4-2023

INTERACTIVE EMIRATE SIGN LANGUAGE E-DICTIONARY BASED ON DEEP LEARNING RECOGNITION MODELS

Ahmed Abdelhadi Abdelhadi

UAEU

جامعة الإمارات العربية المتحدة
United Arab Emirates University

MASTER THESIS NO. 2023: 23

College of Information Technology

Department of Computer Science and Software Engineering

INTERACTIVE EMIRATE SIGN LANGUAGE E-DICTIONARY BASED ON DEEP LEARNING RECOGNITION MODELS

Ahmed Abdelhadi Ahmed Abdelhadi

April 2023

United Arab Emirates University

College of Information Technology

Department of Computer Science and Software Engineering

INTERACTIVE EMIRATE SIGN LANGUAGE E-DICTIONARY BASED ON DEEP LEARNING RECOGNITION MODELS

Ahmed Abdelhadi Ahmed Abdelhadi

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science in Software Engineering

April 2023

United Arab Emirates University Master Thesis
2023: 23

Cover: ESL e-Dictionary System Interface

(Photo: by Ahmed Abdelhadi Ahmed Abdelhadi)

Print: University Print Service, UAEU 2023

Declaration of Original Work

I, Ahmed Abdelhadi Ahmed Abdelhadi, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this thesis entitled "*Interactive Emirate Sign Language E-Dictionary Based on Deep Learning Recognition Models*", hereby, solemnly declare that this is the original research work done by me under the supervision of Dr. Munkhjargal Gochoo, in the College of Information Technology at UAEU. This work has not previously formed the basis for the award of any academic degree, diploma or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my thesis have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation and/or publication of this thesis.

Student's Signature: ____________________

Date: 25/05/2023

Approval of the Master Thesis

This Master Thesis is approved by the following Examining Committee Members:

1) Advisor (Committee Chair): Dr. Munkhjargal Gochoo
   Title: Assistant Professor
   Department of Computer Science and Software Engineering
   College of Information Technology

   Signature ________ Date 2023.04.14

2) Member: Dr. Fady Alnajjar
   Title: Associate Professor
   Department of Computer Science and Software Engineering
   College of Information Technology

   Signature ________ Date 2023 04. 14

3) Member (External Examiner): Dr. Omar Mubin
   Title: Associate Professor
   School of Computer, Data and Mathematical Sciences
   Institution: Western Sydney University, Australia

   Signature ________ Date 18/04/2023

This Master Thesis is accepted by:

Dean of the College Information Technology: Professor Taieb Znati

Signature Taieb Znati

Date 26/05/2023

Dean of the College of Graduate Studies: Professor Ali Al-Marzouqi

Signature Ali Hassan

Date 29/05/2023

Abstract

According to the ministry of community development database in the United Arab Emirates (UAE) about 3065 people with disabilities are hearing disabled (Emirates News Agency - Ministry of Community Development). Hearing-impaired people find it difficult to communicate with the rest of society. They usually need Sign Language (SL) interpreters but as the number of hearing-impaired individuals grows the number of Sign Language interpreters can almost be non-existent. In addition, specialized schools lack a unified Sign Language (SL) dictionary, which can be linked to the Arabic language being of a diglossia nature, hence many dialects of the language co-exist. Moreover, there are not sufficient research work in Arabic SL in general, which can be linked to the lack of unification in the Arabic Sign Language. Hence, presenting an Emirate Sign Language (ESL) electronic Dictionary (e-Dictionary), consisting of four features, namely Dictation, Alpha Webcam, Vocabulary, and Spell, and two datasets (letters and vocabulary/sentences) to help the community in exploring and unifying the ESL. The vocabulary/sentences dataset was recorded by Azure Kinect and includes 127 signs and 50 sentences, making a total of 708 clips, performed by 4 Emirate signers with hearing loss. All the signs were reviewed by the head of the Community Development Authority in UAE for compliance. ESL e-Dictionary integrates state-of-the-art methods i.e., Automatic Speech recognition API by Google, YOLOv8 model trained on our dataset, and an algorithm inspired by bag of words model. Experimental results proved the usability of the e-Dictionary in real-time on laptops. The vocabulary/sentences dataset will be publicly offered in the near future for research purposes.

Keywords: Emirate Sign Language, Automatic speech recognition, ESL data set, ESL electronic dictionary, YOLO.

Title and Abstract (in Arabic)

قاموس تفاعلي إلكتروني للغة الإشارة الإماراتية يعتمد على نماذج التعرف والتعلم التفاعلي

الملخص

بحسب قاعدة بيانات وزارة تنمية المجتمع في دولة الإمارات العربية المتحدة فإن حوالي 3065 شخصاً من ذوي الإعاقة هم من ذوي الإعاقة السمعية (وكالة أنباء الإمارات - وزارة تنمية المجتمع). يجد ضعاف السمع صعوبة في التواصل مع بقية المجتمع. عادة ما يحتاجون إلى مترجمين للغة الإشارة، ولكن مع تزايد عدد الأشخاص الذين يعانون من ضعف السمع، يمكن أن يكون عدد مترجمي لغة الإشارة غير موجود تقريبًا. بالإضافة إلى ذلك، تفتقر المدارس المتخصصة إلى قاموس موحد للغة الإشارة، والذي يمكن ربطه بكون اللغة العربية ذات طبيعة ثنائية اللغة، وبالتالي تتعايش العديد من لهجات اللغة. علاوة على ذلك، لا توجد أعمال بحثية كافية في لغة الإشارة العربية بشكل عام، ويمكن ربط ذلك بعدم وجود توحيد في لغة الإشارة العربية. وعلى ذلك، نقدم قاموس إلكتروني بلغة الإشارة الإماراتية (ESL)، يتكون من أربع ميزات، وهي الإملاء، وكاميرا ويب ألفا، والمفردات، والتهجئة، ومجموعتين من البيانات (الحروف والمفردات / الجمل) لمساعدة المجتمع في استكشاف وتوحيد لغة الاشارة الإماراتية. تم تسجيل مجموعة بيانات المفردات / الجمل بواسطة Azure Kinect وتتضمن 127 علامة و50 جملة، يؤديها 4 إماراتيين يعانون من ضعف السمع. نتج عن ذلك 708 مقطع تسجيل، تمت مراجعة جميع التسجيلات من قبل رئيس هيئة تنمية المجتمع في الإمارات للتأكد من صحة تنفيذها ودقتها. يدمج قاموس ESL الإلكتروني أحدث الأساليب، مثل واجهة برمجة تطبيقات التعرف التلقائي على الكلام من Google، ونموذج YOLOv8 المدرب على مجموعة البيانات الخاصة بنا، وخوارزمية مستوحاة من نموذج حقيبة الكلمات. أثبتت النتائج التجريبية سهولة استخدام القاموس الإلكتروني في أرض الواقع على أجهزة الحاسوب المحمولة. سيتم تقديم مجموعة التسجيلات المفردات / الجمل للجمهور في المستقبل القريب لأغراض البحث.

مفاهيم البحث الرئيسية: لغة الإشارة الإماراتية، التعرف التلقائي على الكلام، قاموس إلكتروني للغة الإشارة الإماراتية، مجموعة التسجيلات للغة الإشارة الإماراتية، YOLO.

Author Profile

Ahmed Abdelhadi Ahmed is currently a Technology specialist and teacher in the Ministry of Education, UAE, and a research assistant with the department of Mechanical engineering in the university of Auckland, New Zealand NZ. He also wrote several engineering books naming few, Intro to Data Science, Co-space Robotics Engineering book v1 & v2, Introduction to Machine learning, for the Ministry of education UAE and Telecommunication and digital Regulations Authority TDRA in UAE. Ahmed is also one of the 8 winning nominees of the first ever Abu Dhabi Artificial Intelligence Summit 2022, to showcase his work on the “Emirate Sign Language” and to showcase the first ever Emirate Sign Language e-Dictionary system. He has also published in the IEEE XPLORE ASET Conference 2023. Moreover, he worked in the research lab of Robotics and Artificial intelligence at UAEU during his Master studies. Ahmed lives in Khorfakkan/Sharjah. He worked as a research assistant and teaching assistant in the University of Khartoum Electrical and Electronics College where he received his Bachelor of Science’s degree with Honor in Electronics and Computer Systems Engineering.

Acknowledgements

I would like to thank my committee for their guidance, support, and assistance throughout my preparation of this thesis, especially my advisor Dr Munkhjargal Gochoo.

Special thanks go to the UAEU CIT administration for their continued guidance, with the upmost respect and homage to Mariam Al Mandhari, the senior administrator.

Dedication

To my beloved Mother, Samira Elhaj Mohamed Elshiekh Sharshab, my inspiration and motivation, to my family, and to myself, for choosing to be a warrior in a garden rather than a gardener in a war.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AP	Average Precision
ArSL	Arabic Sign Language
ASR	Automatic Speech Recognition
CAMSA	The council of Arab Ministers of Social Affairs
CDA	Community Development Authority
DL	Deep Learning
ESL	Emirate Sign Language
FCOS	Fully Convolutional One-Stage
IoU	Intersection over Union
mAP	Mean average precision
mAP	Mean Average Precision
ML	Machine learning
MS COCO	Microsoft Common Object in Context
SL	Sign Language
SLD	Sign Language Dictionary
YOLO	You Only Look Once

Chapter 1: Introduction

Unaddressed hearing loss can impact many aspects of life at an individual level and societal level. Individuals with hearing loss and an inability to communicate properly go uneducated and unschooled, which yields a higher rate of unemployment. And it may also lead to social isolation, loneliness, and stigma [1]. On a societal level, WHO estimates that unaddressed hearing loss poses an annual global cost of US$ 980 billion [1]. As the number of hearing-impaired is estimated to increase [1] the need for a unified and certified electronic Sign Language (SL) Dictionary (SLD) increases with it. One of the main obstacles facing the creation of a certified SLD in the Arab region arises from the Arabic language being a diglossia language as many variations of the language coexist together [2]. However, with the recent advances in technologies in the field of Deep Learning (DL), computer vision has had promising progress in the field of Action Recognition and motion detection using DL and Machine Learning (ML). In computer vision, ML performs an essential role to extract key information from media. Lately, computer vision is leaning toward healthcare [3] and people with hearing loss. Therefore, the research on Sign Language recognition and continuous Sign Language recognition have gained a massive interest which yields the need for a well-rounded dataset for Arabic Sign Language [4] [5]. Up to this date, an Emirate Sign Language (ESL) dataset has not been found.

To cope with the abovementioned issues, we present the ESL e-Dictionary, which consists of features and a dataset, features namely: Dictation, Text-to-Sign, Alpha Webcam, and spelling. The dataset was collected, viewed, and reviewed by certified Sign Language interpreters to ensure the correctness of the 127 signs and 50 sentences used in the ESL e-Dictionary by the head of the Community Development Authority (CDA) Sign Language department in Dubai. The ESL e-Dictionary is inspired by the hearing-impaired community in the UAE as the language is a means to communicate and to overcome communication barriers between hearing-impaired individuals and the rest of society. The Automatic Speech Recognition (ASR) algorithm API is the state of the art in dictation, powered by Google API [6]. The Text-to-Sign feature is inspired by the bag of word algorithm where the entries are looked up in the directory list and matched to the corresponding sign. In addition, the ESL e-Dictionary uses YOLOv8 [7] object detection to detect objects and it was trained to detect Arabic alphabet signs, outperforming YOLOv5l in [8] with the ArSL21l dataset [8]. Moreover, it measures the correctness of the sign demonstrated. Finally, the system can spell words and names. Hence, A hearing-impaired can easily learn and explore new words by just typing them, and the system shall show the corresponding sign, a user can easily say a word and the screen shall show the corresponding signs.

All the above-mentioned features are in real-time and get processed immediately. The system is also designed to help newcomers and children with hearing loss as it can teach them the alphabet, spelling, and the first three vocabulary levels of ESL. Finally, the ESL e-Dictionary is aimed toward educational institutions that teach ESL to become an easy-to-access electronic reference and to ensure the unification of the Emirate Sign Language in the region.

Chapter 2: Methods

2.1 Arabic Sign Language Recognition

Several attempts have been taken to interpret the Arabic language into Arabic Sign Language (ArSL). Typically, scripts are translated into ArSL after the speech has been turned into text. El-Gayyar et al. [9] were successful in achieving 79.8% satisfaction with their translation system via cloud computing. Alfi et al. [10] have described a desktop program for educational environments that produces a series of motionless images of ArSL for selectable text input achieving a generous 96% accuracy. Aouiti and Jemni [11] have made an effort to create an ArSL interpreter web application, based on a dictionary of signs and words. By using an avatar-based representation of the text in Sign Language. Halawani and Zaiton [12] have demonstrated a desktop application that can translate Arabic words into ArSL. A method that transforms Arabic text to ArSL was created by Almohimeed et al. [13] using a corpus of 203 signed sentences and 710 unique signs. Because there are so few signs in the corpus, this system's average accuracy for translating from text-to-Sign Language is only 53.3%. A system developed by Luqman and Mahmoud [14] has 600 phrases and 5,637,151 vocabularies. Their approach, which accomplished a success rate of 82%, is built on three primary stages: morphological analysis (sentence analysis and word extraction), syntactical analysis (word extraction is utilized to generate sentence structure), and ArSL production stage (word extraction is translated into their system).

2.2 Arabic Sign Language Datasets

The Council of Arab Ministers of Social Affairs (CAMSA) is a committee within the Arab states that covers 22 Arab countries, the middle east, and north Africa. Pushed for a standard ArSL but it was met with great resistance, and this can be linked to the Arabic language being a diglossia language as in if an individual from Morocco and another from Egypt met, they will not be able to understand one another or communicate as effectively [2] [15]. But in 2004, CAMSA produced the Arabic Sign Language dictionary which was standardized to the MENA to provide a reference to situations a hearing-impaired may encounter in the Arab world [2]. Despite having a standardized Arabic Sign Language dictionary, various versions of the language can be found in the middle east which is not related to each other's or the Arabic language [2]. Which result to Arab deaf communities to use non unified Arabic Sign Language [16].

In 2021, [16] has promised in their scientific paper to release a dataset called: The Jumla Qatari Sign Language Corpus, in a promise to solve the issue of the lack of unification in the Arabic deaf communities. However, until the date of this work, no such dataset has been published.

In 2018, the UAE launched the Emirate Sign Language (ESL) Dictionary in their official portal under the Zayed Higher Organization (ZHO) of people of determination institution to unify the ESL signs in the UAE [17] yet as far as the search went, no ESL dataset was found.

2.3 Object Detection

Real-time object detection is a very important topic in computer vision, as it is an essential element in computer vision systems. For example, multi-object tracking [18], autonomous driving [19], robotics [20], etc.

Currently, state-of-the-art real-time object detections models are essentially based on Fully Convolutional One-Stage (FCOS) and You Only Look Once (YOLO)

The metric used to evaluate and measure object detection models is the mean Average Precision (mAP) value. Taking into consideration, the Precision value is the measurement of the fraction of the True Positives (TP) to the overall predicted positives, as shown in (1), where FP stands for False Positives.

$$Percision = \frac{TP}{TP+FP} \qquad (1)$$

Recall measures how well you can spot all positives, the formula is shown in (2).

$$Recall = \frac{TP}{TP+FN} \qquad (2)$$

Intersection over Union (IoU), which refers to the overlap between two boxes, as shown in Figure 1, in this case, how much the foreseen boundary overlaps with the actual object boundary. In some datasets the IoU threshold is predefined to determine false negative or a true positive, usually IoU $\geq$ 0.5. The formula of IoU is as seen in (3).

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} \qquad (3)$$

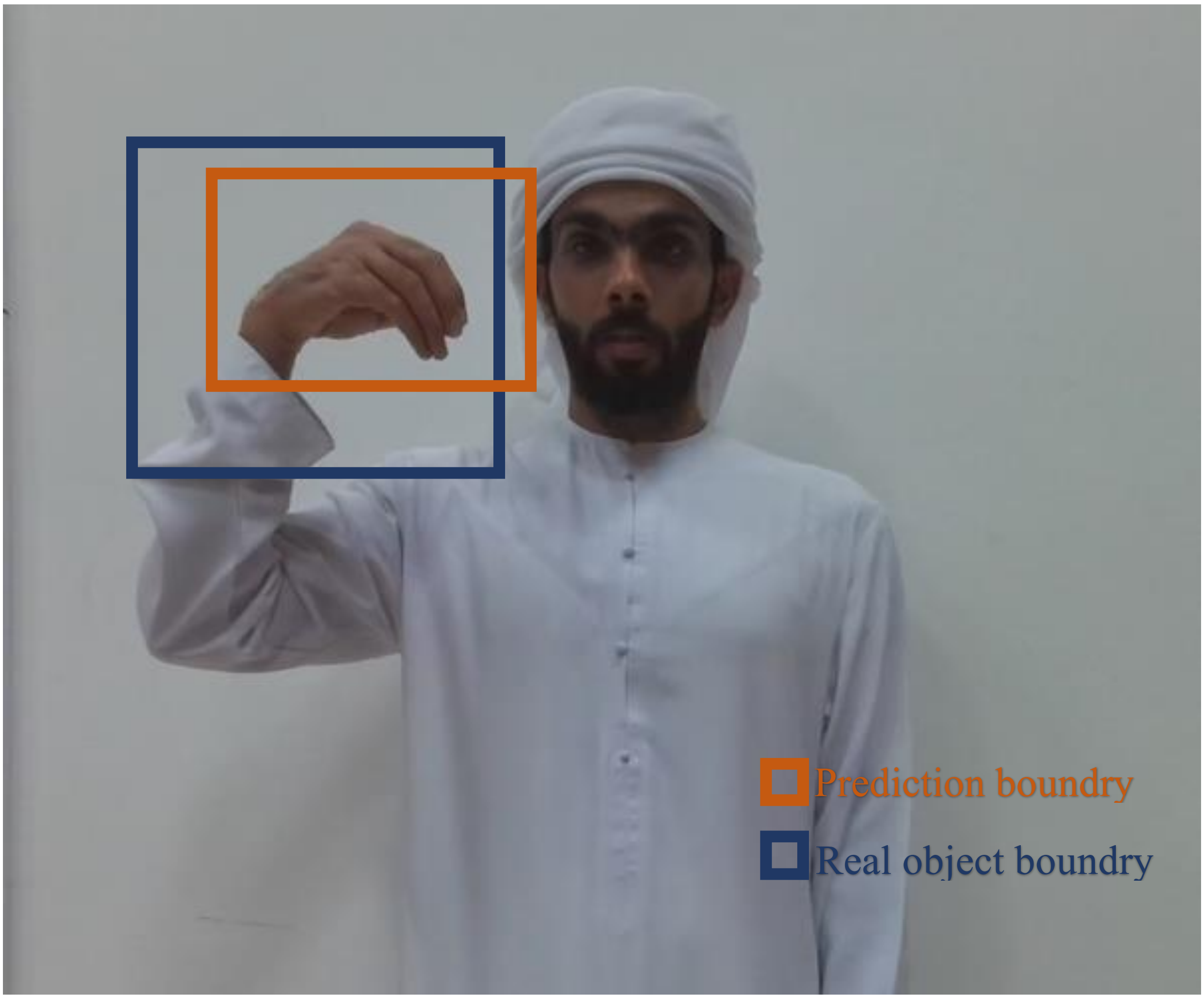

Figure 1: IoU Definition for Showing Boundary Boxes Overlaps

Average Precision (AP) is the area under the precision-recall curve. Where precision is represented in the y-axis and recall is represented in the x-axis, as shown in Figure 2. The general formula for AP is shown in (4).

$$Average\ Precision\ (AP) = \int_0^1 p(r)dr \qquad (4)$$

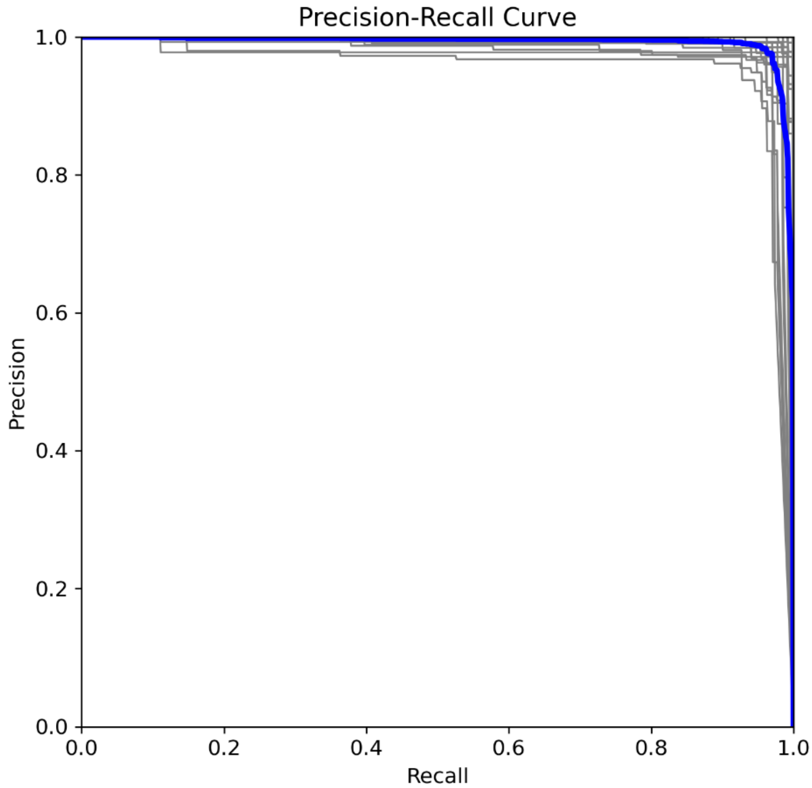

Figure 2: Precision-Recall Curve where the Area Under the Curve is the Average Precision (AP)

Hence, mAP is the average AP for all classes. And this is what most research papers use to evaluate their results on the Common Object in Context (COCO) dataset, a sample of the dataset is shown in Figure 3, which is a visual dataset provided by Microsoft [21].

In the Microsoft COCO (MS COCO) challenge, the guidelines [21] states a threshold for IoU, which is the minimum value of IoU to consider a detection to be positive. The $mAP_{0.5:0.95}$, Refers to a mean average precision with IoU between 0.5 to 0.95 threshold, with a step size of 0.05. There are many other metrics involved and collected for the COCO challenge [21], but the main interest is $mAP_{0.5:0.95}$.

Figure 3: A Sample from the MS COCO Dataset where Objects (Ground Truths) are Highlighted with Annotations of the Objects, Present in the Picture, Annotated at the Top of The Picture [21]

2.3.1 FCOS

FCOS, is an anchor box free object detector [22], in other words, erasing the values of predefined anchor boxes. Hence, discarding the use for the computations related to anchor boxes [22]. Also, removing the hyper-parameters needed in the calculations of anchor boxes [22].

One of the most popular anchor-free Detectors might be YOLO version 1 [23], where instead of utilizing anchor boxes, it foresees bounding boxes at locations close to the center of objects and only those locations near the center are used, since they are believed to be able to produce accurate detection [22]. However, since only the location near the center is used to predict bounding boxes, YOLO version 1 suffers from low recall value [23]. Hence, YOLO version 2 [24] implements anchor boxes.

FCOS, predict bounding boxes using all points in a ground-truth bounding box [22]. Moreover, it discards low-quality bounding boxes by 'center-ness' branch. Hence, FCOS can produce a competitive recall with the likes of YOLO version 2.

The strategy behind 'center-ness' branch, is the addition of a single layer branch, without introducing any hyper-parameters, that goes in parallel with the classification branch, as can be seen on Figure 4.

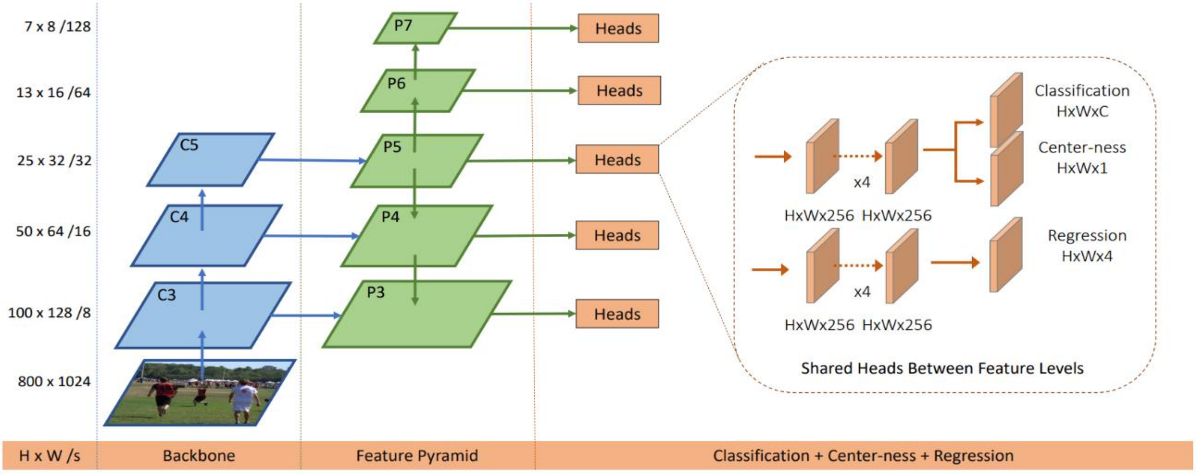

Figure 4: The Network Architecture of the Original FCOS, where Feature Maps of the Backbone Network are Produced by C3, C4, and C5 to P3 to P7 which are Feature Levels Used in the Final Prediction. In the Shared Heads Between Feature Levels the Center-Ness Reside in Parallel to the Classification Layer [22]

The author of [22] introduced 'center-ness' branch to the model as he has observed many of the low-quality bounding box predictions generated by locations very far from the object's center was the top problem of FCOS. The 'center-ness' branch estimates the score associated with the low-quality bounding box via the formula shown in (5).

$$Centerness = \sqrt{\frac{\min(l,r)}{\max(l,r)} \times \frac{\min(t,b)}{\max(t,b)}} \qquad (5)$$

Center-ness decays from 1 to 0 as the bounding box deviates from the center of the object.

In [22] the experiment report indicated that they followed the large-scale detection benchmark COCO [25] common practice for training, validation, and testing.

As ResNet-50 [26] the backbone network and implementing the same hyper-parameters of ResNet-50 [26]. Improvements were made constantly throughout the experiments by moving the center-ness branch to the regression branch, Or sampling only the central parts of the ground-truth boxes for positive samples, etc. Those improvements were referred to as cost-free improvements in [22].

FCOS has two minor differences in comparison to RetinaNet, the use of Group Normalization (GN) into the convolution layers that were introduced, except for layers P6 and P7. Where P5 is used to produce the layers mentioned - before, P6 and P7, and that's the second minor difference.

Finally, at the time of the publication and experiment, the FCOS w/improvements has outperformed all other object-detection models with 44.7 AP, including YOLO version 2 at that time. But since the release of YOLO version 4 [27] no further literature or experiment was done on FCOS at the time of this research. It's worth noting that the reported improvements and AP value were reported after the initial submission.

2.3.2 YOLO

YOLO "You Only Look Once" [7] is an algorithm that allows real-time object detection via utilizing Deep Neural Networks (DNN). It comes with many versions, moreover, each version has many models that differ by size. YOLOv#s, YOLOv#m, YOLOv#s, YOLOv#l, and YOLOv#x, where # stands for the version sequence and the letters s, m, l, and x stands for small, medium, large, extra-large, respectively.

YOLOv4 [27] addressed a very important problem at the time of their publication which was that the most precise present-day neural networks do not function in real-time, additionally, they demand a huge number of Graphics Processing Units (GPU) for the training part of the detection. YOLOv4 [27] addressed such problems via creating a Convolution Neural Network (CNN) that operates on a convolutional GPU in real-time, yet, only requiring a single convolutional GPU.

In [27] they presented two approaches for real-time neural networks:

- For GPU, they use groups of 1 to 8, which is relatively small, in convolutional layers.
- For Vision Processing Unit (VPU), they abstain from using the Squeeze-and-Excitement (SE) blocks in grouped-convolution.

Additional improvements were made for the designed detector [27] to be more adjusted for the training on a single GPU. Those improvements were made in the design as follows:

- A new approach for data augmentation: Self-Adversarial Training (SAT), and Mosaic.
- Genetic algorithms and selection of optimal hyper-parameters.
- A modification to: Spatial Attention Module (SAM), Path Aggregation Network (PAN), and Cross mini-Bath Normalization (CmBN).

Self-Adversarial Training [27] is a data augmentation approach that functions in two parts. The First part the neural network applies altercations and modifications to the original images without applying any alters to the network weights. Hence, the neural network performs an attack on itself. To induce a deception to the neural network that there is no object of interest in the image. In the second part, the neural network is trained to find an object on those altered images in the normal fashion way.

Mosiac [27] is a data augmentation technique that shuffles four training images, therefore, the mixture of four different contexts. This yields detection of objects outside the ordinary context.

They [27] achieved results of 43.5% Average Precision (AP) on the Microsoft COCO dataset [21] at a real-time speed, Outperforming all other object detection models in the COCO challenge, at the time of their publication [27].

The alphabets dataset, ArSL21L [8], was benchmarked in [8] using YOLOv5 as the object detection model, relying on COCOmAP or Mean Average Precision ($mAP_{0.5:0.95}$) since it is the common metric. This paper shall rely on the same methods to bench mark the ArSL21L on the latest model of YOLO, YOLOv7 and YOLOv8.

Chapter 3: Emirates Sign Language E-Dictionary

3.1 Dataset

Our collected dataset for the Emirate Sign Language (ESL) e-Dictionary covers the first 3 levels of ESL. Hence, 127 signs covering the first 3 levels out of 7 levels from the Emirate Sign Language tabulated in Table 1. In addition, our dataset has 50 sentences, tabulated in Table 2, constructed from the 127 signs. Each sign and sentence was recorded 4 times by 4 different hearing-impaired individuals. Figure 5 represents samples of four different signers performing signs.

Figure 5: Signs from Left to Right: Ambassador, Sheikh, Family, and Cook

Each signer signed a consent form to be recorded and documented, and for the collected recordings to serve the prosperity of the academic field. During the recording sessions a certified Sign Language interpreter was present to monitor and ensure each sign/sentence was performed correctly. Finally, the recordings were verified and reviewed by the head of the Sign Language department in the Community Development Authority (CDA) in Dubai, UAE. The recordings are of 10 seconds, 30 frames per second, as signs/sentences can vary from 3 seconds sign to 8 seconds, Azure Kinect v2 camera was used to capture signs/sentences in depth and RGB videos. A recording of each sign and sentence set was chosen according to the level of clarity and accuracy of the demonstration, therefore the ESL e-Dictionary has 177 signs/sentences using the upmost quality samples of the dataset. Samples from the dataset used in the ESL e-Dictionary are shown in the figures from Figure 6 to Figure 13.

Table 1: Signs of Level I, II, and III of the Emirate Sign Language

Emirate Sign Language, Signs of Level I, II, and III									
1	Ambassador	**26**	Day	**51**	Health card	**76**	Zoo	**101**	Translator
2	Astronaut	**27**	Dead	**52**	Hungry	**77**	Pharmacy	**102**	Tuesday
3	Aunt	**28**	Color blue	**53**	Husband	**78**	Photograph	**103**	Uncle
4	Barber	**29**	Delay	**54**	ID Card	**79**	Phycologist	**104**	Sister
5	Birth certificate	**30**	Divorce certificate	**55**	Mother's sister	**80**	Salary certificate	**105**	Lieutenant 1st
6	Be patient	**31**	Doctor	**56**	Injection	**81**	Pray	**106**	Walking
7	Boy	**32**	Eat	**57**	Interviewer	**82**	Pregnant	**107**	Want
8	Brother	**33**	Evening	**58**	Journalist	**83**	Prevent	**108**	Wednesday
9	Burj al Arab	**34**	Family	**59**	Jumeirah	**84**	Read	**109**	Week
10	Burj Khalifa	**35**	Father	**60**	Laugh	**85**	Requests	**110**	Wife
11	Chief	**36**	Fever	**61**	Lawyer	**86**	Sad	**111**	Work
12	Citizenship	**37**	Fire fighter	**62**	Liar	**87**	Play	**112**	Xray
13	Clean	**38**	First aid	**63**	Love	**88**	Saturday	**113**	Year
14	Color black	**39**	Flu	**64**	Volunteer	**89**	Sheikh	**114**	Yesterday
15	Death certificate	**40**	Good conduct certificate	**65**	Marriage certificate	**90**	University certificate	**115**	Mother's brother
16	Color brown	**41**	Generous	**66**	Imam	**91**	Sleep	**116**	0
17	Color green	**42**	Girl	**67**	Me	**92**	Son	**117**	1
18	Color orange	**43**	Global village	**68**	Membership card	**93**	Sorry	**118**	2
19	Color pink	**44**	Friday	**69**	Medicine	**94**	Stingy	**119**	3
20	Color purple	**45**	Grandfather	**70**	Minister	**95**	Sunday	**120**	4
21	Color red	**46**	Grandmother	**71**	Monday	**96**	Think	**121**	5
22	Color white	**47**	Guard	**72**	Morning	**97**	Thursday	**122**	6
23	Color yellow	**48**	Happy	**73**	Mother	**98**	Time	**123**	7
24	Colors	**49**	Hate	**74**	You	**99**	Today	**124**	8
25	Daughter	**50**	Headache	**75**	Nervous	**100**	Tomorrow	**125**	9
126	Zayed grand Mosque			**127**	People of determination card				

Table 2: Sentences Constructed from 127 Words of the First 3 Levels of the ESL

50 Sentences			
1	Seven days a week	**26**	I want a translator
2	My Aunt is pregnant	**27**	The lawyer is nervous
3	The boy is sad	**28**	The girl is happy
4	The mother laughs	**29**	You want a marriage certificate
5	I want an ID card	**30**	The imam prays
6	The mother cleans	**31**	The family prays
7	Walk/head to Zayed grand mosque	**32**	The Girl volunteers
8	I want to eat	**33**	You are stingy/cheap
9	Grandmother is generous	**34**	The day of Friday
10	Today is Saturday	**35**	Father goes Jumeirah
11	Yesterday was Sunday	**36**	Thursday evening
12	Tomorrow is Monday	**37**	Tomorrow is work
13	You be patient	**38**	I want the color red
14	I hate the color yellow	**39**	I want the color brown
15	I am a doctor	**40**	My uncle is an interviewer
16	I want a good conduct certificate	**41**	I need my university certificate
17	You want a health card	**42**	I want a people of determination card
18	I need a membership card	**43**	The day is Wednesday
19	I hate the color black	**44**	I want the color purple
20	I want the death certificate	**45**	Grandfather has a fever
21	The guard has the flu	**46**	I love the color white
22	The injection is at 7 in the morning	**47**	Today is Wednesday
23	Today is Tuesday	**48**	Today is Sunday
24	Today is Monday	**49**	I love the color blue
25	I love the color orange	**50**	I love the color green

Figure 6: Sign for Laugh

Figure 7: Sign for Food

Figure 8: Sign Color Black

Figure 9: Sign for Father

Figure 10: Sign for Family

Figure 11: Sign for Colors

Figure 12: Sign for Fireman

Figure 13: Sign for Flu

3.2 Key Gestures

As humans, distinguish between signs via certain gestures, and can refer to those gestures in the sequence of a sign signing as key-gestures. Figure 14, Figure 15, and Figure 16 below show distinguishable key-gestures of the signs Ambassador, Family, and Zoo.

Figure 14: Key Gestures for the Sign Ambassador

Figure 15: Key Gestures for the Sign Family

Figure 16: Key Gestures for the Sign Zoo

To emphasize on those key-gestures while still maintaining a smooth demonstration to the user, the frame rate has been manipulated to only keep the frames with great change with comparison to the previous frame.

3.2.1 Keyframes

Careaga et al. [28] provide metric-based few shots learning for video action recognition via two stream models, a convolution set and a recurrent neural network video encoder structure. It was evaluated on the Kinetics 600 [29] dataset. They showed a result of 84.2% accuracy and 59.4%, on the test set and specific test set for a 5-shot and 5-way task, respectively. Hence, as a video enhancement technique, a low frame rate on the few-shot algorithm can result in a higher accuracy [30][31]. Therefore, each recording of our dataset went through the process of converting the recordings to Graphics Interchange Format (GIF) images [32] and was tuned to 10 keyframes per second. The 10 frames were chosen as follows: the second frame of every 3 sequential frames. e.g., in Figure 17 the gray boxes show a 30-frame sequence of a one second, and the frames marked red are the chosen frames in the given sequence. The number of low frames per second ensures the efficiency of the system and help to explore in the direction of the few-shot learning model. Therefore, our dataset has 10 frames per second and 10-second 177 Gifs/videos.

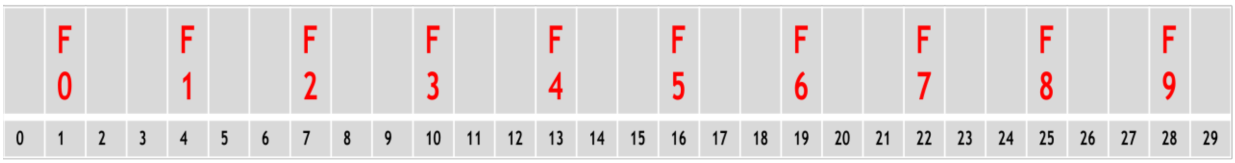

Figure 17: Chosen Frames Sequence are Indicated by F0 to F9, which Mark the Chosen Frames Per Every 30 Frames

Chapter 4: System Design

The system includes four main features: Dictation, Vocabulary, Alpha Webcam, and Spell feature. Dictation and Vocabulary are features that allow the user to look up the directory. The system has two appearance modes: light and dark, and it can behave according to the system's appearance as well. Additionally, the status of the mic is present for the user to monitor the system activity. The logo acts as a window to present the signs to the user. The system interface is shown in Figure 18.

The flowchart of the system is shown in Figure 19, excluding the spell feature, it will be shown in the upcoming sections.

Figure 18: ESL E-Dictionary Main Window, the Logo Acts as the Channel to Return Sign/Sentence to the User. The Left Panel has the Main Features and Controls of the System. On the Right Panel there is a Mic Status Monitor and a Dictation Box to Show the Spoken Phrase/Word

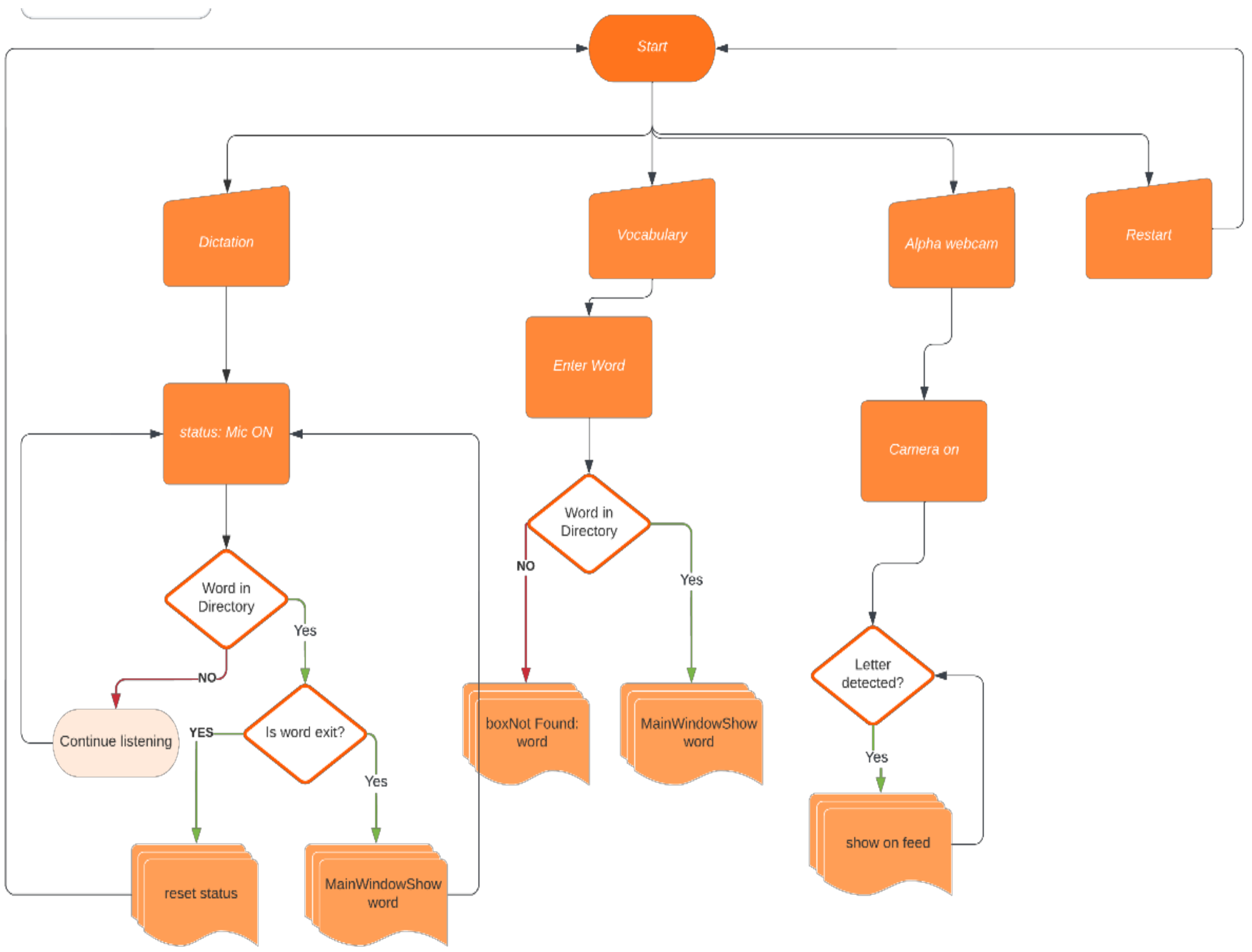

Figure 19: The Flowchart of the System Main Functions Excluding the Spell Feature, which is Shown in Figure 25

4.1 Dictation

ASR is a "machine-based process of decoding and transcribing oral speech" (Levis & Suvorov, 2012, p. 1) that's built into many technologies such as telebanking and customer services. Google ASR API offers a substantial increase in accuracy when it comes to speech recognition among its peers [6][33]. The process for the dictation tool is shown in Figure 20. Clicking on the dictation button activates the feature as the microphone turns on and waits for voice input, the Mic status changes to MIC ON. The input is then interpreted to the Arabic text, which shows on the dialog box "you said: +Arabic Text". The system tags the input and searches for the tag in the Signs directory. If the tag is found in the directory, the system searches in the signs' recordings database to fetch the corresponding sign. Once fetched, the system plays the corresponding sign in the logo frame replacing the logo with the sign, as shown in Figure 21.

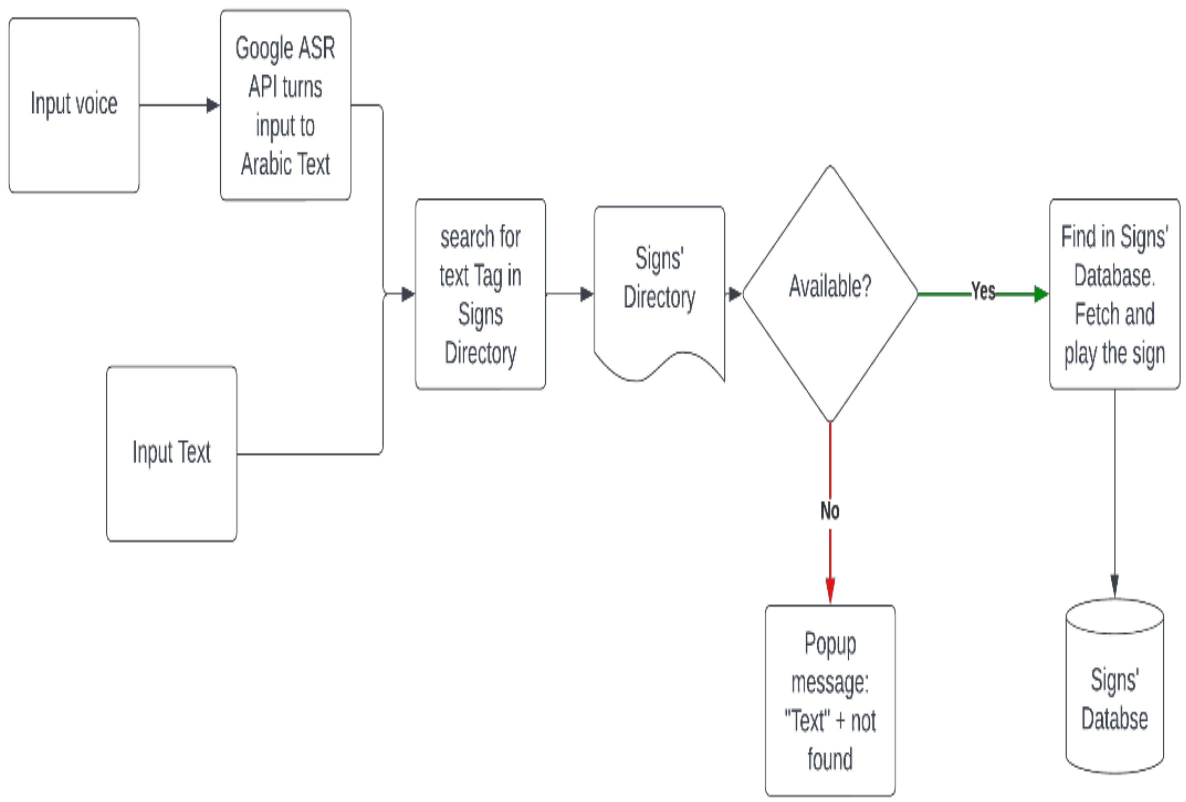

Figure 20: Process for the Dictation Feature and the Vocabulary Feature

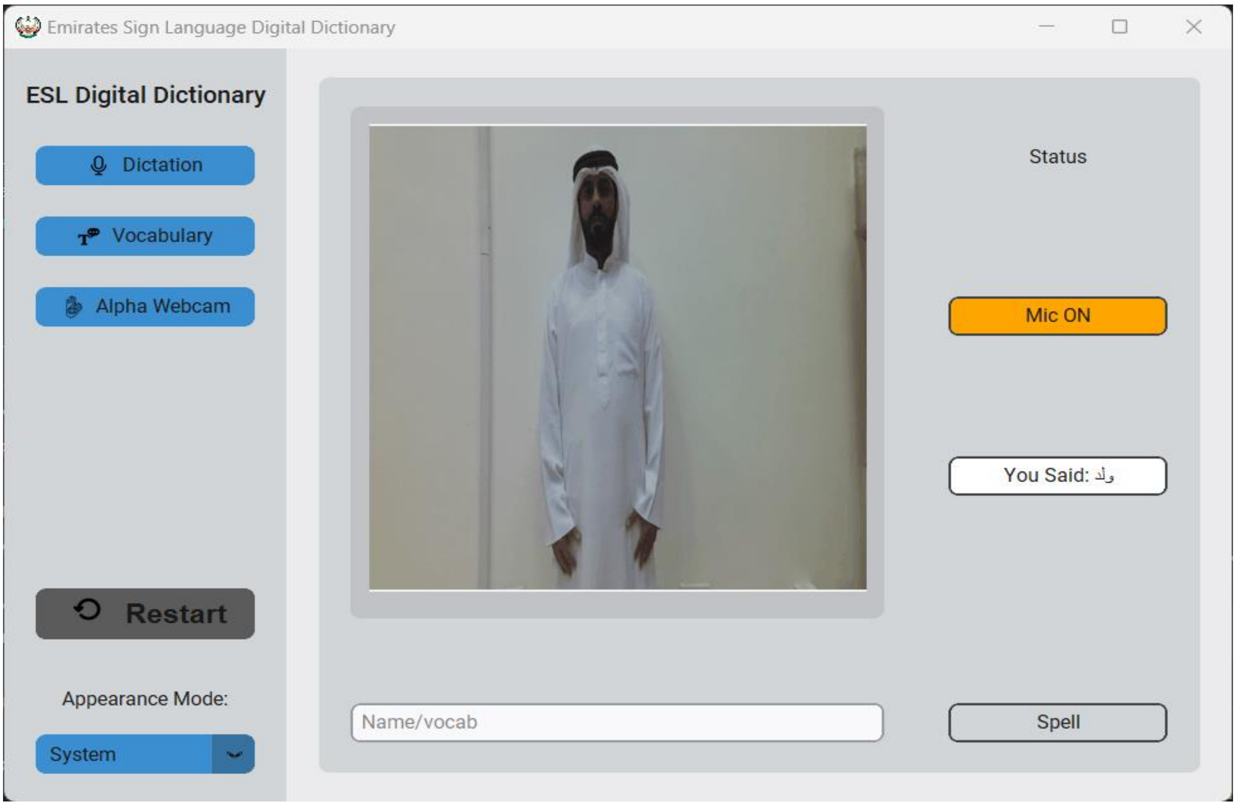

Figure 21: When Dictation is On, the Status Monitor for the Mic Turns Orange with ON Status and the Voice Input is Passed to the Dictation Box to Show in Text and the Logo is Replaced by the Interpretated Sign

4.2 Vocabulary

The Vocabulary feature allows the user to search for signs in a text fashion, where the user inputs the word in the entry box shown in Figure 22, and the system fetches and plays the corresponding sign if found.

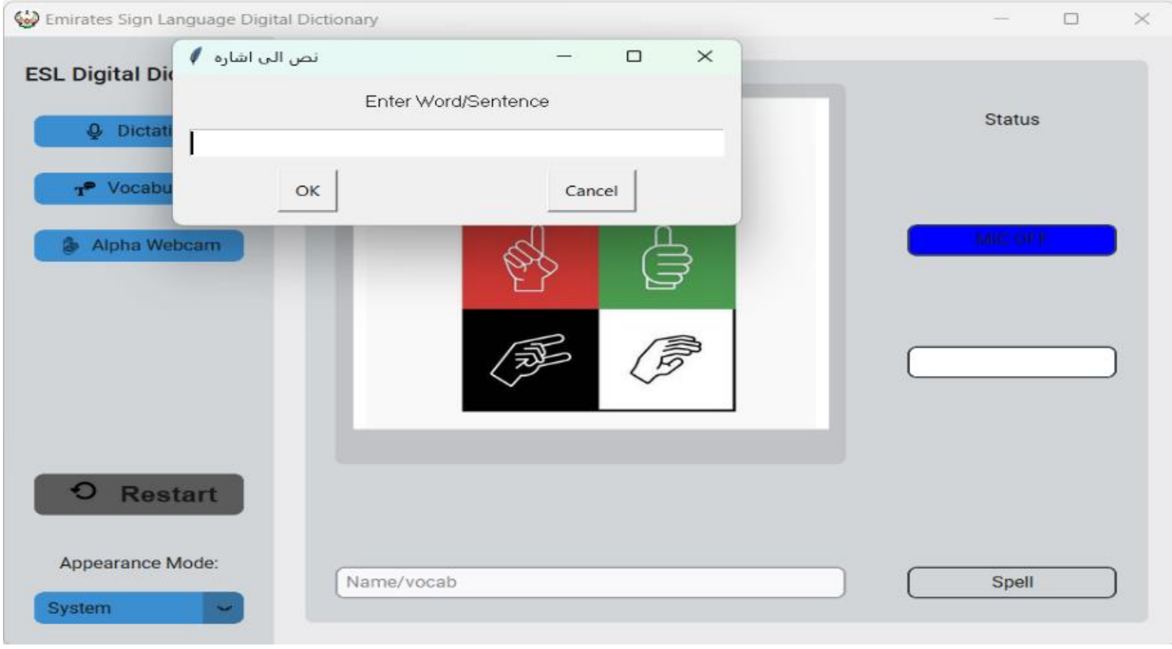

Figure 22: Vocabulary Entry Box has a Text Input Field where the User Can Look Up Vocabulary and Sentences in the Sign Language

4.3 Alpha Webcam

The YOLOv8 model is a state-of-the-art machine learning and object detection model to detect objects and recognize them [7]. The model used in the ESL E-Dictionary, shown in Figure 23, achieved the COCOmAP of 0.86. The feature mentioned detects signs demonstrated via the webcam and shows how accurately the sign is being performed. So far, the Alpha Webcam only works on Arabic/Emirate SL alphabets. Dataset samples are shown in Figure 24.

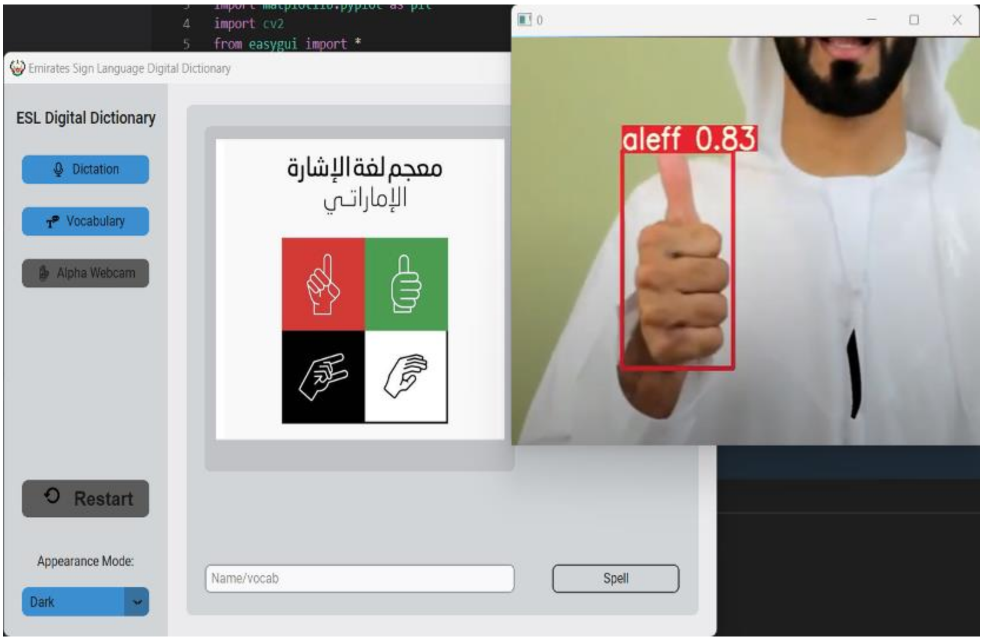

Figure 23: Alpha Webcam Feature, where the Device Webcam Pop-Up Window will be Called Once Clicked on the Alpha Webcam Button

Figure 24: Arabic/Emirate Sign Language Alphabets from ArSL21L [8]

4.4 Spell

The flowchart of the spell feature is shown in Figure 25. The user enters the word or name to be spelled in the Alphabet of the ESL, in the entry box, and the entry is first checked whether the entry is valid/empty then if not, its entry is processed, and corresponding letters are shown to the user sequentially.

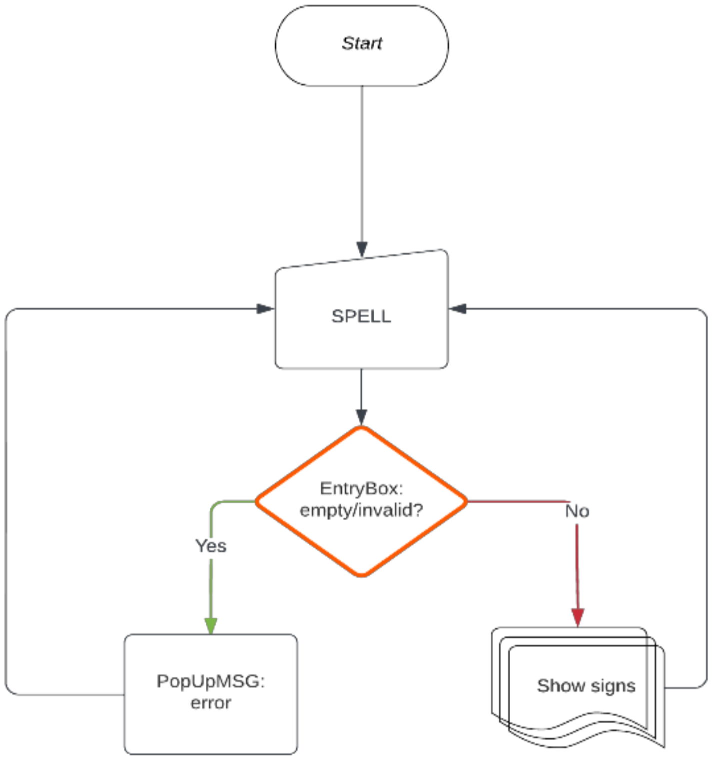

Figure 25: Spell Feature Flowchart

Chapter 5: ArSL21L Benchmark on YOLO Latest Models

ArSL21L [8], shown on Figure 24, has 14202 images of 32 signs that were signed by 50 signers, was evaluated on YOLOv5 [8], which led to the results shown in Table 3 YOLOv5l scored the highest $mAP_{0.5:0.95}$ scoring 0.8306 among its peers. The aim is to compare the results obtained by [8] to the latest models available YOLOv7 [34] and YOLOv8 [7]. The data was split in a similar manner to [8] with 9955 images for the training set by the random selection of 35 signers. And for the testing, 4247 images by the remaining 15 signers. Finally, the learning rate was kept the same at 0.001 Adam Optimizer for 300 epochs.

Table 3: Results on ArSL21L Dataset [8]

Model	Precision	Recall	$mAP_{0.5}$	$mAP_{0.5:0.95}$
YOLOv5s	0.953	0.9408	0.9784	0.7661
YOLOv5m	0.968	0.9468	0.9842	0.7768
YOLOv5l	0.9787	0.9766	0.9909	**0.8306**
YOLOv5x	0.9758	0.9743	0.9896	0.8224

5.1 YOLOv7

The ArSL21L dataset was evaluated on YOLO model version 7 [34] which is one of the latest models of YOLO and object detection models up to this date. Table 4 tabulate the performance evaluation via Precision, Recall, mAP metrics of both versions of the YOLOv7 models. Both versions were trained for 300 epochs at first, but they have not convergent. Therefor we increased the number of epochs to 600 epochs in the training stage which had both versions convergent at around 500 epochs for the classification loss metrics. Shown in Figure 26. There are three types of loss, the box loss, objectness loss, and classification loss. The box loss refers to how accurately the algorithm shall pin the center of an object and how good the foreseen box covers the object [35]. Objectness is a metric for the probability a region of interest contains an object [35]. A high objectness means a high probability of an object to be in the image window. Classification loss show metrics for how accurately the algorithm can spot the correct object [35]. YOLOv7 [34], the standard model for 640 image size, scored 0.8228 $mAP_{0.5:0.95}$ and 0.9909 $mAP_{0.5}$. On the contrary, YOLOv7x scored a $mAP_{0.5:0.95}$ score of 0.8305 and a $mAP_{0.5}$ score of 0.9914.

In other words, YOLOv7x has outperformed the standard YOLOv7 model. Table 5 shows the evaluation metrics for all 32 classes that are in the ArSL21L for the YOLOv7x model, as well as Figure 27 shows the confusion matrix of the YOLOv7x model during testing stage.

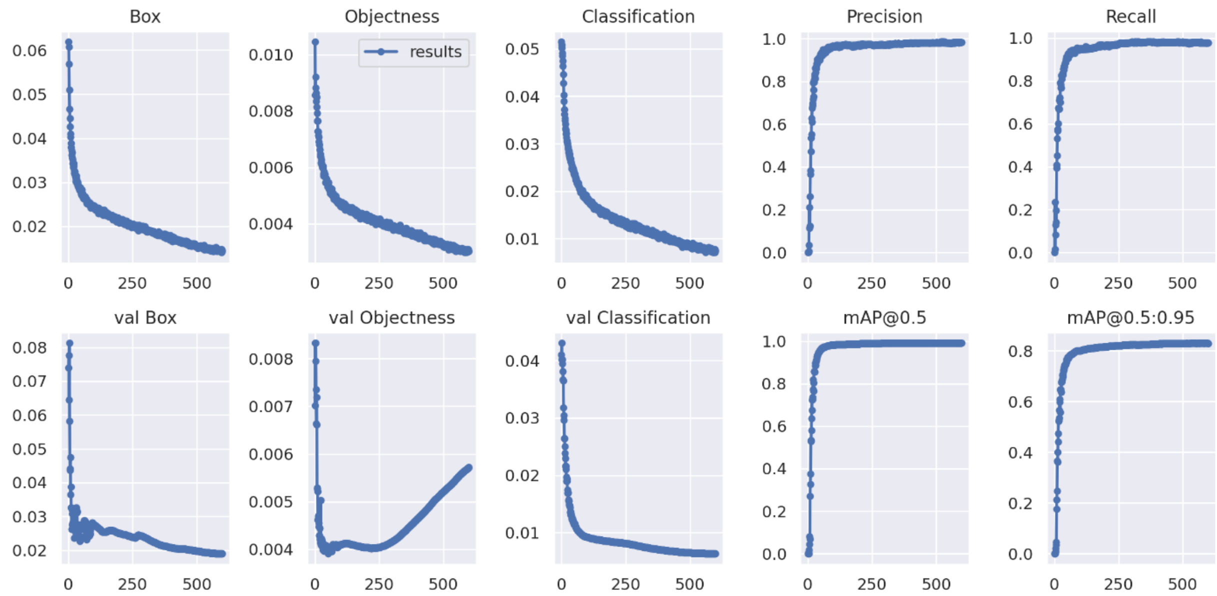

Figure 26: YOLOv7x Plots of Box Loss, Objectness Loss, Classification Loss, Precision, Recall and mean Average Precision (mAP) During the Training, where the X-Axis Refers to the Number of Epochs

Table 4: Results of YOLOv7 Models on ArSL21L.

Model	**Precision**	**Recall**	**mAP0.5**	**mAP0.5:0.95**
YOLOv7	0.9803	0.9832	0.9909	0.8286
YOLOv7x	0.9857	0.9821	0.9914	**0.8305**

Table 5: Evaluation Results of YOLOv7x for 32 Signs

Class	Labels	P	R	mAP0.5	mAP0.5: 0.95	Class	Labels	P	R	mAP 0.5	mAP0.5: 0.95
Ain	133	0.999	0.985	0.996	0.809	Laam	132	0.992	0.944	0.994	0.859
Al	136	0.991	0.993	0.993	0.915	Meem	135	1	0.98	0.996	0.843
Aleff	136	0.999	0.993	0.997	0.892	Nun	123	0.992	0.976	0.994	0.833
Bb	137	0.991	1	0.996	0.892	Ra	123	0.984	0.999	0.992	0.788
Dal	111	0.969	0.937	0.982	0.792	Saad	135	0.977	0.941	0.992	0.798
Dha	135	0.992	0.993	0.996	0.842	Seen	135	0.999	0.985	0.996	0.882
Dhad	130	0.979	0.969	0.983	0.84	Sheen	135	0.999	1	0.997	0.888
Fa	135	0.92	0.936	0.972	0.811	Ta	135	0.97	1	0.995	0.854
Gaaf	134	0.977	0.955	0.992	0.817	Taa	135	0.993	0.989	0.996	0.803
Ghain	135	1	1	0.996	0.859	Thaa	137	0.995	0.978	0.996	0.858
Ha	135	0.998	0.993	0.996	0.808	Thal	135	0.988	0.948	0.985	0.801
Haa	135	0.957	0.983	0.987	0.747	Toot	135	1	0.989	0.996	0.834
Jeem	135	0.992	0.952	0.991	0.781	Waw	121	0.967	0.97	0.981	0.799
Kaaf	135	0.977	1	0.994	0.884	Ya	134	0.947	0.978	0.992	0.833
Khaa	135	1	0.982	0.996	0.796	Yaa	135	0.973	1	0.997	0.814
La	135	1	0.995	0.996	0.868	Zay	135	0.947	0.935	0.947	0.729
					Average		4252	0.983	0.977	0.991	**0.83**

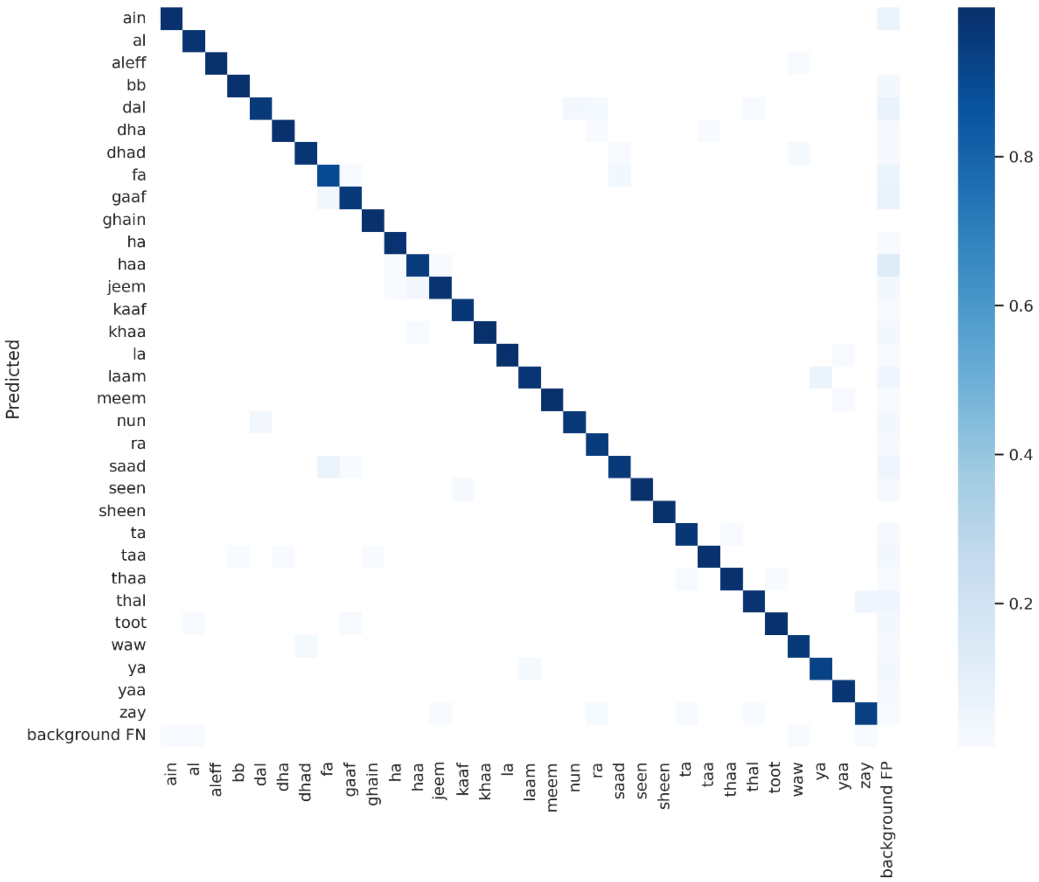

Figure 27: Confusion Matrix of YOLOv7x where Vertical Axis Refers to Predicted Results and Horizontal Axis Refers to Ground Truths

5.2 YOLOv8

YOLOv8 [7] is the latest, state-of-the-art, model in the YOLO series, up to this date, currently it has five versions, which all are models aimed at 640 pixels datasets. It was designed to boost performance and accuracy [7]. Table 6 summarizes the attributes of the five YOLOv8 models.

Table 6: YOLOv8 Versions Summary [7]

Model	size (pixels)	mAPval 50-95	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)	FLOPs (B)
YOLOv8n	640	37.3	80.4	0.99	3.2	8.7
YOLOv8s	640	44.9	128.4	1.20	11.2	28.6
YOLOv8m	640	50.2	234.7	1.83	25.9	78.9
YOLOv8l	640	52.9	375.2	2.39	43.7	165.2
YOLOv8x	640	53.9	479.1	3.53	68.2	257.8

5.2.1 Training

All five versions of YOLOv8 were trained for 300 epochs, Table 7 summarizes how they performed. With YOLOv8x outperforming the rest of the models with $mAP_{0.5:0.95}$ score of 0.86077.

Table 7: Results of YOLOv8 Model Variations Training on ArSL21L

Model	Precision	Recall	mAP0.5	mAP0.5:0.95
YOLOv8n	0.97992	0.97677	0.99031	0.83984
YOLOv8s	0.9828	0.97845	0.98975	0.84669
YOLOv8m	0.98375	0.9821	0.99171	0.85687
YOLOv8l	0.98496	0.98287	0.99035	0.85815
YOLOv8x	0.98561	0.98269	0.99036	**0.86077**

5.2.2 Performance evaluation

Focal loss allows the model to focus on hard misclassified examples, as the confidence of predicting the class correctly increases, the scaling factor of focal loss decays to zero [36].

In the validation stage the model showed great improvements in terms of precision, recall, and mean average precision, mAP, until it started to plateau at around 100 epochs, shown in Figure 28. As well as, the box, focal, and classification losses of the validation stage also showed a quick fall until around 100 epochs where it started to plateau as well. We utilized early stopping to determine the best weights.

Post training YOLOv8x, we introduced the new and unseen set of the dataset, to make predictions, examples shown in Figure 29.

The results of the testing of YOLOv8x are tabulated in Table 8. As can be observed YOLOv8x, which scored $mAP_{0.5:0.95}$ score of 0.86, has outperformed all previous models, the YOLOv5l, which the ArSL21l was originally tested on, scored $mAP_{0.5:0.95}$ of 0.83 [8], and YOLOv7x, which scored a $mAP_{0.5:0.95}$ score of 0.83 as well, at 600 epochs. Figure 30 shows the confusion matrix of YOLOv8x.

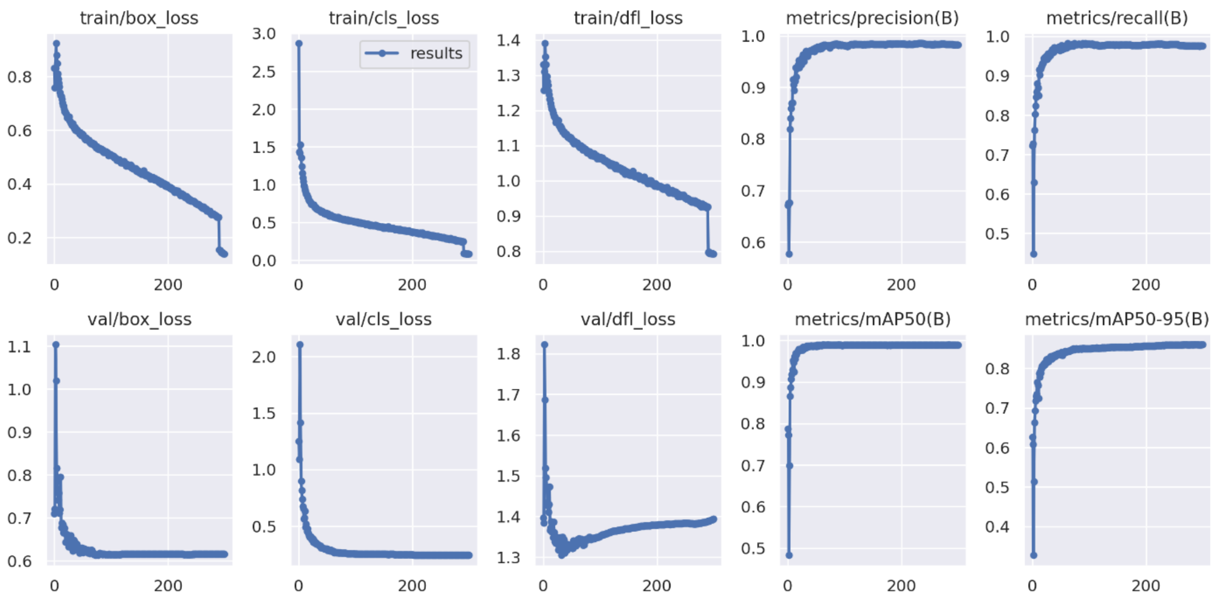

Figure 28: YOLOv8x Plots of Box Loss, Classification Loss (CLS), Distribution Focal Loss (DFL), Precision, Recall and mean Average Precision (mAP) During the Training where the X-Axis Refers to the Number of Epochs

Figure 29: Examples from the Test Dataset Showing the Performance of YOLOv8x

Table 8: Evaluation Results of YOLOv8x for 32 Signs

Class	Labels	P	R	mAP0.5	mAP0.5 :0.95	Class	Labels	P	R	mAP0.5	mAP0.5 :0.95
ain	133	1	0.981	0.995	0.852	laam	132	0.984	0.947	0.992	0.871
al	136	0.992	0.993	0.989	0.93	meem	135	1	0.983	0.995	0.868
aleff	136	1	0.972	0.995	0.915	nun	123	0.971	0.992	0.994	0.868
bb	137	1	0.991	0.995	0.92	ra	123	0.998	0.992	0.995	0.849
dal	111	0.99	0.893	0.977	0.823	saad	135	0.985	0.963	0.994	0.837
dha	135	0.992	0.993	0.994	0.879	seen	135	0.997	0.963	0.985	0.889
dhad	130	0.96	0.962	0.986	0.867	sheen	135	1	0.995	0.995	0.922
fa	135	0.963	0.948	0.99	0.85	ta	135	0.974	1	0.994	0.884
gaaf	134	0.992	0.967	0.994	0.862	taa	135	1	0.981	0.995	0.858
ghain	135	0.994	1	0.995	0.875	thaa	137	1	0.976	0.995	0.885
ha	135	0.978	0.993	0.994	0.856	thal	135	0.984	0.939	0.975	0.838
haa	135	0.97	0.993	0.975	0.751	toot	135	1	0.994	0.995	0.872
jeem	135	0.991	0.963	0.989	0.806	waw	121	0.967	0.958	0.984	0.823
kaaf	135	0.963	1	0.983	0.895	ya	134	0.957	0.978	0.991	0.871
khaa	135	0.985	0.993	0.995	0.811	yaa	135	0.978	0.993	0.994	0.86
la	135	0.999	1	0.995	0.899	zay	135	0.948	0.946	0.959	0.758
					Average		4252	0.985	0.976	0.99	**0.861**

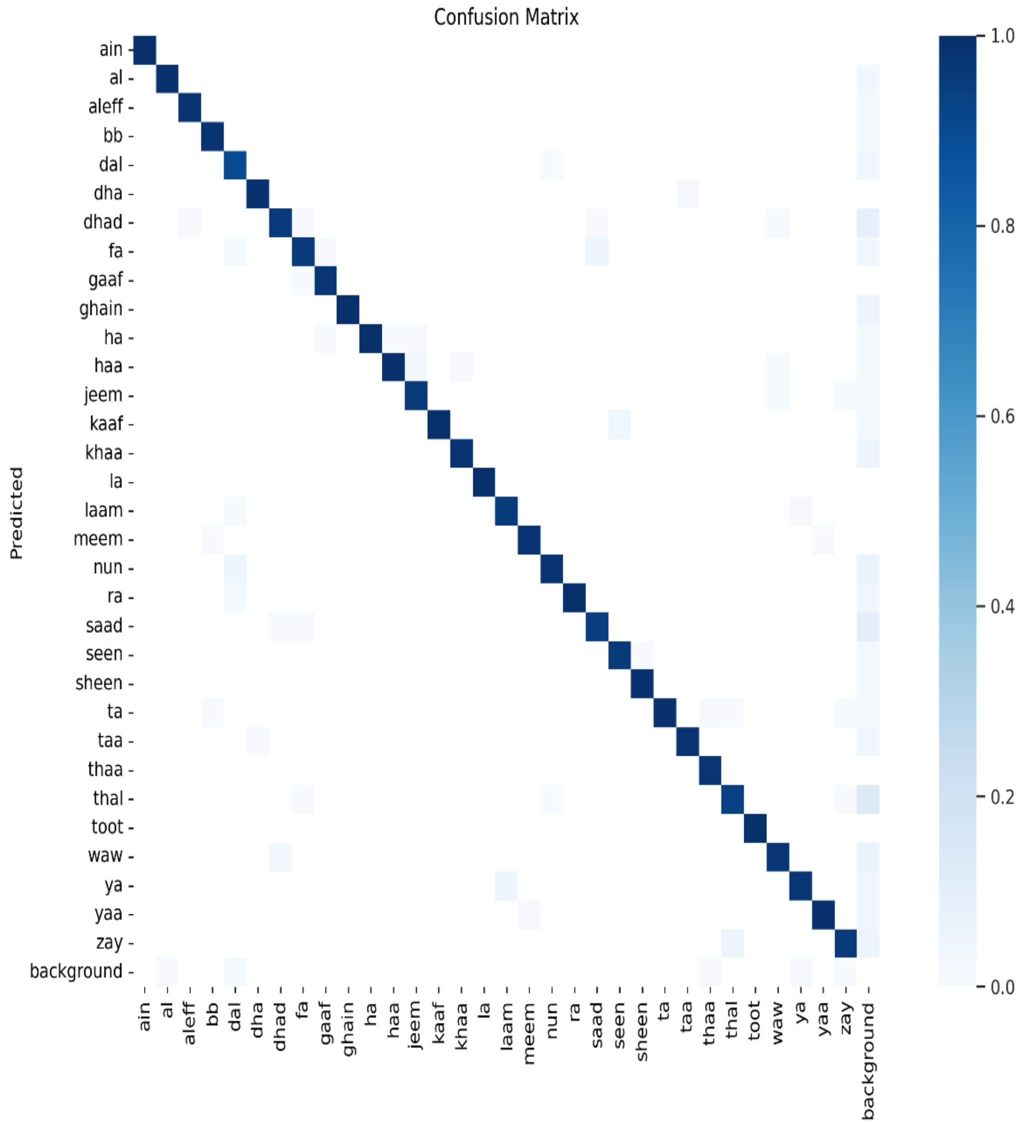

Figure 30: Confusion Matrix of YOLOv8x where Vertical Axis Refers to Predicted Results and Horizontal Axis Refers to Ground Truths During Testing

Chapter 6: Performance Evaluation

Evaluating a Multiway-to-ESL is difficult due to the absence of any proper methodology. Hence, to evaluate our system, we gathered 40 people that are by any means related to the Emirate Sign Language (ESL) field. The volunteers are between the age of 12 to 40 years old, which include 8 university students, 2 professors, and 14 Emirate schools' staff (teachers and administration staff). 6 hearing-impaired individuals, 3 of the 6 are employees and the remaining 3 are students in a specialized school for hearing-impairment. 2 certified Emirate Sign Language interpreters. Finally, 8 customer support employees. We had the volunteers create a list of words and phrases they would use in their daily interactions or common to use in their line of work. We cross-referenced the lists and matched them to the ESL e-Dictionary Dataset, removing any sort of dialectics of the Arabic language. Finally, Matched sentences were then given back to the volunteers to interact with the system. The device used is a mid-level laptop with the following specifications:

- Processor: Intel(R) Core (TM) i7-10750H CPU @ 2.60GHz 2.59 GHz.
- RAM: 16.0 GB.
- System type: 64-bit operating system, x64-based processor.
- GPU: Nvidia GeForce GTX 1650 4 GB.

6.1 Dictation Performance

Automatic Speech Recognition (ASR) for Arabic is very tricky since most Modern Standard Arabic (MSA) data that is used by the ASR community is collected from Broadcast News (BN) and Broadcast Conversations (BC) [37]. Therefore, to minimize Word Error Recognition (WER) for dictation, we had the volunteers speak closely to the mic and in a clear environment to minimize background noise and mimic BC and BN environment. Speech is then transferred into text which is then looked up in the ESL e-Dictionary directory. However, Google strips the dictated speech from punctuation and dictate speech as pronounced, as shown in Table 9, the directory had to match the ASR by Google algorithm dictation syntax, for the system to be able to fetch the correct sign.

Table 9: Voice Dictation and ESL e-Dictionary Directory.

Speech	ASR API by google dictation	ESL e-Dictionary Directory
ألوان	الوان	الوان
برج خليفة	برج خليفه	برج خليفه

6.2 Text-to-Sign Performance

Vocabulary is a feature that allows the user to look up signs via Text-to-Sign means. But since the Arabic language has many variations and it is a dialect type of language, we asked the volunteers to avoid using punctuation and to type the text as they pronounce it. Some of the volunteers found it difficult in means of removing punctuations and spelling words as they are pronounced, e.g., "شهادة" is an Arabic word that means a certificate, if such a word is entered, the system shall notify the user "word not found". Since the signs directory is written matching the Google ASR API dictation of the Arabic language. Therefore, spelling the word as it's pronounced, "شهاده", would result for the word to be found and the correct sign demonstrated.

6.3 Response Performance

A real-time system is a system that responds within a short time [38]. Hence, the most dominant factor of our system is: timeliness. It's expected from a real-time system to return results within milliseconds. To evaluate the response performance of our system, we averaged the response time of when the system received the command until it has fetched the corresponding sign.

6.3.1 Dictation response time

10 random voice commands were given to the system consequently, the response time was recorded at two different stages. The First stage (A) is as follows: from the system being ready to accept speech to the system identifying the corresponding sign from the ESL e-Dictionary directory. The Second stage (B): from the system receiving the dictation from the ASR API to the system showing the first frame of the corresponding sign.

Table 10 shows the commands given and the average response time of the system in both stages. The ESL e-Dictionary's average response time in stage (A) is 1.564546 seconds, and in stage (B) is 0.268299.

Table 10: Average Dictation Response Time

Command No.	Command	Stage (A) Response Time	Stage (B) Response Time
1st	ولد	1.19177	0.27004
2nd	عمه	2.03029	0.26458
3rd	برج خليفه	2.09917	0.27658
4th	حارس	1.24846	0.26190
5th	جميرا	1.65654	0.27905
6th	طبيب	1.65602	0.26571
7th	محامي	1.12132	0.27026
8th	بنت	2.42882	0.26272
9th	جده	1.10891	0.26554
10th	جد	1.10416	0.26661
Average Response Time		**1.564546**	**0.268299**

6.3.2 Text-to-Sign response time.

To measure the response time of the vocabulary feature, we shall calculate the time it takes from, receiving the request to translate to the system showing the first frame of the corresponding sign. 10 random words were used, as shown in Table 11, and the average response time of the system is 0.36415 seconds.

Table 11: Average Text-to-Sign Response Time

Text No.	Text	Response Time
1st	ولد	0.35869
2nd	برج العرب	0.331286
3rd	برج خليفه	0.3611937
4th	حارس	0.3722663
5th	جميرا	0.3522024
6th	شهاده ميلاد	0.333455
7th	طباخ	0.498286
8th	بنت	0.340995
9th	جده	0.340996
10th	الوان	0.3522026
Average Response Time		**0.3641573**

Chapter 7: Discussions

The Arabic language is a diglossia language [2] with very complex syntaxes. Moreover, the Arabic Language is a dialect type of language, e.g., an Arabic speaker of a different nationality might use terminologies and gestures that's hard for an individual of a different Arab nationality to understand.

During the data collection process, as hearing impairment individuals were interacting, we observed that few gestures only is needed to make up a sentence. And we believe this is how the Sign Language community overcomes the very hard complexity of the Arabic language syntax. In other words, a single sign can refer to many things depending on the context of the conversation. For example, the sign "wants" the meaning of it depends solely on the context, as in some scenarios it meant "I want" while in others it meant "I need" or "get".

Hence, in the system we removed syntax and punctuation, and stripped the directory of any context. For instance, for a user to get the signing of "I want a translator", the Arabic language syntax would have the signer signing 4 signs, as shown in Figure 31: "I", "want", "A person", and "to translate". However, in real life scenarios the signer would sign only 2 signs, as shown in Figure 32: "want", and "translate", and depending on the context of the conversation the hearing impaired would understand.

Therefor during the testing of the system, we had users narrow down their sentences and communication to the key words of what they are trying to communicate to the hearing impairment individual using words from the directory list of the ESL e-Dictionary.

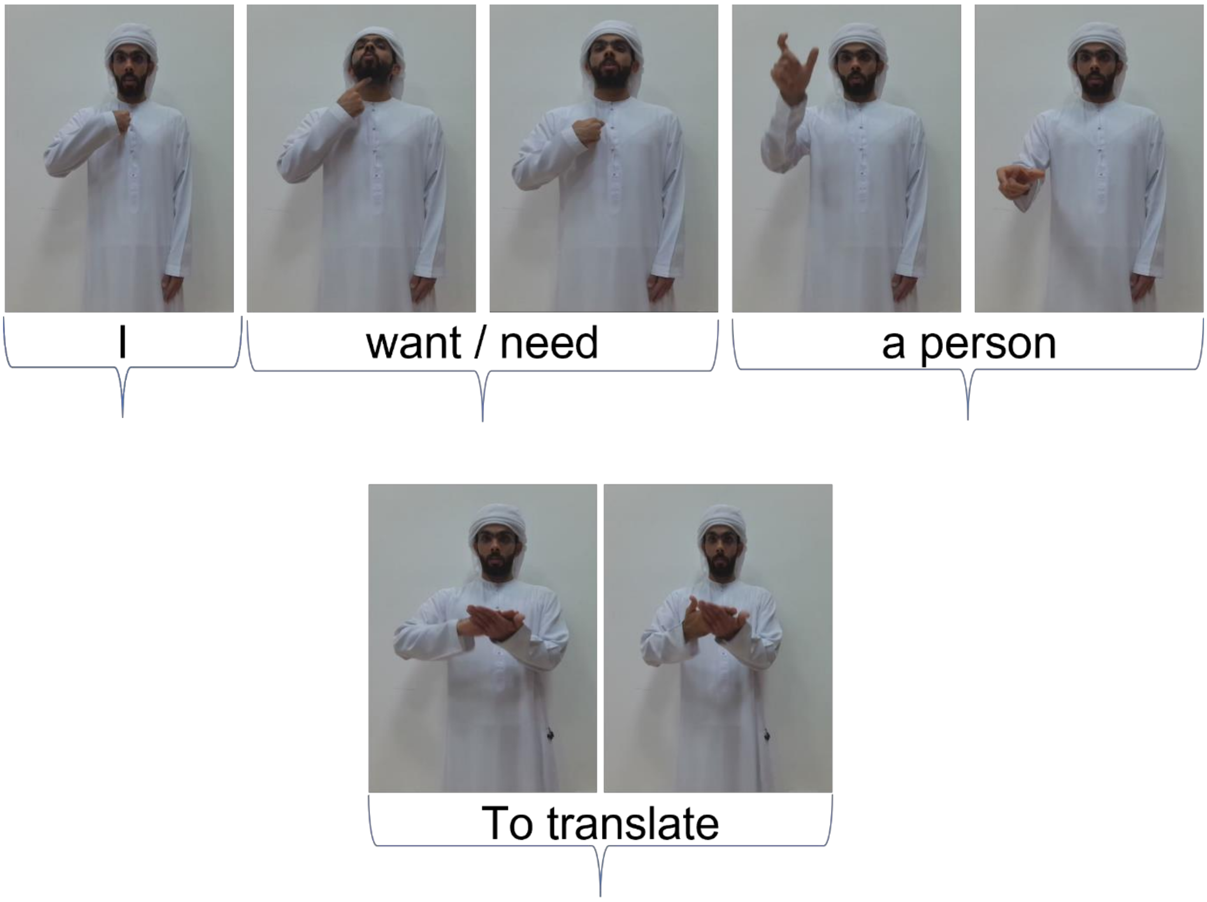

Figure 31: The Expected Signing of the Sentence "I want a translator"

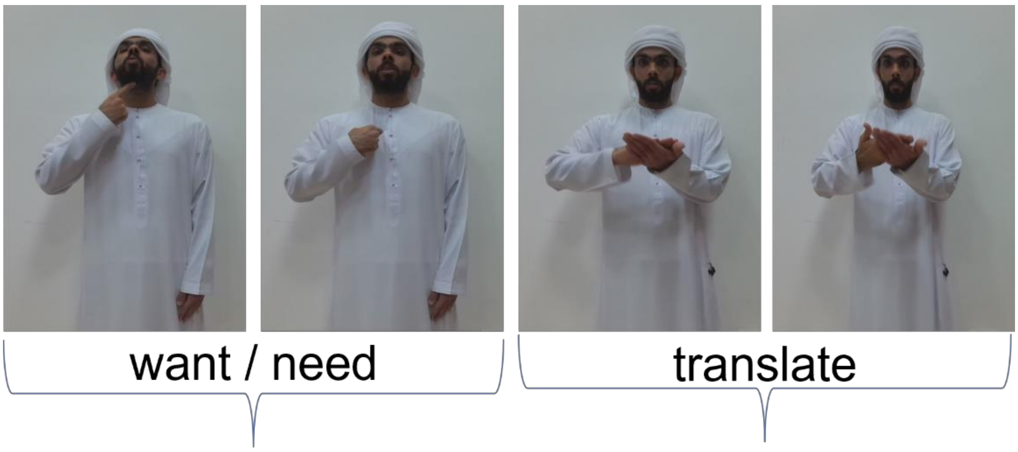

Figure 32: Real Life Signing Scenario of the Sentence "I want a translator"

Moreover, it was observed that the signers always emphasize the key gestures of a sign, to be able to differentiate between two signs that have almost the same demonstration. Figure 33 shows key gestures for the color brown sign, and Figure 34 show the key gestures for the color yellow.

As can be observed, there are many similarities between the two signs, where the main gesture is to get your hand to your nose, indicated by K1 and K2, the gesture that follows shall determine which sign the signer is trying to sign. The signer in Figure 34 repeats the key gestures K2 and K3 one more time to differentiate between the color brown and yellow in the signing process. This was a common practice among all signers in many signs.

But we chose to neglect that emphasis on key gestures as we were following the Zayed Higher Organization (ZHO) of people of determination Sign Dictionary.

Figure 33: Key Gesture, From Left to Right K1, K2…K4, for Color Brown Sign

Figure 34: Key Gestures, From Left to Right K1, K2…K6, for Color Yellow Sign

Chapter 8: Conclusions and Future Works

Automatic Speech Recognition (ASR) processing APIs and algorithms require a huge amount of data to achieve high accuracy, APIs like Google ASR API, and Microsoft Azure ASR API, which constantly collect user data, provide to some extent a suitable ASR capability. Yet, we can't always ensure the level of accuracy of the API, due to the huge amount of variation that occurs while pronouncing a word. Hence, the dictation feature of most systems powered by those APIs remain in the hands of those companies in terms of accuracy and processing speed. Moreover, for the system to be usable in dictation mode we had to limit voice inputs to as few words as possible in an input.

Secondly, the Arabic Language, as mentioned, is of a dialect nature, which would lead the APIs to misrecognize the speech, which would result in the system failing to interpret the input. Furthermore, attempting to collect as many sentences and phrases as possible is an impossible task and word-by-word processing can seem like the most convenient way for now.

Moreover, since the Arabic language is a diglossia language [2] with high level of complexity in its syntaxes. We face the issue of words that are of noun nature to be used as verbs or adjectives depending on the context of the conversation, therefor in our system any word to be inputted, text or voice, must be returned to its original form which is a noun. For instance, the word "Walk" the original form is "مشي" which is a noun, but it has many other forms in the Arabic language to be used as a verb. Hence, our system only has the origin of the word, and the understanding of the sign being a verb, or a noun is referred back to the context of the conversation.

Finally, signs can differ from one region to another even inside the same country, for example, a hearing-impaired individual in Sharjah/UAE would sign the word "lying" differently than a hearing-impaired in Dubai/UAE. As well as the instructors found in the schools teaching Sign Language can be influenced by the region or the signs of their birth country if they were expats. Therefore, following the Zayed Higher Organization (ZHO) of people of determination official ESL dictionary was the best attempt to establish the ESL e-Dictionary. However, signs in ZHO are being constantly added, removed, and swapped, therefore constant follow-ups and updates are required to ensure the correctness of the ESL e-Dictionary.

Further work can be done by enhancing the dataset of the ESL e-Dictionary in terms of increasing its size and exploring in the direction of the few-shot learning algorithm to accomplish a real time continues Sign Language interpreter. Moreover, creating our own ASR API using the Arabic version of UAE would reduce the Word Error Rate (WER) in speech dictation. As well as searching in the context understanding of the Arabic language would improve our system in term of reducing WER and creating a real time continues Sign Language interpreter for the Arabic language.

References

[1] World Health Organization, “Deafness and Hearing Loss,” Fact Sheets, 1st of April 2022, [Online], Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss [Accessed January 8, 2023].

[2] Center for Strategic and International Studies, “Reading the Signs: Diverse Arabic Sign Languages,” [Online], Available: https://www.csis.org/analysis/reading-signs-diverse-arabic-sign- languages [Accessed January 8, 2023].

[3] M. A. Abdel-Fattah, “Arabic Sign Language: A Perspective,” *Journal of Deaf Studies and Deaf Education*, vol. 10, No. 2, pp. 212-221, 2005.

[4] W. Aditya, T. Shih, T. Thaipisutikul, A. Fitriajie, M. Gochoo, F. Utaminingrum, and C. Lin, “Novel Spatio-Temporal Continuous Sign Language Recognition Using an Attentive Multi-Feature Network,” *Sensors*, vol. 22, no. 17, pp. 6452, 2022.

[5] W. Suliman, M. Deriche, H. Luqman, and M. Mohandes, “Arabic Sign Language Recognition Using Deep Machine Learning,” *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, pp. 1-4, 2021.

[6] S. McCrocklin, A. Humaidan, and I. Edalatishams, “ASR Dictation Program Accuracy: Have Current Programs Improved,” *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, pp. 191-200, 2019.

[7] G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by Ultralytics (Version 8.0.0),” [Computer software], Available: https://github.com/ultralytics/ultralytics [Accessed December 20, 2022].

[8] G. Batnasan, M. Gochoo, M. -E. Otgonbold, F. Alnajjar, and T. K. Shih, “ArSL21L: Arabic Sign Language Letter Dataset Benchmarking and an Educational Avatar for Metaverse Applications,” *2022 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1814-1821, 2022.

[9] M. El-Gayyar, A. Ibrahim, and M. E. Wahed, “Translation from Arabic speech to Arabic Sign Language based on cloud computing,” *Egyptian Informatics Journal*, vol. 17, pp. 295-303, 2016.

[10] A. Alfi, M. El Basuony, and S. El Atawy, “Intelligent Arabic Text to Arabic Sign Language Translation for Easy Deaf Communication,” *International Journal of Computer Applications*, vol. 92, pp. 22-29, 2014.

[11] N. Aouiti and M. Jemni, “For a Translating System from Arabic Text to Sign Language,” *Universal Learning Design: Proceedings of the Conference Universal Learning Design*, pp. 33-38, 2014.

[12] S. Halawani and A. Zaiton, “An Avatar-Based Translation System from Arabic Speech to Arabic Sign Language for Deaf People,” *International Journal of Information Science Education*, pp. 13-20, 2012.

[13] A. Almohimeed, M. Wald, and R. Damper, “Arabic Text to Arabic Sign Language Translation System for the Deaf and Hearing-Impaired Community,” *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pp. 101-109, 2011.

[14] H. Luqman and S. Mahmoud, “Automatic Translation of Arabic Text to Arabic Sign Language,” *Universal Access in the Information Society*, pp. 939-951, 2018.

[15] K. Al-Fityani and C. Padden, “Sign Language Geography in the Arab World,” pp. 433-450, 2010.

[16] A. Othman and O. El Ghoul, “Intra-Linguistic and Extra-Linguistic Annotation Tool for the “Jumla Dataset” in Qatari Sign Language,” *8th International Conference on ICT & Accessibility (ICTA)*, pp. 01-06, 2021.

[17] Wikipedia, “Emirate Sign Language,” Article, 19th of December 2022, [Online], Available: Https://en.wikipedia.org/wiki/Emirati_Sign_Language [Accessed January 10, 2023].

[18] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, “Towards Real-time Multi-Object Tracking,” *Computer Vision–ECCV 2020: 16th European Conference Proceedings*, pp. 107-122, 2020.

[19] D. Feng, A. Harakeh, S. Waslander, and K. Dietmayer, “A Review and Comparative Study on Probabilistic Object Detection in Autonomous Driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 9961-9980, 2022.

[20] E. Martinez-Martin and A. P. del Pobil, “Object Detection and Recognition for Assistive Robots: Experimentation and Implementation,” *IEEE Robotics and Automation Magazine*, vol. 24, no. 3, pp. 123-138, 2017.

[21] COCOdataset, “Common Objects in Context challenge,” [Online], Available: https://cocodataset.org/#guidelines [Accessed January 10, 2023].

[22] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully Convolutional One-Stage Object Detection,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627-9636, 2019.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[24] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *Proceedings of IEEE Conference in Computer Vision and Pattern Recognition*, pp. 7263-7271, 2017.

[25] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Lawrence, "Microsoft COCO: Common Objects in Context," *Computer Vision–ECCV 2014: 13th European Conference Proceedings*, pp. 740-755, 2014.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.

[27] A. Bochkovskiy, C. Wang, and H. Mark, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv preprint arXiv:2004.10934, 2020.

[28] C. Careaga, B. Hutchinson, N. Hodas, and L. Phillips, "Metric-based Few-Shot Learning for Video Action Recognition," arXiv preprint arXiv:1909.09602, 2019.

[29] J. Carreira, E. Noland, A. Bank-Horvath, C. Hillier, and A. Zisserman, "A Short Note about Kinetics-600," arXiv preprint arXiv:1808.01340, 2018.

[30] Z. Zhu, L. Wang, S. Guo, and G. Wu, "A Closer Look at Few-Shot Video Classification: A New Baseline and Benchmark," arXiv preprint arXiv:2110.12358, 2021.

[31] Y. Bo, Y. Lu, and W. He, "Few-Shot Learning of Video Action Recognition Only Based on Video Contents," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 595-604, 2020.

[32] M. Gygli, S. Yale, and C. Liangliang, "Video2gif: Automatic Generation of Animated Gifs from Video," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1001-1009, 2016.

[33] A. Abdellatif, K. Badran, D. E. Costa, and E. Shihab, "A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering," *IEEE Transactions on Software Engineering*, vol. 48, no. 8, pp. 3087-3102, 2022.

[34] C. Wang, A. Bochkovskiy, and H. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-art for Real-time Object Detectors," arXiv preprint arXiv:2207.02696, 2022.

[35] M. Kasper-Eulaers, N. Hahn, S. Berger, T. Sebulonsen, Ø. Myrland, and P. E. Kummervold, "Short Communication: Detecting Heavy Goods Vehicles in Rest Areas in Winter Conditions Using YOLOv5," *Algorithms*, vol. 14, no. 4, pp. 114, 2021.

[36] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980-2988, 2017.

[37] F. Biadsy, P. J. Moreno, and M. Jansche, "Google's Cross-Dialect Arabic Voice Search," *2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4441-4444, 2012.

[38] J. E. Cooling, *Software Design for Real-Time Systems*, London: International Thomson Computer Press, pp. 2-15, 2013.

UAE UNIVERSITY MASTER THESIS NO. 2023:23

Unaddressed hearing loss can impact many aspects of life at an individual level and societal level. And the research on Arabic sign language is very limited due the dialect nature of the language. Presenting a system and an approach to unify the sign language in the region based on deep learning models and object detection.

Ahmed Abdelhadi Ahmed received his Master of Science in Software Engineering from the Department of Computer Science and Software Engineering, College of Information Technology at UAE University, UAE. He received his Bachelor of Science in Electrical and Electronics Engineering from the College of Engineering, University of Khartoum, Sudan.

www.uaeu.ac.ae

Online publication of thesis:
https://scholarworks.uaeu.ac.ae/etds/