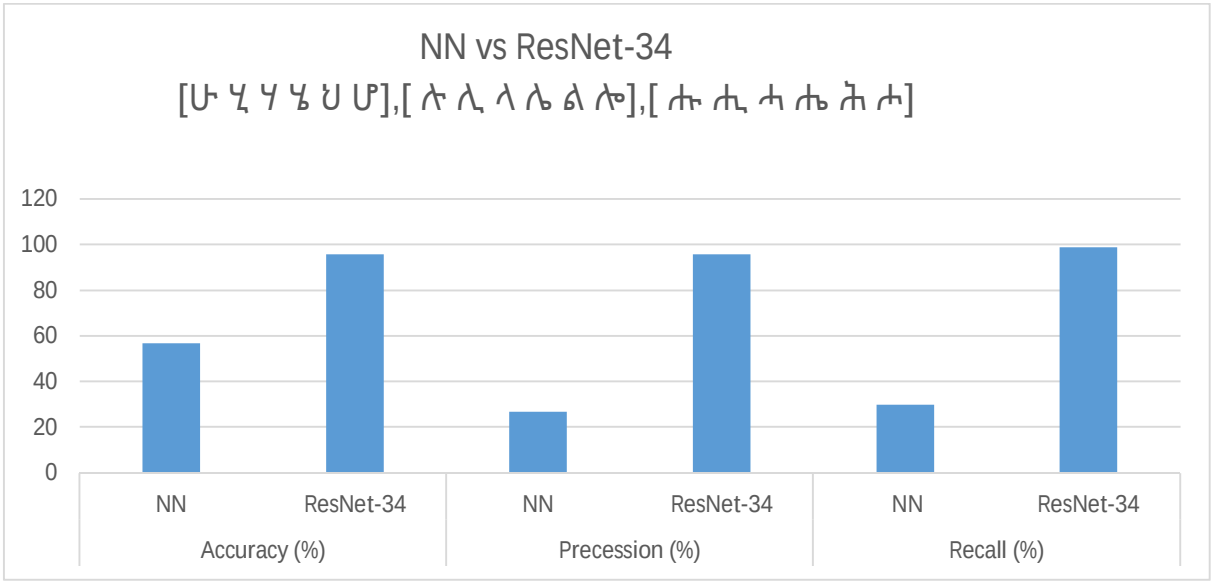


Figure 5-10: Comparison of NN and ResNet-34 bar graph

The above bar graph shows deeper network is excellent in recognition of 2D, 3D data. NN help to decision making process of a certain system whereas deep neural network can make a decision by themselves. The proposed classification algorithm ResNet-34 more efficient than NN.

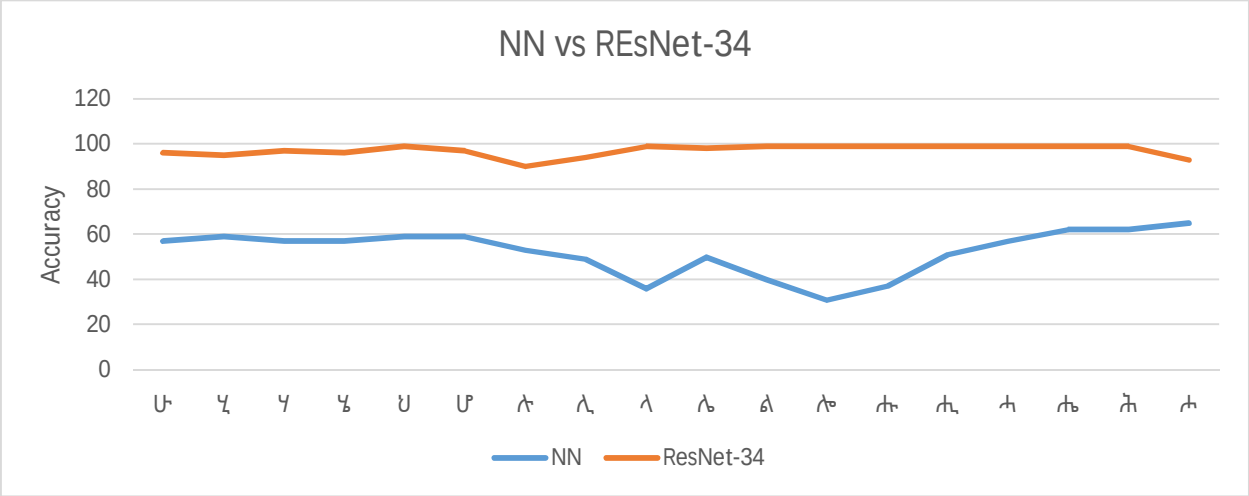


Figure 5-11: NN VS ResNet-34

Classification accuracy is simply the rate of correct classifications, either for an independent test set, or using some variation of the cross-validation idea. A good accuracy in a classification problem is 100%. The above graph shows ResNet-34 score nearly 100% whereas NN score at the maximum of 65%. The correct classifier is ResNet-34 than NN.

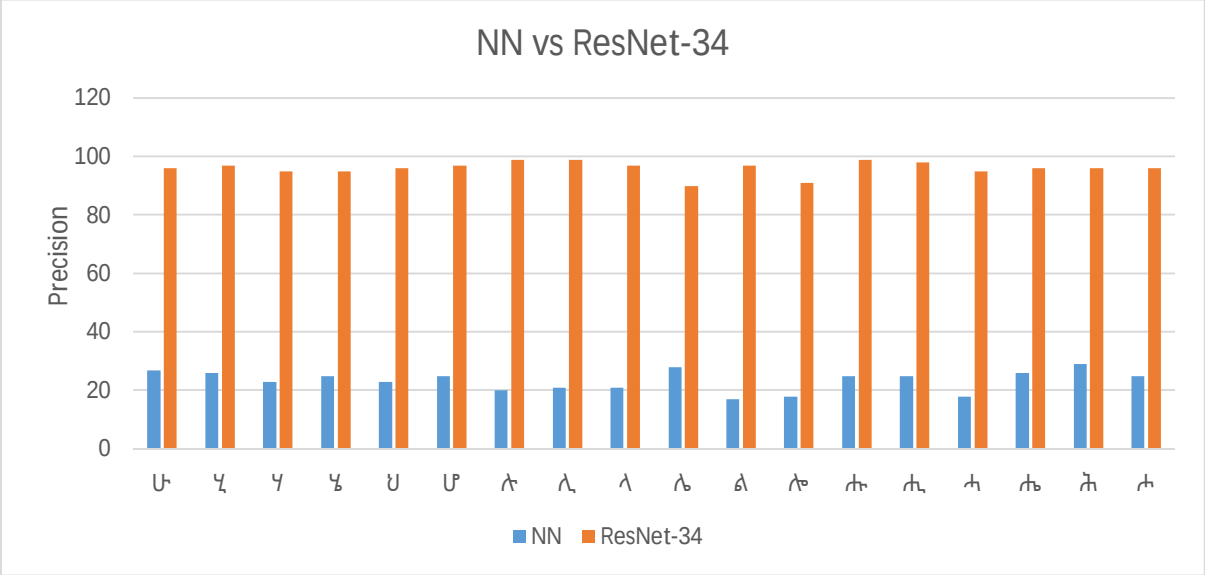


Figure 5-12: Precision for NN Vs ResNet-34

Precision help to answer what proportion of positive identifications was actually correct. The precision result of NN and the proposed algorithm ResNet-34 is quite different.

Precision takes all relevant predictions in to account. Positive predictive value of the data is higher in the case of ResNet-34 as shown below.

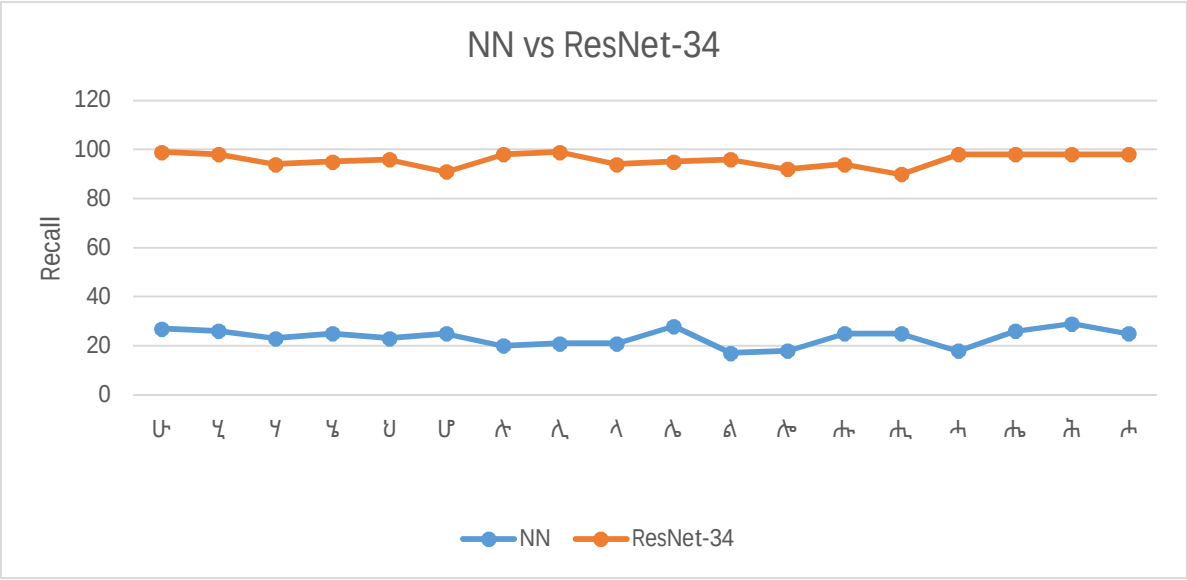


Figure 5-13: Recall for NN Vs ResNet-34

How many of the true positive were found represent recall? Recall actually calculates how many of the actual positives our model capture through labeling it as positive (True positive). The proportion of actual positives that was identified is better than the proposed model ResNet-34 than NN.

Table 5-8: Classification result comparison for Derived Amharic sign letters (ResNet-34 vs SVM)

	Accuracy (%)		Precession (%)		Recall (%)		F1 score (%)	
	Based on [4], SVM	The proposed system (ResNet-34)	Based on [4], SVM	The proposed system (ResNet-34)	Based on [4], SVM	The proposed system (ResNet-34)	Based on [4], SVM	The proposed system (ResNet-34)
ሁ	65	86	29	93	27	73	28	81
ሂ	72	85	35	98	35	100	35	99
ሃ	70	87	25	75	27	100	26	86
ሄ	74	86	32	89	38	100	35	94
ህ	76	89	31	100	31	100	31	100
ሆ	73	87	22	93	30	100	26	96
ሉ	72	80	29	57	40	100	33	72
ሊ	67	84	24	98	33	100	28	99
ላ	58	89	18	100	31	60	23	66
ሌ	62	88	29	75	27	100	28	86
ል	70	89	29	100	24	84	26	91
ሎ	68	89	32	82	29	100	30	90
ሑ	64	99	32	89	22	100	26	94
ሒ	66	99	30	98	20	100	24	99
ሓ	62	99	28	76	22	80	25	78
ሔ	65	99	31	96	26	100	28	99
ሕ	68	99	34	84	24	100	28	91
ሖ	69	93	21	96	31	98	25	99

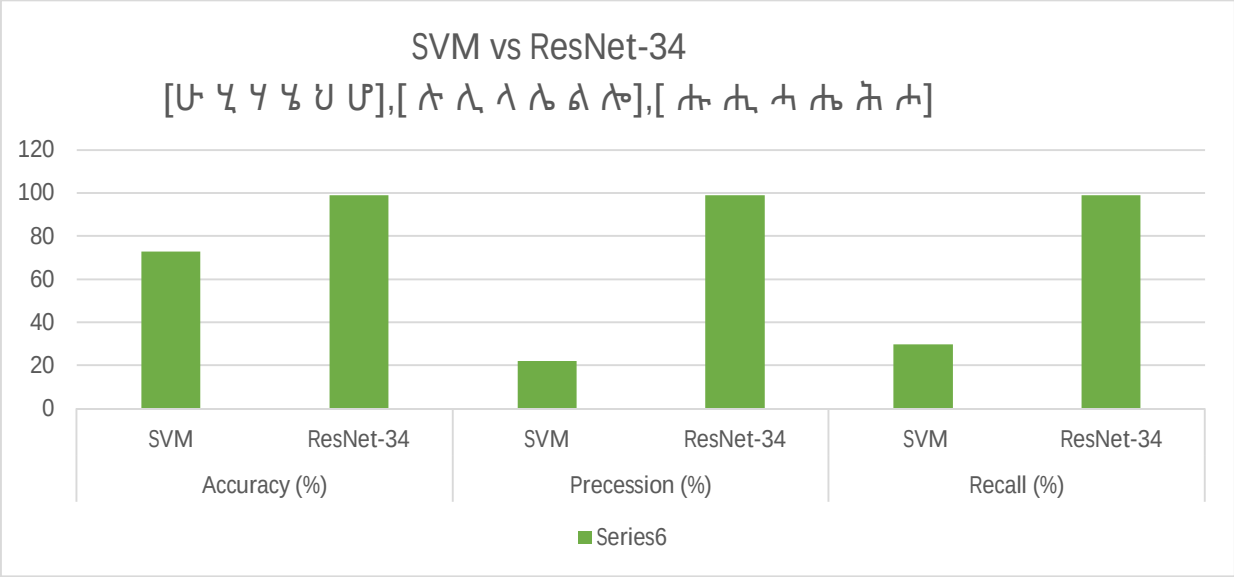


Figure 5-14: Comparison graph for SVM Vs ResNet-34

Figure 5-14: Comparison graph for SVM Vs ResNet-34

One of the advantages of machine learning algorithms like SVM is fast processing and real-time predictions. Machine learning algorithms work on practical scenarios gest like sign language recognition. But, in the case of adding more layer their efficiency is decline. Deep learning solves such problem by creating residual connection. So, the above graph shows SVM score at the maximum of 74% accuracy whereas ResNet-34 score nearly 100%.

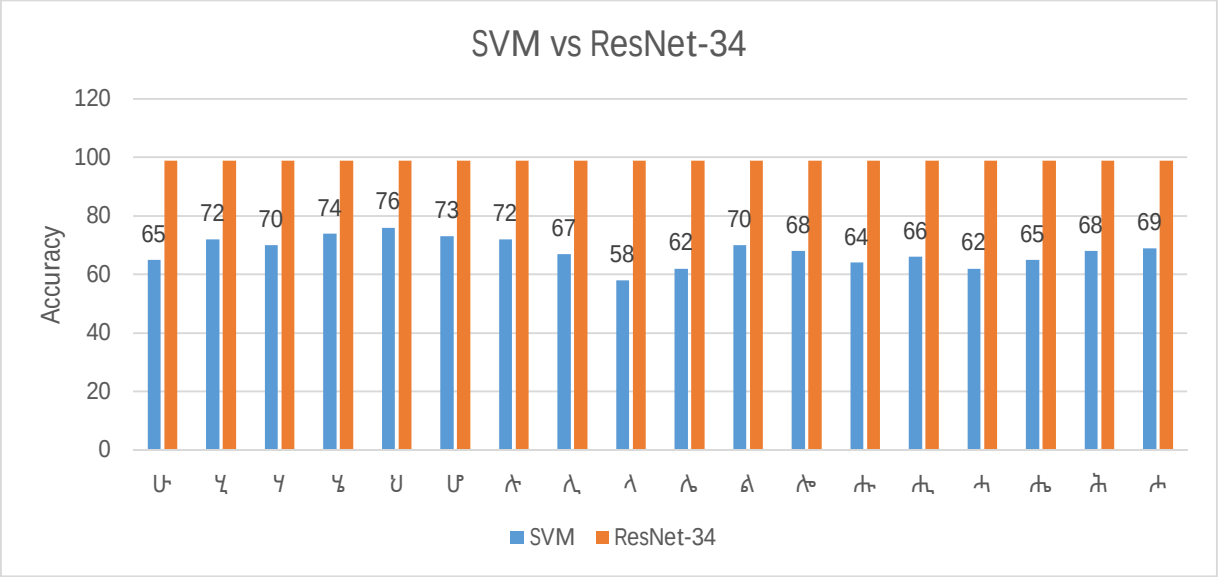


Figure 5-15: Accuracy for SVM Vs ResNet-34

The accuracy scored by SVM is between 60% - 75%. Whereas the accuracy achieved by ResNet-34 is nearly 99%. After the emerging of deep neural network, Residual network (ResNet) is a powerful algorithm for image (2D). Now a days it is also power full for 3D image classification. This is shown by the above graph clearly.

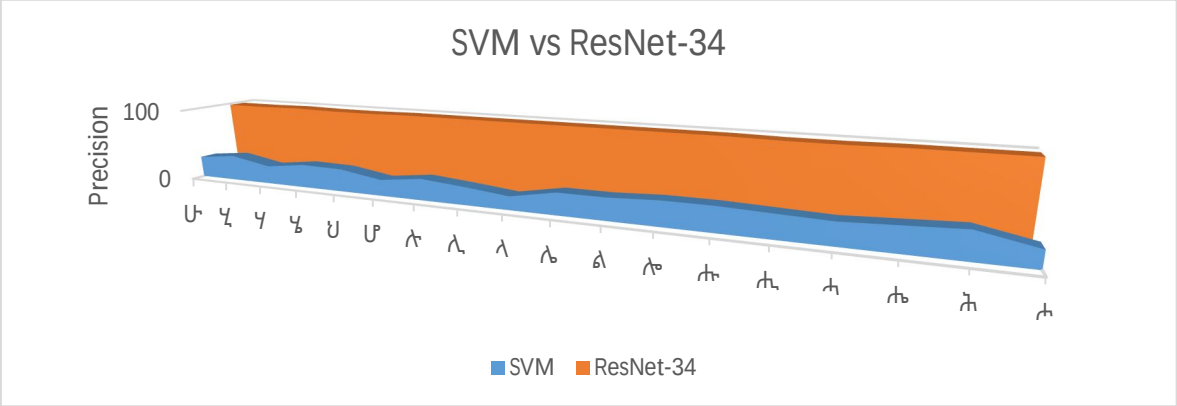


Figure 5-16: Precision for SVM Vs ResNet-34

SVM is old machine learning algorithm like NN. The precision result of SVM is much lower than ResNet-34. Precision refers how close measurements are to each other. The measurement result is similar in the case of ResNet-34.

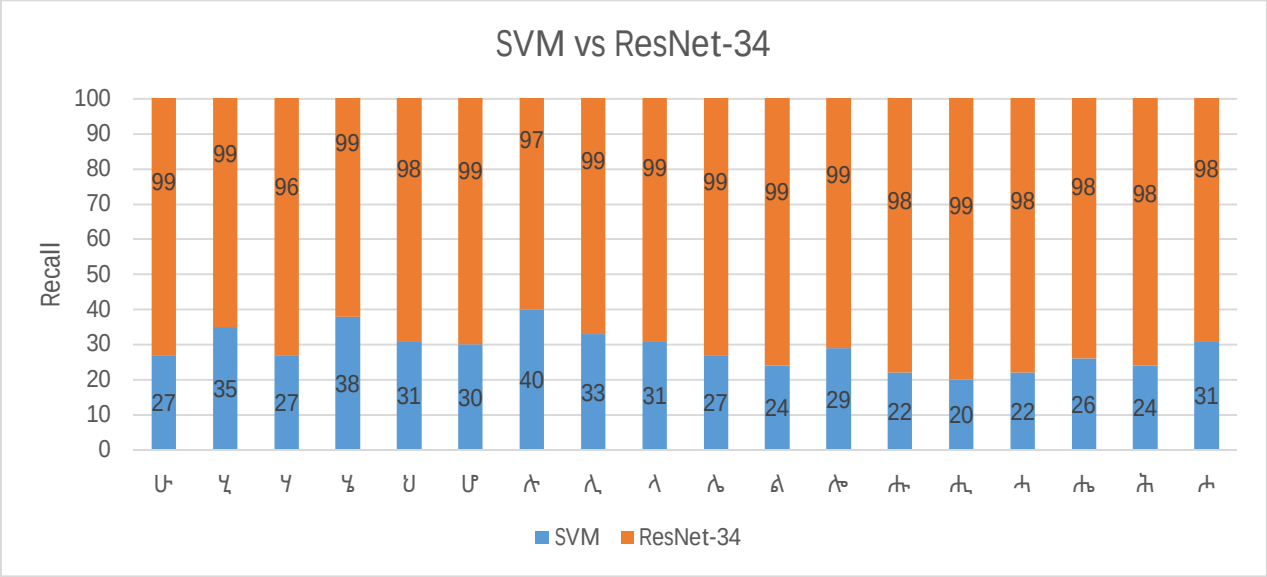


Figure 5-17: Recall for SVM ResNet-34

Recall also gives a measure of how accurately our model is able to identify the relevant data. Based on classification ResNet-34 is achieve great result than SVM.

Experiment 3: was done to test other language on the proposed model ResNet-34. We chose Argentinian sign language dataset which is available on free. Used Argentinian dataset is explained in session 5.2. The classification result is presented as follows.

	precision	recall	f1-score
Accept	1.00	0.68	0.81
Appear	0.95	0.60	0.73
Argentina	1.00	0.88	0.93
Away	0.47	0.79	0.59
Barbecue	0.61	0.58	0.60
Bathe	0.67	0.75	0.71
Biruthday	0.93	0.67	0.78
Bitter	0.82	0.70	0.75
Born	0.73	0.60	0.66
Breakfast	0.62	0.75	0.68
Bright	0.31	0.93	0.46
Buy	0.31	0.84	0.45
Candy	0.80	0.53	0.64
Catch	0.87	0.33	0.48
Chewing-gum	0.51	0.73	0.60

Colors	1.00	0.51	0.68
Copy	0.89	1.00	0.94
Country	0.80	0.77	0.78
Dance	0.87	0.57	0.69
Deaf	0.45	0.49	0.47
Drawer	0.33	0.65	0.44
Enemy	0.33	0.73	0.45
Find	0.88	0.22	0.35
Food	0.99	0.83	0.90
Give	0.47	0.73	0.57
Green	0.40	0.85	0.55
Help	1.00	0.30	0.46
Hungry	1.00	0.81	0.89
Last name	0.45	0.32	0.37
Learn	0.78	0.75	0.77
Light-blue	0.69	0.75	0.72
Man	0.43	0.75	0.55
Mock	0.87	0.74	0.80
Music	0.77	0.44	0.56
Name	0.79	0.49	0.60
None	0.87	0.37	0.52
Patience	0.55	0.75	0.64
Perfume	0.30	0.33	0.31
Photo	0.41	0.69	0.52
Realize	0.98	0.45	0.62
Red	0.94	0.66	0.77
Rice	0.47	0.72	0.57
Run	0.67	0.93	0.78
Ship	0.95	0.71	0.81
Shut down	1.00	0.10	0.19
Skimmer	0.52	0.47	0.49
Son	0.66	0.42	0.51
Spaggettui	0.78	0.73	0.76
Sweet milk	0.48	0.87	0.62
To land	0.43	0.44	0.43
Trap	0.29	0.24	0.26
Uuruguay	0.56	0.05	0.09
Water	0.83	0.69	0.75
Where	0.88	0.46	0.60
Women	0.80	0.50	0.61
Yellow	0.92	0.83	0.87
Yogurt	0.53	0.70	0.60
call	0.92	0.41	0.57
opaque	0.99	0.50	0.66

Figure 5-18: Argentinian sign language result

Argentinian sign language is not a family of Ethiopian sign language that is way the proposed model classification result is lower than classification result of Amharic sign words. But the top ten accuracy is 98% similar to the classification result of Amharic sign words. So, the model is efficient for other language too.

Experiment 4: was done on CNN-RNN model used by other researcher by using Amharic sign word dataset and compere the result to the first experiment. The comparison result is shown below.

Table 5-9: LSTM Vs ResNet-34

	Precession LSTM (%)	Precession ResNet-34 (%)	Recall LSTM (%)	Recall ResNet-34 (%)	F1 score LSTM (%)	F1 score ResNet-34 (%)
ለምን	84	94	84	99	84	97
ሌላ	85	100	83	99	88	100
ልጅ	85	100	85	100	85	1.00
መልካም	83	100	83	100	86	100
መሰቅ	80	84	86	100	81	91
መቼ	84	76	81	100	84	86
መናደድ	82	100	82	100	87	100
መደሰት	88	100	85	100	84	100
ሚስት	83	60	83	15	83	24
ማን	77	94	80	100	78	0.97
ማዘን	86	100	86	100	86	100
ምሽት	68	100	68	100	68	100
ምን	84	92	80	100	83	96
ምንም	83	93	79	1.00	83	82
ሰላም	85	100	85	100	85	100

ሰነፍ	76	96	76	100	76	98
ሳምንት	88	97	81	100	80	99
ሴትአያት	80	99	80	98	80	99
ስም	56	100	56	100	56	100
ስንት	85	99	85	100	85	100
ቀን	83	84	83	100	83	91
በልልኝ	84	100	84	100	84	100
ባል	81	85	87	097	84	0.90
ቤተሰብ	85	99	85	100	85	1.00
ብር	83	94	80	1.00	83	0.97
ትላንት	88	100	88	00	88	00
ነው	90	82	90	100	90	90
ነገ	77	95	74	100	79	98
ንጋት	83	100	83	100	83	100
አመሰግናለሁ	88	93	88	100	88	96
ዓመት	86	100	80	100	86	100
አባት	90	94	90	100	90	97
አንተ	87	96	81	100	84	98
አንቺ	92	100	92	100	92	100
አክስት	82	0.98	80	1.00	82	0.99
አዎ	83	100	83	100	83	100
አይደለም	81	100	81	100	81	100
እህት	78	82	78	96	78	88
እሱ	80	100	80	100	80	100
እስዋ	83	96	83	100	83	98

እሺ	88	100	86	100	80	100
እነሱ	87	100	84	100	80	100
እናት	82	99	80	0.98	82	0.99
እኔ	80	100	89	100	84	100
እኛ	81	94	81	100	81	097
እያንዳንዱ	89	98	89	100	89	99
እግዚአብሄር	88	100	88	100	88	100
እጮኛ	80	100	80	100	80	100
ከትናንትበስ ቲያ	83	100	83	100	83	100
ከነገወድያ	82	100	85	100	82	100
ወር	83	100	86	100	88	100
ወንድአያት	87	61	84	1.00	82	0.76
ዐነት	84	99	88	100	89	100
ዘመድ	85	100	85	100	85	100
የት	81	100	81	100	81	100
የትኛው	84	95	86	1.00	84	98
ይቅርታ	82	99	82	100	82	100
ጉዞ	76	100	77	100	76	100
ጎበዝ	83	100	79	100	82	100

The above comparison table clearly shows ResNet-34 classification result is better than LSTM classification result.

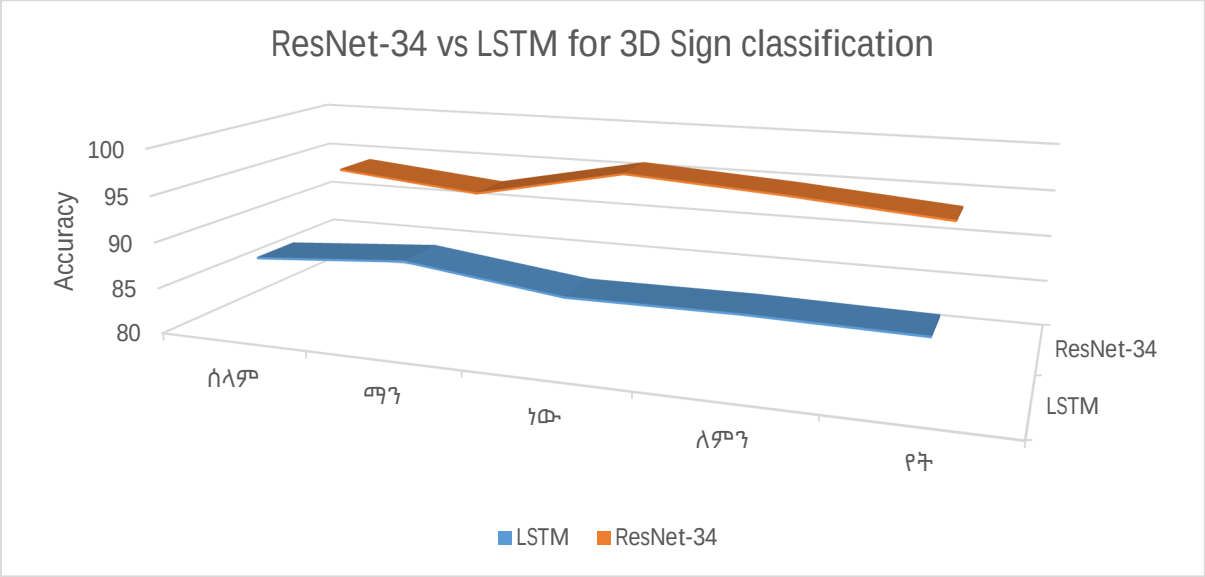


Figure 5-19: Accuracy for LSTM Vs ResNet-34

This is because LSTM is originally profound for speech recognition whereas CNN work on image recognition and classification. Based on [29], LSTM score 72.3% for sign sentences and 89.5% for sign word. This day, 2D CNN is increased in to 3D CNN. So CNN handle a sequence of frames effectively than LSTM and other old machine leaning algorithms.

- ✓ **5.6. Threats to Validity**

5.6.1. Internal Threats to Validity

Inward dangers to legitimacy are dangers which are brought about by instrumentation issues and uncontrolled factors. The instrumentations utilized in the analyses were PCs and other simultaneous cycles run simultaneously the investigations were running on the PCs. The time needed to remove the highlights was truly huge which takes over a day on the elite PC regardless of whether time isn't the significant worry of this examination. Giving the component extraction a shot, a superior performing PC may bring about lower time prerequisite.

5.6.2. External Threats to Validity

The dataset utilized in this investigation isn't so a lot higher. Profound learning approach need large information to score high precision. The nature of the dataset isn't likewise

acceptable on the grounds that it recorded by advanced cell it is smarter to recorded sign video by utilizing computerized camera that improve the nature of the dataset.

✓ 5.7 Discussion

The proposed Amharic language through signing recognizer is evaluated from the capacity of learning machines (ResNet-34) to perceive Amharic language through signing into the relating text character. Learning-capacity of the chose machines is evaluated via preparing the 80 % of the dataset, validation takes 10 % and getting their accuracy by utilizing the leftover 10% for testing.

The exploratory consequence of ResNet-34 showed huge execution more prominent than 95% for recognition of the chose 60 sign words. Regardless of whether profound learning need enormous dataset, the dataset utilized in this exploration was restricted by framework memory (hard disk). During the process of immense data, RAM and processer speed likewise notice an ability issue. We attempt to take care of this issue by receiving a batch of data rather than processing the whole data. The issue of speed and memory may not be serious issue for Amharic gesture-based communication recognition. The fundamental concern is obtaining a model that recognizes Amharic sign language with better accuracy.
Word level Amharic language through signing recognition isn't done at this point however past endeavors to perceive Amharic gesture-based language recognition accomplished great outcome. One such endeavor is that of Neguse Kefiyalew who has accomplished the recognition rate of 74 % utilizing SVM. He centers on character level and perceive all Amharic characters just as three derived Amharic character of ሀ , ለ , ሐ .

The answer of the research question which was mentioned in chapter one is the system correctly recognize sign words based on ResNet-34 and also this research answer deep learning is better than machine learning algorithms by comparing ResNet-34 with SVM and NN which was done by the previous researcher on three derived Amharic characters based on video sequence.

CHAPTER SIX: CONCLUSION AND FUTURE WORK

✓ 6.1 Conclusion

Hearing issue is basic in Ethiopia. A huge number of individuals are living with this issue. These individuals utilize Amharic language via gestures for communication reason. In any case, it is hard to speak with individuals who can't realize Amharic gesture-based communication. There is no framework that can interpret sign words or sentence to corresponding text or sound. This hole makes the existence of hearing impeded individuals extremely challenging. In Ethiopia, there is an exploration work that make an interpretation of Amharic characters sign to Amharic text yet at the same time an issue of acknowledgment of sign words. This investigation is applied to take care of the issue of sign word acknowledgment utilizing profound learning algorithms.

In this examination, an endeavor has been made to plan and execute a framework which is capable for perceiving Amharic communication via gestures words. The framework has five significant parts: Video dataset assortment, video to frame change, frame preprocessing, Feature extraction and Classification. Frame preparing incorporate cropping and RGB to grayscale transformation. Feature extraction and classification is finished by utilizing profound learning residual network ResNet-34.

The preliminary task is video dataset assortment which is finished by utilizing Samsung mobile. Prior to begin to record, select 60 frequently utilized sign words for this exploration work. Furthermore, discover the signers who sign the chose words. At that point, record each word with a length of 3-4s. In the wake of recoding all video the next task is video to frame change by utilizing diverse methods. The framework begins by accepting RGB video frames of an Amharic sign words. However, two preprocessing steps are essential. RGB to grayscale change and Cropping. RGB to grayscale transformation help to minimize the competition. The proposed feature extraction and classification model get 224 x 224 elements of frames so editing is vital.

Before feature extraction and classification, the algorithm split the dataset into training, validation and testing. Feature extraction is finished by utilizing powerful profound learning algorithm ResNet-34. Concentrate the element of each frame and saved by .csv format. After finishing of feature extraction, classification is finished by a similar algorithm ResNet-34.

The last piece of the framework is sign recognition. For sign order, ResNet-34 classifier is utilized. Test result show that the ResNet-34 classifier accomplished a general exactness of 95%. The explanation behind accomplishing very well is comes fundamentally from the pre-owned algorithm ResNet, ResNet is ground-breaking for picture acknowledgment. Input to ResNet-34 for this examination is a batch of frames which is a bit contrast from getting a single frame. The second factor of accomplishing an excellent outcome is a dataset utilized is recording and gathering appropriately. Utilize 50 recordings for each sign words.

Taking everything into account, our created framework that perceives 60 sign words by utilizing profound learning algorithm ResNet-34.

- ✓ **6.2 Future Work**

In this paper, we score an outcome for acknowledgment of sign words, yet at the same time there is a hole on conveying full gesture-based communication acknowledgment administration. Coming up next are a portion of the suggestion that the analysts propose for future work:

- Improve the proposed plan for better speed highlight extractor and arrangement calculation.

- Enhancing the acknowledgment framework by actualizing on many sign words by utilizing incredible gadget.

- This work has utilized ResNet-34 and has a decent assessment result for acknowledgment. Consequently, stretching out this work to expression and sentence level for the investigation of the language can be a solid match for future explores.

- •Our work introduced around single direction correspondence which implies it just makes an interpretation of the sign to message. Along these lines, we proposed that other scientist will plan the framework which should work like two-route correspondence to make an interpretation of sign to text and the other way around.

- In this exploration work, the primary test was getting word level sufficient sign recordings. There ought to be an information base that incorporates all the Amharic sign words with high frequencies.

REFERENCE

[1] Robert Smith, HamNoSys4.0 for Irish Sign language Workshop Hand book, Ireland: Dublin City University, 2010

[2] Eyasu Hailu, Sign language news, Addis Ababa University, 2009.

[3] Legesse Zerubabel, "Ethiopian Finger Spelling Classification: A Study To Automate Ethiopian Sign Language," Master's Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2008.

[4] Nigus Kefyalew Tamiru, “Amharic Sign Language Recognition based on Amharic Alphabet Signs” Master's Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2018.

[5] በኢትዮጵያመስማትየተሰናቸውማህበር፣ “የአማርኛየምልክትቃንቃመስማትናመናገርለተሰናቸው” 2003 ዓም

[6] French Sign Language at *Ethnologue* (18th ed., 2015)

[7] About American Sign Language, Deaf Research Library, Karen Nakamura

[8] Indian Sign Language at *Ethnologue* (22nd ed., 2019)

[9] *"Archived copy" (PDF). Archived from the original (PDF) on 14 December 2013. Retrieved 14 December 2013*

[10] Yang Quan and Peng Jinye, "Chinese Sign Language Recognition for a Vision-Based Multi-features Classifier," in International Symposium on Computer Science and Computational Technology, Shaanxi Xi’an, P. R. China, 2008.

[11] Atiqur, Ahsan, Ibrahim and Sujit, "Recognition of Static Hand Gestures of Alphabet in Bangla Sign Language," IOSR Journal of Computer Engineering (IOSRJCE), vol. 8, no. 1, pp. 7-13, 2012.

[12] Abcdf Costello”American sign alphabet recognition ” (2008:xxv)

[13] Tefera Gimbi, "Recognition of Isolated Signs in Ethiopian Sign Language," Master's Thesis,Addis Ababa University, 2014

[14] *"Amharic". Ethnologue. Retrieved 8 December 2017*

[15] *Ankita Wadhawan1, Parteek Kumar* “Deep learning-based sign language recognition system for static signs” 2019

[16] Harish chandra thuwal, adhyan srivastava “real time sign language gesture recognition from video sequences” , 2017

[17] dongxu.li, cristian.rodriguez, xin.yu, hongdong.li “Word-level Deep Sign Language Recognition from Video:A New Large-scale Dataset and Methods Comparison” 2019

[18] Zirui Jiao, Naiming Yao, Hui Chen, Fengchun Qiao, Zhihao L, Hongan Wang” An Ensemble of VGG Networks for Video-Based Facial Expression Recognition”2018

[19] Yanqiu liao, Pengwen xiong, Weidong min, weiqiong min, and jiahao lu, “dynamic sign language recognition based on video sequence with blstm-3d residual networks” 2019

[20] መኮንንሙላት፣አሳይጉታ፣ሚርያሂማነን፣ “የምልክትቃንቃመማርያለጀማረዎች” 2008 ዓም

[21] Lee S, Kwon H, Han H, Lee G, Kang B. A space-variant luminance map based color image enhancement. IEEE T Consum Electr 2010; 56: 2636-2643.

[22] https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148, [Oct. 24, 2020]

[23] Ahmed Ali Mohammed Al-Saffar, Hai Tao, Mohammed Ahmed Talab "Review of deep convolution neural network in image classification" 2017

[24] D. Scherer, A. Muller and S. Behnke, "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition," 20th International Conference on Thessaloniki, Greece, 2010

[25] Prasun Roy∗, Subhankar Ghosh∗, Saumik Bhattacharya∗ and Umapada Pal "Effects of Degradations on Deep Neural Network Architectures", 2019

[26] *Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun***"**Review of Deep Residual Learning for Image Recognition**"**

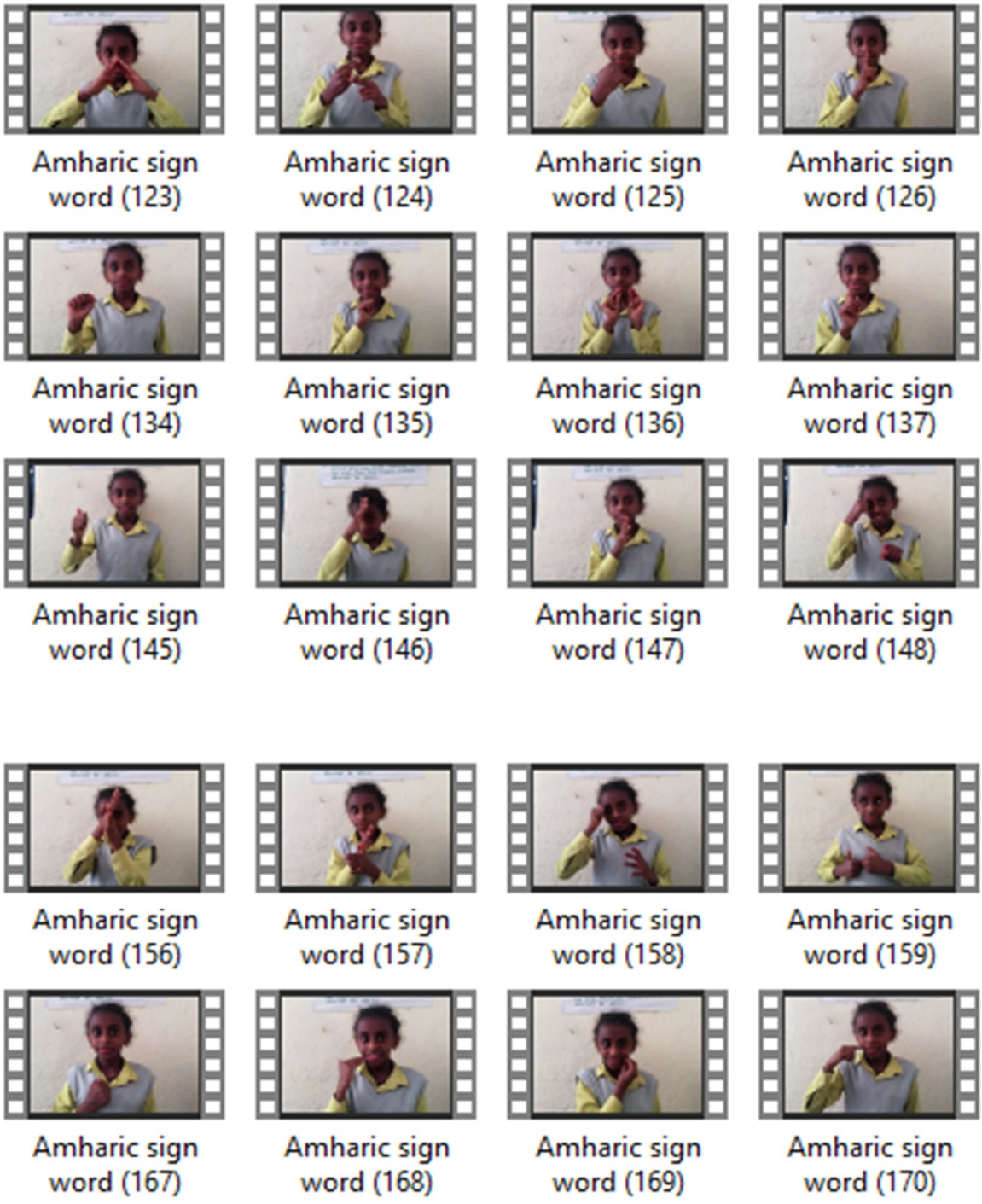
[27] Milan Kumar Asha Paul , Janakiraman Kavitha and P. Arockia Jansi Rani," Key-Frame Extraction Techniques: A Review" 2018

[28] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks (CNN). In NIPS, 2012.

[29] Anshul Mittal, Pradeep Kumar, Partha Pratim Roy, Raman Balasubramanian and Bidyut B. Chaudhuri "A Modified-LSTM Model for Continuous Sign Language Recognition using Leap motion" 2019

Appendixes

APPENDIX A: SAMPLE DATA USED FOR SYSTEM DESIGN





Amharic sign word (128)



Amharic sign word (129)



Amharic sign word (130)



Amharic sign word (131)



Amharic sign word (132)



Amharic sign word (139)



Amharic sign word (140)



Amharic sign word (141)



Amharic sign word (142)



Amharic sign word (143)



Amharic sign word (150)



Amharic sign word (151)



Amharic sign word (152)



Amharic sign word (153)



Amharic sign word (154)



Amharic sign word (161)



Amharic sign word (162)



Amharic sign word (163)



Amharic sign word (164)



Amharic sign word (165)

APPENDIX B: PYTHON CODE

I. video to frame convertion

```
# assumption only first 51 frames are important
while count < 51:
    ret, frame = cap.read()  # extract frame
    if ret is False:
        break
    framename = os.path.splitext(video)[0]
    framename = framename + "_frame_" + str(count) + ".jpeg"
    hc.append([join(gesture_frames_path, framename), video, frameCount])

    if not os.path.exists(framename):
        frame = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
        lastFrame = frame
        cv2.imwrite(framename, frame)

        if cv2.waitKey(1) & 0xFF == ord('q'):
            break
        count += 1
```

II. feature extraction

```
    # open the output CSV file for writing
    csvPath = os.path.sep.join([config.BASE_CSV_PATH,
            "{}.csv".format(split)])
    csv = open(csvPath, "w")
    # loop over the images in batches
    for (b, i) in enumerate(range(0, len(imagePaths), config.BATCH_SIZE)):

            print("[INFO] processing batch {}/{}".format(b + 1,
                    int(np.ceil(len(imagePaths) / float(config.BATCH_SIZE)))))
            batchPaths = imagePaths[i:i + config.BATCH_SIZE]
            batchLabels = le.transform(labels[i:i + config.BATCH_SIZE])
            batchImages = []
            # loop over the images and labels in the current batch
            for imagePath in batchPaths:
                    image = load_img(imagePath, target_size=(224, 224))
                    image = img_to_array(image)
                    image = np.expand_dims(image, axis=0)
                    image = preprocess_input(image)
                    # add the image to the batch
                    batchImages.append(image)
            batchImages = np.vstack(batchImages)
            features = model.predict(batchImages, batch_size=config.BATCH_SIZE)
            features = features.reshape((features.shape[0], 7 * 7 * 2048))
            # loop over the class labels and extracted features
            for (label, vec) in zip(batchLabels, features):
                    # construct a row that exists of the class label and
```

```
                    # extracted features
                    vec = ",".join([str(v) for v in vec])
                    csv.write("{},{}\n".format(label, vec))

        # close the CSV file
        csv.close()

# serialize the label encoder to disk
f = open(config.LE_PATH, "wb")
f.write(pickle.dumps(le))
f.close()
```

III. classfication

```
# train the network
print("[INFO] training simple network...")
H = model.fit(
        x=trainGen,
        steps_per_epoch=totalTrain // config.BATCH_SIZE,
        validation_data=valGen,
        validation_steps=totalVal // config.BATCH_SIZE,
        epochs=25)

# make predictions on the testing images, finding the index of the
# label with the corresponding largest predicted probability, then
# show a nicely formatted classification report
print("[INFO] evaluating network...")
predIdxs = model.predict(x=testGen,
        steps=(totalTest //config.BATCH_SIZE) + 1)
predIdxs = np.argmax(predIdxs, axis=1)
print(classification_report(testLabels, predIdxs,
        target_names=le.classes_))
```