Using Sign-Language as an Input Modality for Microtask Crowdsourcing

Master's Thesis

Aayush Singh

Using Sign-Language as an Input Modality for Microtask Crowdsourcing

by

Aayush Singh

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday August 26, 2022 at 10:00 AM.

Student number:	5208122	
Project duration:	December 1, 2021 – August 26, 2022	
Thesis committee:	Prof. dr. ir. Geert-Jan Houben,	TU Delft, Responsible Professor
	Dr. Ujwal Gadiraju,	TU Delft, Daily Supervisor
	Dr. Frank Broz,	TU Delft, Committee Member
	Sebastian Wehkamp,	ML6, External Supervisor

Preface

This thesis marks the end of my MSc in Computer Science at the Delft University of Technology. Through this thesis, I would like to bring attention to deaf and mute people's involvement in technological landscapes, such as crowdsourcing, through sign languages. Although the microtasks in crowdsourcing are open for everyone's participation, it would invite more participants and lower barriers to participation if the microtasks are made more comfortable and natural for them, in terms of task input.

The thesis journey has been a great experience altogether. Firstly, I want to thank my daily supervisor Dr. Ujwal Gadiraju, ML6 supervisor Sebastian Wehkamp for their guidance throughout, and Alexandru Balan for giving me the opportunity to work on this topic. Their advice, knowledge, and support made our meetings and work a great way to learn and an enjoyable experience overall. It helped me stay motivated and improve the results.

Next, I want to thank Prof. Dr. Geert-Jan Houben and Dr. Frank Broz, for joining the thesis committee and for their interest in my work.

Last but not the least, I am grateful to my parents and my friends for their encouragement and support.

Aayush Singh
Delft, The Netherlands
August 2022

Abstract

Several input types have been developed in different technological landscapes like crowdsourcing and conversational agents. However, sign language remains one of the input types that has not been looked upon. Although numerous amount of people around the world use sign language as their primary language, there have not been many efforts to include them in these technological landscapes. In this thesis, we hope to draw attention to and take a step towards the inclusion of deaf and mute people in microtask crowdsourcing. We identify some of the existing technical and research gaps in the current architectures for Sign Language Recognition/Translation in a real-time setting. Next, we determine various microtasks which can be adapted to use sign language as input, keeping in mind the challenges it introduces. We, then, investigate the effectiveness of a system that uses sign language as input by building a web application - SignUpCrowd - for microtask crowdsourcing, namely Visual Question Answering and Tweet Sentiment Analysis tasks, and comparing it with already prevalent input types such as text and click. This comparison with different popular input types will help understand how much of a difference there is for sign language as input. In addition, it will also show the preference of input types for the particular microtasks. For this, we developed three web applications with different input types and conducted a between-subject experimental study on Prolific wherein a number of workers (N=240) were asked to perform the above-mentioned tasks using sign language, text, and click input. Our results indicate that, in terms of task completion time and task accuracy, sign language as an input modality in microtask crowdsourcing is not significantly different from other, commonly used, input types. We also noticed that people's input type preference for the given microtasks for sign language was more than text input. Although people with no knowledge of sign language found it difficult, this input modality aims at a different target audience. This shows us that there is scope for sign language as an input type for microtask crowdsourcing among people, and paves the way for more efforts for the introduction of sign language in real-world applications.

Contents

List of Figures

List of Tables

1 Introduction

Crowdsourcing has become a universal technique for gathering data from people. It opens up the possibility to reach out to different kinds of people all around the world. This variety of data collection through crowdsourcing helps in better generalization of machine intelligence. It has been used in different domains of computer science, in Demartini et al., 2017 the authors have presented how human computation, by means of crowdsourcing, is used in novel architectures for hybrid human-machine information systems. The work explores different applications where hybrid human-machine information systems have been applied in the domain of information retrieval, natural language processing, semantic web, machine learning, and multimedia to better solve existing problems. In Gadiraju and Yang, 2020, the authors have argued for greater utilization of crowd computing, beyond the practice of leveraging it for training data creation. Among other suggestions, they have highlighted and encouraged the focus on employing crowd computing for problems related to open-ended crowd knowledge creation and conversational human-AI interaction. Some studies have also tried to address different social issues through crowdsourcing. In Abbas et al., 2020, the authors leveraged real-time crowdsourcing to handle and process complex therapeutic conversational tasks for social robotics, Softbank's Pepper robot. Across all of the previous research, it can be seen that there are various types of crowdsourcing such as *macro-task* (similar to outsourcing, e.g. Product Innovation), *microtask* (crowdsourcing to divide typical tasks into more manageable, discrete components, e.g. Data Validation), *crowdfunding* (funding for a venture by raising money from the crowd, e.g. Project Fundraising), *contests* (involving crowd for new innovations and ideas in the form of competitions or volunteer campaigns, e.g. Logo Designs) and possibly many more. For each type of crowdsourcing task, there exist numerous platforms with varying levels of support and input methods as per the task requirement. Our focus is mainly on microtask crowdsourcing. Microtask crowdsourcing can be interpreted as the process of dividing a huge task into several, rapid, little microtasks and assigning them to a vast, unidentified group of workers. It is often used to improve the quality of machine-run algorithms in order to combine both the scalability of machines over large amounts of data as well as the quality of human intelligence in processing and understanding data (Von Ahn and Dabbish, 2008). For example, in model creation for a self-driving car, the car should be able to identify several objects on the road, say traffic lights. Then, the task could be broken down for the crowdworkers into identifying traffic lights in different types of images.

There can be several types of task categories in microtask crowdsourcing. In Gadiraju et al., 2014, the authors broke it down into six broad categories: a) Information Finding, b) Verification and Validation, c) Interpretation and Analysis, d) Content Creation, g) Surveys and f) Content Access. The categories are considered to be high-level and are not just limited to them. They can be a combination of different task categories depending on the task requirement. For all of these different tasks and task categories, there are different input methods available. Some input methods are limited to some tasks due to task requirements whereas some of them are preferred over the others, mainly because of ease of use. Among them, some of the popular input types are click, text, and speech input. Some new input modalities have also been introduced in some research. Like in Alallah, Neshati, Sakamoto, et al., 2018 and Alallah, Neshati, Sheibani, et al., 2018, the authors examined the popularity and social acceptability of head-worn display inputs. Their results indicate that data collected via a crowdsourced experiment and a laboratory-style setting did not differ at a statistically significant level. They conclude that the results provide initial support for crowdsourcing platforms as viable options for conducting social acceptability research. It shows that introduction of new input modalities has been received quite well by the crowdsourcing community.

Introduction to new input types for disabled people or particular types of tasks can invite more participation. This increase in participation will involve different kinds of people and their different perspectives on the tasks. As the greatest strength of crowdsourcing is its greater and wider reach, hence one could assume that opening this avenue for more non-identical people would benefit both parties. In this thesis study, we take a step in this direction and propose that this new audience could be the deaf and mute people in society through the introduction of sign language as a new input modality.

Sign Language (SL) is the primary language for the deaf and mute community. According to the World Federation of the Deaf (UN-ISLD, 2021), there are more than 70 million deaf people around the world that use sign language. It is a natural and complete language that has its own linguistic intricacies. Every spoken language has its own sign language, like American Sign Language (ASL), Chinese Sign Language (CSL), German Sign Language (DGS), and so on. In total, there are around 300 different sign languages. Sign languages are not a one-to-one mapping of spoken languages. They have their own definite grammar. For instance, a well-constructed question must be accompanied by the correct eyebrow position. When a person is asking questions related to who, where, what, why, and when, here the eyebrows are expected in a certain position. If the question is regarding a yes/no situation, the eyebrows are expected in some particular way. SL does not only use hand gestures to communicate but also includes facial expressions, hand movements and positions, and body posture. Any change in them can change the entire meaning of the sign. Some more facts about SL are shown in Figure 1.1. That is why it is generally hard for someone with no knowledge of sign languages to understand them. All of these factors make translation into spoken language difficult. There are mainly two research areas going on in Sign Language interpretation, i.e. Sign Language Recognition (SLR) and Sign Language Translation (SLT). Both of them are explained in the next chapter.

Existing research which is at the intersection of crowdsourcing and sign languages focuses on developing ways to build a corpus for various sign languages utilizing crowdsourcing techniques (Riemer Kankkonen et al., 2018). The paper by Farooq et al., 2021 investigates the idea of engaging the deaf community for the development and validation of a corpus for a sign language and its dialects. The authors propose a framework for building a corpus for sign languages by bringing into use the power of crowdsourcing. Lowering the barriers to participation is an important step toward ensuring the sustainability of the crowdsourcing paradigm (Kittur et al., 2013).

1.1. Problem Statement

In spite of the fact that there are so many people utilizing sign language, there are not many platforms in any technological landscape where sign language literate people can participate and contribute. For example, conversational agents and crowdsourcing platforms have different types of inputs like text, voice, and emoji but, do not include sign languages. One might argue the need for sign language as a new input modality or question why would anyone even want to use sign language for completing crowdsourced tasks. To answer that, simply because it is more natural and

FIVE THINGS YOU DIDN'T KNOW ABOUT THE DEAF

THEY WOULD BE THE THIRD MOST POPULATED COUNTRY IN THE WORLD

There are 466 million people with deafness in the world, so if they were all in the same country, they would be the third most populated country in the world behind China and India. 70 million of these people use Sign Language.

THEY ARE BORN FROM HEARING FAMILIES

Inbreeding is common among Deaf people. This shocks some hearing people who believe they are perpetuating deafness. The truth is that no one can be blamed: between 90 and 95% of deaf children are born from hearing couples.

THEY HAVE GOT BETTER PERIPHERAL VISION

Studies show that Deaf people reach the levels of peripheral vision of an adult in adolescence and that these levels in adulthood remain higher than those of hearing people.

MAYBE THEY DRIVE BETTER

Statistics show that they are as good or better drivers than hearing people. In the United States, they can even drive heavy commercial vehicles since 2013.

THEY ARE NOT SICK AND THEY DO NOT DIE BEFORE

Deafness is not a disease nor is it dangerous to health: statistics show that death rates in deaf people are the same as in hearing people.

Figure 1.1: Facts about Sign Language. Adapted from figure in UnUsualVerse, 2020

comfortable for someone to communicate in their primary language. Like people who can speak and hear properly like to speak/write in their first language, it should also be the case when people are not able to hear and/or speak properly and communicate via sign language.

Our study investigates the possibility of introducing sign language as a new input modality in microtask crowdsourcing. Here are the main research questions we focus on:

- **RQ1:** What are the existing technical and research gaps in the current architectures for Sign Language Recognition/Translation for real-time human interaction?
- **RQ2:** What is the effectiveness of sign language as an input modality in microtask crowdsourcing?

First, we looked at a state-of-the-art architecture available for SLR and SLT and applied the architecture in a real-time setting. Then, we identified the possible technical and research gaps in those architectures when applied in a real-time setting. After identifying the gaps, we looked at different types of tasks in crowdsourcing that can be adapted in some way to include sign languages. Then, to make our argument, for SL input in crowdsourcing platforms, stronger we performed an experiment with crowdworkers on a system, SignUpCrowd, which we built for microtask crowdsourcing using SL as input. We present our analysis of the data gathered and compare it with other input types, text, and click, for the same set of tasks.

1.2. Thesis Contribution

- C1: A comprehensive analysis of technical and research gaps in the current architectures for SLR and SLT (RQ1)
- C2: A microtask crowdsourcing system with SL input type (RQ2)
- C3: A comparative study on how SL input type compares to other input types, like text and click, under the same task setting (RQ2)

All the code and data used for the implementation is publicly available[1].

1.3. Research Outline

- Chapter 2 discusses the history and some of the previous work done in the domain of Sign Language. It also explains the difference between SLR and SLT, and how the requirement for the dataset for the two changes. Finally, it presents how crowdsourcing is currently being utilized with Sign Languages.
- Chapter 3 discusses the methods used to answer the research questions. First, we look into the architecture used to answer RQ1 and then narrow down the categories of the task that can be adapted for SL input, based on the gaps identified. Then, we explain how systems with different input types and the experiments on them were set up.
- Chapter 4 presents the results that we obtained from the experiments, utilizing different parameters.
- Chapter 5 discusses what those results mean i.e. explaining its relevance and limitations with respect to other input types.
- Chapter 6 discusses the conclusions we draw from the results and discussions, and provides an outlook of what future work can be done.

[1] https://osf.io/n8pca/?view_only=fc7bf6ab55d6482f83ff2729c25b937f

2
Background and Related Work

In general, research in sign language has been going on for more than a decade now. As the methods of recognition became sophisticated, deeper segregation of the problem was established. This introduced two main research problems in this domain: 1) Sign Language Recognition and 2) Sign Language Translation, the difference is shown in Figure 2.1. The latter is still considered to be a new problem as it was recently proposed in Camgoz et al., 2018a. This chapter first looks upon the history of SL and then focuses on some of the previous research done in the area of Sign Language. In addition to that, it also explores some of the datasets that are currently available and how crowdsourcing techniques are utilized to create SL datasets.

2.1. History of Sign Language

For most of modern history, spoken languages have been widely used. Its widespread use made it challenging for people to even introduce sign languages, let alone use them. There was a false belief that learning sign language may prevent speech development. For instance, teaching sign languages was outlawed in 1880 by the Second International Congress on Education of the Deaf, a significant international gathering of deaf educators, in favor of speech therapy. It was not until the groundbreaking work on American Sign Language (ASL) by Stokoe Jr, 1960 that signed languages started gaining recognition as natural, independent, and well-defined languages, which then inspired other researchers to further explore signed languages as a research area. However, antiquated notions that pushed away signed languages continue to do harm and subjects many to linguistic neglect (Humphries et al., 2016). Several studies have shown that deaf children raised solely with spoken languages do not gain enough access to a first language during their critical period of language acquisition (Murray et al., 2020). This language deprivation can lead to life-long consequences on the cognitive, linguistic, socio-emotional, and academic development of the Deaf (Hall et al., 2017).

Signed languages are the primary languages of communication for the Deaf and are at the heart of Deaf communities. In an increasingly digitized world, the technological community has a crucial responsibility to include signed languages in its research and allow their participation. Failing to recognize signed languages as full-fledged natural language systems in their own right has had detrimental effects in the past.

2.2. Sign Language Recognition (SLR)

Sign Language Recognition is about recognizing actions from sign language. It is considered to be the naive gesture recognition problem but not just limited to alphabets and numbers. It focuses on

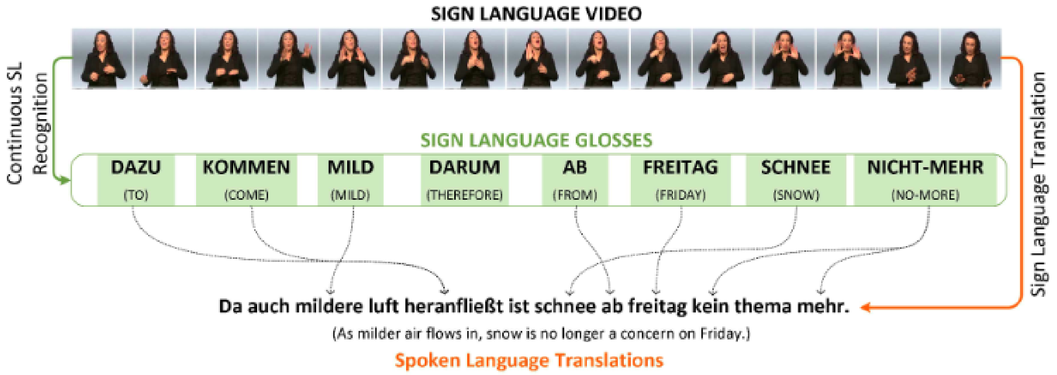

SIGN LANGUAGE VIDEO
Continuous SL Recognition
SIGN LANGUAGE GLOSSES
DAZU (TO)
KOMMEN (COME)
MILD (MILD)
DARUM (THEREFORE)
AB (FROM)
FREITAG (FRIDAY)
SCHNEE (SNOW)
NICHT-MEHR (NO-MORE)
Sign Language Translation
Da auch mildere luft heranfließt ist schnee ab freitag kein thema mehr.
(As milder air flows in, snow is no longer a concern on Friday.)
Spoken Language Translations

Figure 2.1: Difference between Sign Language Recognition and Translation. Adapted from figure 1 in Camgoz et al., 2018a

recognizing a sequence of continuous signs but disregards the underlying rich grammatical and linguistic structures of sign language that differ from spoken language. Much previous work has been around isolated SLR and continuous SLR. Early research focused on recognizing individual basic hand gestures with the help of special gloves or sensors (Starner and Pentland, 1997, Imagawa et al., 1998, Brashear et al., 2003). Starner et al., 1998 and Mehdi and Khan, 2002 looked upon recognizing sign language in a controlled setting where the user was required to have some wearable or sensor gloves on to make tracking easy. There has also been the use of a depth camera, Kinect. In the paper by Lang et al., 2012, they use Kinect and claim that its use makes real-time 3D reconstruction easily applicable, including hidden Markov models with a continuous observation density for recognition. These detections were mainly looking at isolated sign languages.

In continuous SLR (CSLR), Koller et al., 2016 has also utilized the Hidden Markov Model (HMM) framework in the context of SLR. It treats the outputs of the Convolutional Neural Network (CNN) as true Bayesian posteriors and trains the system as a hybrid CNN-HMM in an end-to-end fashion. The architectures that employed hidden Markov models have been noted to have limited capacity to capture temporal information. In Cui et al., 2017 a recurrent CNN based architecture is used. It introduces a three-stage optimization process for training their deep neural network architecture. SubUNets in Cihan Camgoz et al., 2017 inject domain-specific expert knowledge into the system regarding suitable intermediate representations. The authors make use of transfer learning between different interrelated tasks, aiming at exploiting a wider range of more varied data sources. There have been some great results from using Iterative Training. In Cui et al., 2019, deep convolutional neural networks with stacked temporal fusion layers as the feature extraction module, and bidirectional recurrent neural networks as the sequence learning module have been introduced in addition to an iterative optimization process. The training process of first training the end-to-end recognition model for the alignment proposal, and then using the alignment proposal as strong supervisory information to directly tune the feature extraction module, is run iteratively to achieve improvements in the recognition performance. This iterative training scheme is found to partially solve the problem of overfitting while also costing more training time. Min et al., 2021 revisits the iterative training scheme and proposes to enhance the feature extractor with alignment supervision.

Recent innovations have made advantage of a variety of the signer's characteristics, such as numerous visual cues (i,e., hand movement, facial expression, and body posture). Zhou et al., 2020 introduces a spatial-temporal multi-cue (STMC) network to solve the vision-based sequence learning problem. This research creates separate modules (spatial multi-cue and temporal multi-cue) to decompose visual features of different cues and explores the collaboration of multiple cues. Min et al., 2021 comprises two auxiliary losses, one of which focuses on visual features only. In this thesis study, we utilize the Visual Alignment Constraint (VAC) architecture (Min et al., 2021) to do the analysis of SLR/SLT in a real-time setting.

For the evaluation of SLR models, there are several metrics. Most of the metrics are similar to the metrics used for speech-to-text accuracy. One of the most straightforward and popular metrics is Word Error Rate (WER). WER is the measure of performance of the Sign Language Recognition model. It calculates how many “errors“ or “mismatches“ are in the predicted text by the model, compared to the ground-truth annotations. Lower WER indicates a more accurate model. Here is the formula to calculate WER, where *S* is the number of substitutions, *D* is the number of deletions, *I* is the number of insertions, and *N* is the number of words in ground-truth.

$$WER = \frac{S + D + I}{N} \tag{2.1}$$

2.3. Sign Language Translation (SLT)

Sign Language Translation is about interpreting the sign language in terms of natural language, whatever the language may be, with its grammar. The primary objective of SLT is to translate sign language videos into spoken language forms, taking into account the different grammatical aspects of the language. This problem is comparatively new and not much research has been done in this area. However, recently it has gained some focus and there has been ongoing research in order to obtain spoken language from sign language videos.

As per our best knowledge, this problem was first introduced by Camgoz et al., 2018b where the authors not only introduced the problem but along with that a new dataset was introduced, RWTH-PHOENIX-Weather 2014T which contains video segments, gloss annotations, and spoken language translations. Camgoz et al., 2020 builds upon the previous work in SLT and proposes an architecture that jointly learns Continuous Sign Language Recognition and Translation while being trainable in an end-to-end manner. This end-to-end training is achieved by using a Connectionist Temporal Classification (CTC) loss to bind the recognition and translation outputs in a single unified architecture. There have also been attempts to utilize several NLP techniques to achieve better performance in translation. Some research is pointing in the direction of NLP and tokenization methods to appeal to the NLP community (Yin et al., 2021). In Yin and Read, 2020, the authors combines the STMC architecture, from Zhou et al., 2020, with a transformer to achieve improvement on the current state-of-the-art architecture for SLT.

For evaluation of SLT models, a common metric that is utilized for comparing a predicted translation to reference translations is the Bilingual Evaluation Understudy (BLEU) Score. It mainly compares the n-grams of the predicted translation with the n-grams of the ground-truth translation and counts the number of matches. These matches are position-independent (Papineni et al., 2002). A higher BLEU score indicates a better translation model. Table 2.1 shows how different BLEU scores (in percentages) can be interpreted (Google-BLEU, 2022).

BLEU Score	Interpretation
<10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
>60	Quality often better than human

Table 2.1: BLEU Score Interpretation.

Dataset Name	Language	Vocabulary	Samples
Isolated Signs Datasets			
MS-ASL [Joze and Koller, 2018]	American	1,000	25,000
WLASL [LI et al., 2020]	American	2,000	21,083
DEVISIGN [Chai et al., 2014]	Chinese	4,414	331,050
BosphorusSign [Camgöz et al., 2016]	Turkish	855	24,161
BosphorusSign22k [Özdemir et al., 2020]	Turkish	744	22,542
INCLUDE [Sridhar et al., 2020]	Indian	263	4,287
Continuous SL Datasets			
RWTH-BOSTON-104 [Dreuw et al., 2007]	American	104	201 sentences
RWTH-BOSTON-400 [Dreuw et al., 2008]	American	483	843 sentences
How2Sign [Duarte et al., 2021]	American	15,686	35,000 sentences
CSL-Daily [Zhou et al., 2021]	Chinese	2,000	21,000 sentences
RWTH-PHOENIX-Weather 2014 [Koller et al., 2015]	German	2,048	6,841 sentences
RWTH-PHOENIX-Weather 2014 T [Camgoz et al., 2018b]	German	2,887	8,257 sentences

Table 2.2: Sign Language Datasets

2.4. Benchmark Datasets

The available datasets are also similar to the research categories in Sign Language. The contents of the dataset vary as per the research category. For research in Isolated Sign Language, the datasets mostly contain different words for different domains and their sign language sequence. Moreover, some of these datasets, like Chai et al., 2014 and Özdemir et al., 2020, also contain data regarding depth and pose, providing for insightful research. On the other hand for Continuous Sign Language Recognition, there are sentences present with sign language videos and gloss annotations, like in Koller et al., 2015 and Zhou et al., 2021. One of the most widely used datasets for CSLR is RWTH-PHOENIX-Weather 2014 introduced by Koller et al., 2015. It contains videos from German weather forecasts recorded at 25fps, with the size of frames being 210 by 260 pixels. An extension to this dataset is the RWTH-PHOENIX-Weather 2014 T introduced by Camgoz et al., 2018b. It also contains the transcription of the original German speech which helps in SLT. There is also a newly published data, How2Sign dataset by Duarte et al., 2021 which is for the English language, covering a vast variety of vocabulary. However, until March 2022 there were some annotations for the dataset missing and were being processed. Some of the other datasets that are being utilized for research are shown in Table 2.2.

2.5. Crowdsourcing and Sign Languages (SL)

To support the research in SL, several different repositories of video gestures are available for many sign languages of the world. In this area, there have been several research attempts to utilize sign languages in crowdsourcing. They focus mainly on how to develop datasets for different sign languages using crowdsourcing techniques. In Farooq et al., 2021, the authors present a framework for building a parallel corpus for sign languages by exploiting the powers of crowdsourcing. They developed a word-level parallel corpus comprising the gestures of almost 700 words of Pakistan Sign Language (PSL) and a sentence-level translation corpus comprising more than

8000 sentences for different tenses for PSL. The study by Tanaka et al., 2020 examines the use of crowdsourcing in the conversion of sign language to text. Their system divides live sign language videos into shorter segments, distributing them to workers. Then the worker interprets and types the segments into text, the system generates captions through the integration of these texts. They claim that the system allows the interpretation of sign language-to-caption text, and also provides an opportunity for deaf and mute individuals to assist those that are unable to read sign language.

In terms of introducing a new input modality for microtask crowdsourcing, it is important to realize the factors that might influence the crowdworker's response and affinity for the input modality. One of the important factors is task clarity. In Gadiraju, Yang, and Bozzon, 2017, the authors highlight the importance of clarity of the properties of a task in crowdsourcing. This defines the performance of workers because their understanding of tasks will transfer into their response to tasks. The authors show results that clearly picture the importance of clarity as an explicit property of microtask crowdsourcing. In Nouri et al., 2021, the authors employ natural language processing techniques to aid in identifying clarity flaws in microtask descriptions. Hence, this opens up an interesting research area in terms of microtask clarity upon the introduction of a new input modality. Another important factor that is quite evident, but is considered invisible is the work environment. It can also be considered as a barrier to participation just because of a lack of resources some workers might fail to do a microtask. In Gadiraju, Checco, et al., 2017, the authors reveal the significant role of work environments in the shaping of crowd work. Through their experiments on how workers deal with UI design choices and how it influences the quality of the work produced, they found substantial evidence that confirms the “invisible“ role of work environments in shaping crowd work.

3
Method and Experimental Setup

Our aim was to determine the gaps, technical and research-based, in current architectures for SLR/SLT. And to find out how effective is a sign language input-based system for microtask crowdsourcing, discussing whether it would be more appealing for people with knowledge of sign language. Thus for the first part, we utilized a state-of-the-art architecture for SLR and analyzed it in a real-time setting. For the second part, we developed three web applications having different methods of input to complete the task, 1) Sign Language[1]; 2) Text[2]; and 3) Button Click[3]. To safeguard the validity of the comparison, the main workflow of the task was kept similar for all the input modalities. The code used for the implementation is made publicly available[4].

3.1. Understanding the Technical and Research Gaps (RQ1)

The main goal of RQ1 is to understand the current state-of-the-art architectures for SLR/SLT in a real-time setting. This would involve focusing on utilizing one of the architectures and experimenting with some actual SL videos. It would not only help us understand how good or bad the architecture is, but would also help us point out the gaps, in technical and research terms, for taking SL input in real-world applications. In the current research landscape of SL, there are mainly two research papers that have achieved the lowest Word Error Rate (WER, the lower the better): Spatial-Temporal Multi-Cue (STMC) for CSLR by Zhou et al., 2020 and Visual Alignment Constraint for CSLR (VAC-CSLR) by Min et al., 2021. Due to the unavailability of public code for the former, we utilized VAC-CSLR. This architecture focuses on the problem of SLR by achieving a 21.2 WER on the RWTH Phoenix Weather 14 Dataset. To make use of this architecture for the SLT problem, we added a transformer for translation over the VAC-CSLR architecture, as shown in Figure 3.1. Moreover, RWTH Phoenix Weather 14T dataset was used to train both networks separately. Table 3.1, shows key statistics of the dataset. In the table, OOV stands for Out-Of-Vocabulary words which appear in the development and test sets, but not in the training set of the dataset. The architecture is based on a two-step, Sign-to-Gloss Gloss-to-Text, translation where the first step is to obtain glosses from the video sequence, and in the next step, the glosses are converted into spoken language sentences. After the training and testing phase, the model was utilized in a real-time setting. It was tested on the different videos with translation happening on the go, in sets of frames using OpenCV. And, MediaPipe (Lugaresi et al., 2019) was used to identify when to start and end a sign sequence.

[1] https://seventh-port-334508.nw.r.appspot.com
[2] https://app2-service-dot-seventh-port-334508.nw.r.appspot.com
[3] https://app3-service-dot-seventh-port-334508.nw.r.appspot.com
[4] https://osf.io/n8pca/?view_only=fc7bf6ab55d6482f83ff2729c25b937f

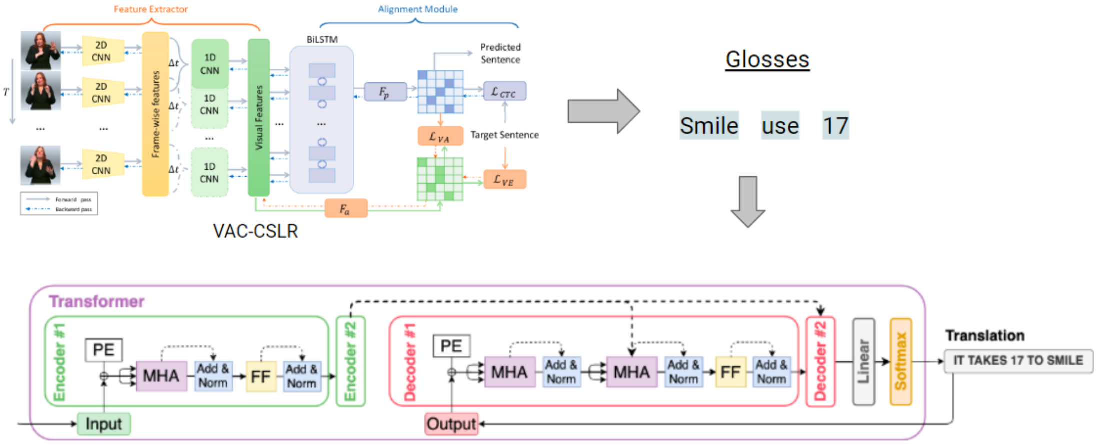


Figure 3.1: VAC-CSLR + Transformer network for SLT

	German Sign Gloss			German		
	Train	Dev	Test	Train	Dev	Test
segments	7,096	519	642	7,096	519	642
vocabulary	1,066	393	411	2,887	951	1,001
total words	67,781	3,745	4,257	99,081	6,820	7,816
total OOVs	-	19	22	-	57	60

Table 3.1: Statistics of the RWTH-Phoenix-Weather 2014T Dataset

3.1.1. Visual Alignment Constraint (VAC) Network + Transformer

The first stage required utilizing the VAC network (Min et al., 2021) to obtain glosses from the video sequences. The Visual Alignment Constraint network focuses on enhancing the feature extractor with alignment supervision by proposing two auxiliary losses: the Visual Enhancement (VE) loss and the Visual Alignment (VA) loss. The VE loss provides direct supervision for the feature extractor, which itself is enhanced with the addition of an auxiliary classifier F_a on visual features V to get the auxiliary logits $\widetilde{Z} = (\widetilde{z}_1, ..., \widetilde{z}_t) = F_a(V)$. This auxiliary loss makes the feature extractor make predictions based on local visual information only. Then, to compensate for the contextual information that VE loss lacks, the VA loss is proposed. The VA loss is implemented as a knowledge distillation loss (Hinton et al., 2015), which regards the entire network and the visual feature extractor as the teacher and student models, respectively. The final objective function is composed of the primary Connectionist Temporal Classification (CTC) loss, the visual enhancement loss, and the visual alignment loss:

$$\mathcal{L} = \mathcal{L}_{CTC} + \mathcal{L}_{VE} + \mathcal{L}_{VA} \tag{3.1}$$

In the second stage to obtain translation from glosses, a two-layered Transformer was used to maximize the log-likelihood

$$\sum_{(x_i, y_i) \in D} logP(y_i | x_i, \theta) \tag{3.2}$$

where D contains gloss-text pairs (x_i, y_i).

We referred to the original Transformer (Vaswani et al., 2017) implementation for more details.

3.1.2. Experiment Setup on the VAC + Transformer Network

After hyper-parameter tuning and model validation, the model was applied to different videos from the published datasets and clips from various SL-friendly news channels. Due to the fact that the models were trained on a German SL dataset, therefore the videos were mainly selected from German SL sources. We utilized random videos from RWTH-Phoenix-Weather 2014 and RWTH-Phoenix-Weather 2014-T dataset, and also took SL snippets from Tagesschau[5], a news show in Germany, for evaluation. These videos were not very long, just a sentence long (so, up to 8-10 seconds). In the translation pipeline, a video is broken down into frames of images and on every image a MediaPipe holistic model is run, which identifies key points from the image. If the identified key points contain left or right-hand key points then the SLR model starts taking frames for prediction. This set of frames is decided on the basis of key-point detection of the left or right hand from the MediaPipe holistic model. Then, after we get the glosses from the VAC model, these glosses are passed to the Transformer model which provides the spoken translations. The final translations were compared to the actual text for the SL video sequence. In addition to this, we also applied different transformations to the frames captured from videos. Here are the transformations that were applied:

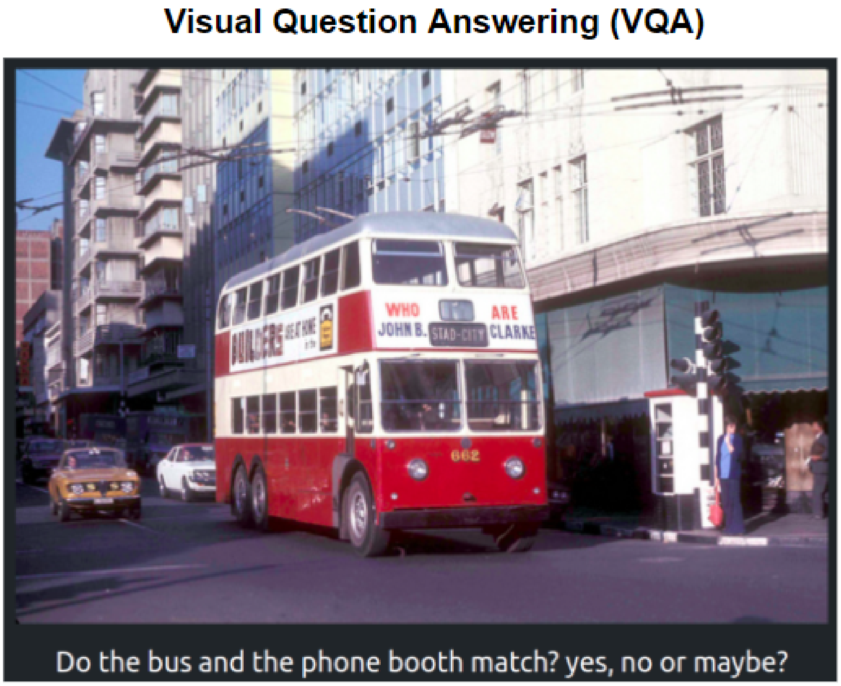
- Segmentation masks: A mask is used to segment an image. It is used to identify the parts of an image containing a particular object, in this case, a human. It was mainly used to avoid noise in the images, with the background being insignificant for prediction.
- Image rotations: It is a common image processing operation. The image is rotated at various angles to capture the different aspects of the image features in different orientations.
- Image resizing: In this, the size of the image was changed by the central cropping method at different dimensions.
- Image scaling: This is different from image resizing as it happens on the entire image by resampling. The images were scaled randomly between 0.5 to 1.5 intervals.

3.2. System and Task Description (RQ2)

Nowadays, all kinds of tasks are being crowdsourced, from quality assurance to product development. In the study by Gadiraju et al., 2014, the authors have categorized the tasks into 6 high-level classes: 1) Information Finding (IF), 2) Verification and Validation (VV), 3) Interpretation and Analysis (IA), 4) Content Creation (CC), 5) Surveys (S), and 6) Content Access (CA). The tasks are not just limited to these classes but can also be a mix of them. We utilize this as a reference for determining the types of tasks where Sign Language can be introduced as an input modality. Apart from IF and CA, all of the other tasks seem to have the potential to include the new input modality, i.e. SL. This is because tasks under IF and CA, involve navigation through the internet, whereas all the other tasks are more in a question-answer setting where instead of saying or writing the answer, using sign language might suit some workers. However the categories list is not fully exhaustive, therefore a combination of IF or CA with other categories might also open up the possibility for the introduction of a new input modality. In this regard, we decided to choose basic tasks from two of these classes (VV and IA).

There are two types of tasks present for each of the applications, namely Visual Question Answering (VQA: *Class VV*) and Tweet Sentiment Analysis (TSA: *Class IA*), the example shown in Figure 3.2 and Figure 3.3 respectively. In total, there were 16 sub-tasks to be completed for each worker. The sub-tasks were equally distributed in each batch of tasks, but they were randomly

[5] https://www.tagesschau.de/


Visual Question Answering (VQA)

Do the bus and the phone booth match? yes, no or maybe?

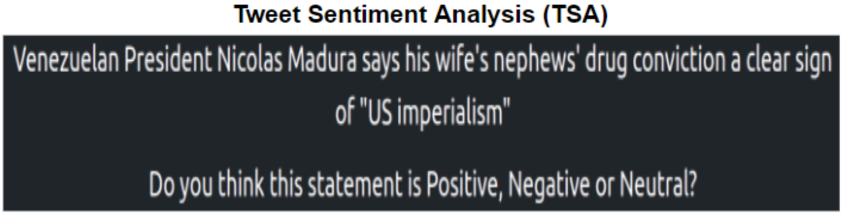
Figure 3.2: Example of VQA Sub-Task

arranged for completion. So, each crowdworker was expected to complete 16 sub-tasks, a combination of VQA and TSA in random order. Along with the basic task description, there are also instructions to help a crowdworker understand the task better, including some sign language examples.

For the Visual Question Answering task, a picture would be shown to the worker. Along with the picture, there would be a question regarding the picture (e.g., “Do you see a body of water in the picture? “). The answer to the question will be a “YES“, “NO“ or “MAYBE“. The answer from the worker would then be captured via the input type of the web application.

For the Tweet Sentiment Analysis task, a text/tweet would be shown to the user. The user will be required to assess the sentiment behind the text/tweet (for e.g., “This time tomorrow...we’ll have the Iron on. Iron Maiden pieces Drops tomorrow nights.“) by choosing one of “POSITIVE“, “NEGATIVE“ or “NEUTRAL“ options. The answer would again be captured via the input type of the web application.

In the SL input web application, the worker had 15 seconds to answer the question asked in each sub-task. After those 15 seconds, it would move to the next sub-task. Apart from this, there was also a trial mode available for people with no knowledge of American Sign Language (ASL) which can be used as many times as they like. This trial mode had less number of tasks, 5, but the rest of the conditions remained similar. On the other hand in the text and button click input web application, the worker had to answer the question and then move on to the next sub-task. For all the conditions, after the completion of the task, the worker was provided with a survey link that had questions about the user experience of the task.



Tweet Sentiment Analysis (TSA)

Venezuelan President Nicolas Madura says his wife's nephews' drug conviction a clear sign of "US imperialism"

Do you think this statement is Positive, Negative or Neutral?

Figure 3.3: Example of TSA Sub-Task

3.2.1. Task Participants and Quality Control

We recruited 240 participants (80 for each input type) from the Prolific crowdsourcing platform[6]. The number of participants or sample size was decided on the basis of A priori Power Analysis (Effect size, f = 0.25) done on G*Power (Faul et al., 2009). For participant selection, there was a criterion of approval rate of more than 50%. The task with SL input required workers to have a camera in their system, otherwise, there were not any other technical restrictions kept from our side. From the total 240 participants selected, some of them were removed due to incomplete responses and we were left with 210 participants (70 for each input). Out of 210 workers, 53% were female, and 47% were male. Their average age was 27.6 years old (SD=8.91). The participants were rewarded by the minimum reward rate of £6 per hour and the tasks lasted for around 10 minutes on average per participant. Among 70 workers for the SL input, there were 12 workers who had knowledge of some kind of sign language. Workers who participated in one condition were not allowed to participate in the other condition using Prolific's built-in screening feature. To prevent malicious activity on the microtasks, we had attention check questions in the user experience form (Gadiraju et al., 2015). The questions within the questionnaire were rephrased and randomly placed in the questionnaire.

3.2.2. Measures

The effectiveness of an application can be measured in several ways, depending on the type of application. For introducing a new input modality, we wanted to achieve effectiveness in terms of work quality and user experience. So, we determined the effectiveness of the SignUpCrowd application by measuring the following factors:

- Quality of work: Determining how accurate the responses from the crowdworkers are. The dataset, COCO dataset[7] for VQA (Goyal et al., 2017) and tweet_eval sentiment dataset[8] (Rosenthal et al., 2017) from Hugging Face for TSA, used for the microtasks had ground truth labels present with them. The responses captured from the tasks through sign language were compared with the actual labels.

- User satisfaction of the system: Determining the user experience of the application. After the task, the workers were given a survey form that had questions related to task experience, time allotted, preference towards different input types, and level of sign language interpretability. The questions consisted of 12 items in which the workers were asked to pick the most suitable level of agreement for the statement (e.g. "The system was able to correctly interpret the signs I made for the sub-tasks."; 1 = *strongly disagree* and 5 = *strongly agree*). The user experience

[6] https://www.prolific.co/

[7] https://visualqa.org/

[8] https://huggingface.co/datasets/tweet_eval

survey had a high level of internal consistency (Cronbach's α = 0.89). Some of the questions which were specified for the Sign Language input type were replaced.

- Comparison with other input types: Any new input modality will only be effective if it solves a problem and has at least the same usability as the existing input types. The comparison was made on the basis of completion time, response accuracy, and also on user experience.

3.2.3. System Implementation

For the development of microtask crowdsourcing applications (input types: SL, text, and click) for the experiment, several technologies were utilized. Here are the technologies used and their description of their use in the applications:

- Flask + Javascript

  For the web application, we utilized Flask which is a web framework written in Python. It provides flexibility in implementation and technical experimentation, and fast integration with new solutions. For the front-end interaction, we used Javascript. Javascript is a scripting or programming language that allows the development of different features on a web page. Along with them, HTML and CSS were also used for giving structure and style to the web pages.

- MediaPipe

  MediaPipe (Lugaresi et al., 2019) provides ML solutions like face tracking, hand tracking, pose detection, segmentation, object detection, etc. for live and streaming media. We used the MediaPipe Holistic pipeline to capture the pose, hand, and face components. For each of the components, there are different models being optimized. For pose and face landmarks, it uses the BlazeFace model (Bazarevsky et al., 2019), 33 and 468 landmarks respectively. And for hands, it uses a single-shot detector palm detection model (Liu et al., 2016), 21 landmarks per hand, as shown in Figure 3.4. We utilized it to collect key points or landmarks as training data for different words necessary for the task. Apart from collecting landmarks data, it also helped in identifying when to start predicting during a live video stream, based on hands landmarks.

- OpenCV

  OpenCV is an open-source library that is usually used for image processing and computer vision tasks. We also used it for capturing video from the system's camera and saving it in frames of image (30fps).

- Google Cloud Platform (GCP)

  The applications were deployed on Google Cloud Platform using App Engine. Google App Engine (GAE) is a platform-as-a-service product that provides the option of scalable and fully-managed deployment of web applications. It allows the developers to focus on the code of the application and GAE itself handles the platform and infrastructural concerns.

- SLR Model Description

  For the application using SL input, SignUpCrowd, we developed a model for recognizing the signs necessary for task completion. The model architecture was inspired by different skeleton-based architectures for SLR. We had to reduce the number of layers and parameters in the original architectures as the data in our case was less. The final model had two LSTM layers and three Dense layers trained with Adam optimizer (learning rate = 0.001), as shown in Figure 3.5.

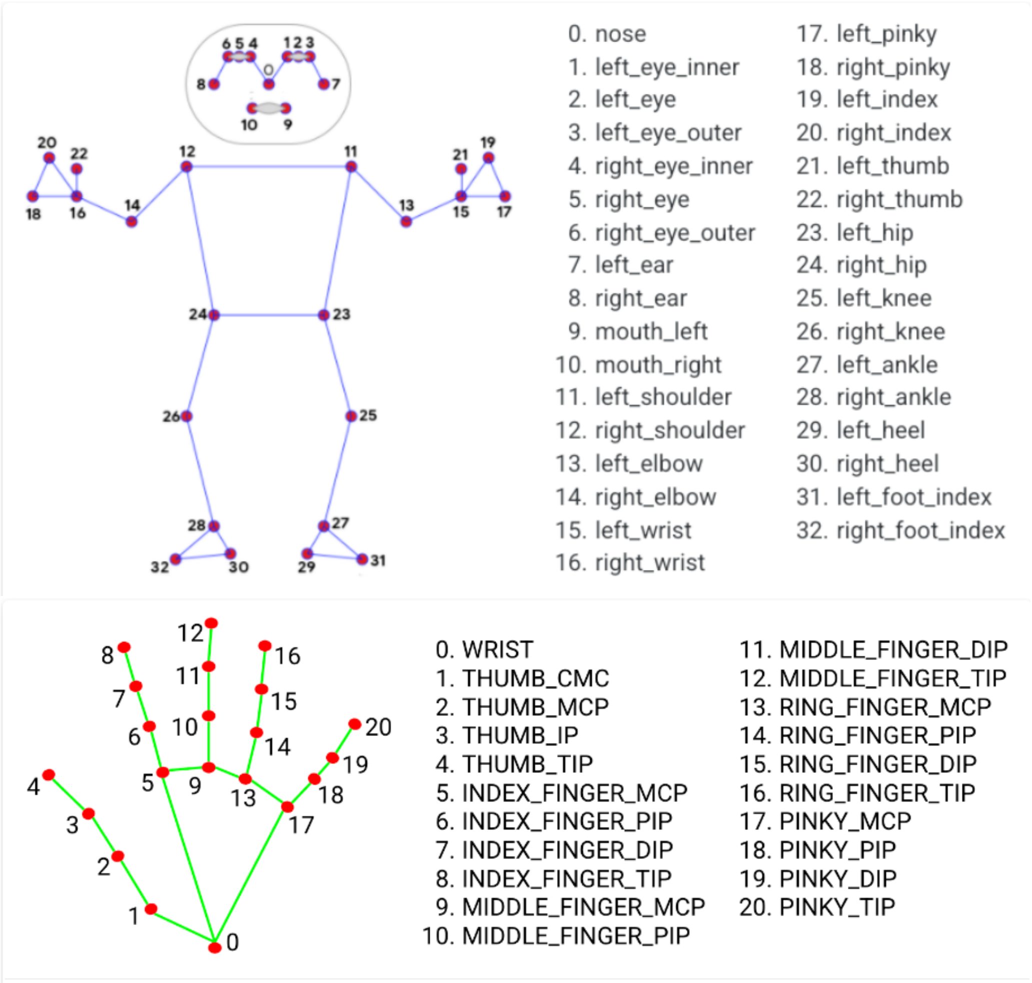
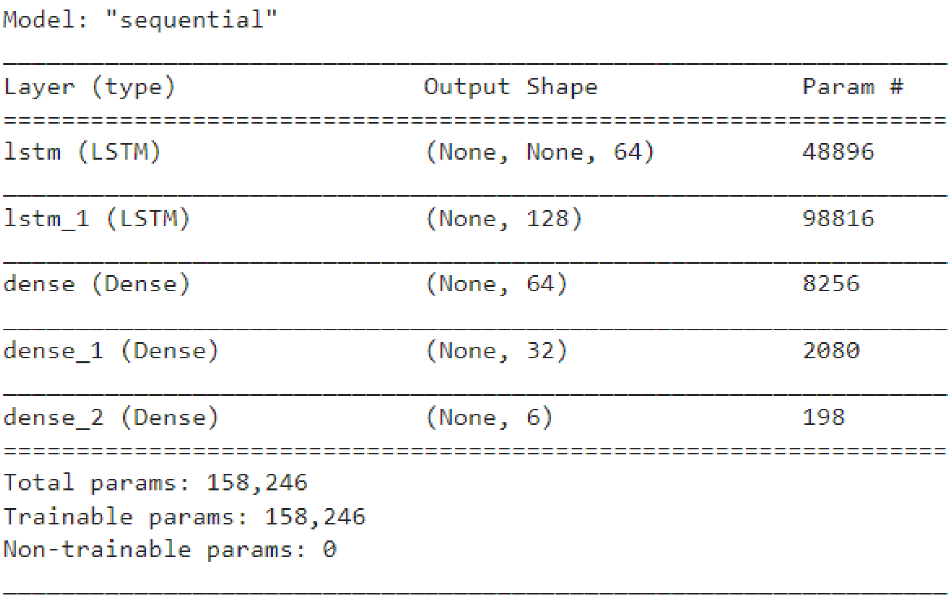


Figure 3.4: Pose and Hand landmarks in MediaPipe Holistic model

.

3.2.4. Procedure

The entire task for a participant is divided into 2 stages: actual task and post-task questionnaire. For SL input, there is an optional stage available where the participant can try out the microtasks and camera settings before actually starting the task. When the worker lands on the home page of the web application, there are instructions present to inform about the respective microtasks and in case of SL input, about the different signs necessary to complete the microtasks. After that, the worker can start the task, in case of SL input the worker can also choose to try out the signs with microtasks as many times as they want. Then during the task, different microtasks are shown as explained previously. When the task ends, the worker is given a link to a post-task questionnaire. This questionnaire has some basic user details and task experience questions. Upon submission of the questionnaire, the worker is given the Prolific task completion link to get paid. Appendix A and B has details about the post-task questionnaire and web application workflow respectively.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
lstm (LSTM)                  (None, None, 64)          48896     
_________________________________________________________________
lstm_1 (LSTM)                (None, 128)               98816     
_________________________________________________________________
dense (Dense)                (None, 64)                8256      
_________________________________________________________________
dense_1 (Dense)              (None, 32)                2080      
_________________________________________________________________
dense_2 (Dense)              (None, 6)                 198       
=================================================================
Total params: 158,246
Trainable params: 158,246
Non-trainable params: 0
_________________________________________________________________
```

Figure 3.5: Model summary used for SL input application

.

3.2.5. Potential Biases

There have been studies where it has been seen that crowdsourced data that comprises a subjective component is potentially affected by the inherent bias of crowd workers who contribute to the tasks. In Hube et al., 2019, the authors aim to understand the influence of workers' own opinions on their performance in the subjective task of bias detection. Their findings reveal that workers with strong opinions tend to produce biased annotations and such bias should be mitigated to improve the quality of the data collected. In Draws et al., 2021, the authors propose a 12-item checklist adapted from business psychology to combat cognitive biases in crowdsourcing. We utilize this checklist to point out the potential biases in the data collected in our study out of the 12 items.

- Sunk Cost Fallacy: This cognitive bias is about "*Is the time required to complete my task and what it requires from crowd workers clear at the onset?*". Especially in the case of the SL input web application, SignUpCrowd, the workers might not be completely aware of the total time required to complete the task. As there is also a section to try out the application workflow, hence they could be spending more time on the task than expected, just to understand the SL and application better. For the application for other input types, this would not be the case as the task is quite straightforward with no try-it-out section.

- Loss Aversion: This cognitive bias focuses on "*Does my task design give crowd workers a reason to suspect that they may not get paid (fairly) after executing my task?*". There is a possibility that this bias might occur due to the extra understanding required to complete the task. As most of the participants do not have knowledge about SL, hence it is possible that the time spent to gain a basic understanding might make them suspicious and susceptible to this bias. Therefore, we tried to provide as concise and direct information as needed for the task as possible.

4 Results and Analysis

In this chapter, we will go through the results of different experiments performed to answer the research questions. First, we look at the technical and research gaps identified from the experiment on several videos in a real-time setting. Next, we present the results obtained from the analysis of the experiments done by crowdworkers on applications with different input types.

4.1. Technical and Research Gaps

To understand the current state of research in SLR and SLT for real-time application, we utilized the combination of one of the state-of-the-art architectures for SLR with transformers. After conducting several experiments on different videos, we found some gaps in the technical and research aspects of current architectures for SLR/SLT. Here are the technical and research gaps we found after our analysis:

1. **Limited Number of Datasets Available**:

   To increase the performance of SLR models, it is important to have enough datasets available. As per the current research in SL, almost all research papers mention the need for more data to progress the research quality. The datasets available are mostly of alphabets, numbers, and individual words. There are also datasets for Continuous SLR that contain gloss representations for the SL sequences, but, for SLT, spoken translations are also required. There are very few datasets that contain spoken translations as well in the dataset. The main reason is that the SLT problem is comparatively new and also for spoken translation annotations, human SL interpreters are required to translate the entire video dataset. It is important because the problem of SLT is crucial for real-world applications which connect people with SL knowledge to the ones that do not.

   Another aspect of limited datasets available is that most of the SL corpora discussed in this paper and various other papers are either unavailable for use due to the presence of corrupted or unreachable data, or available under heavy restrictions and licensing terms. SL data is particularly challenging to anonymize due to the need for valuable facial and other physical features in signing videos, therefore restricting its open distribution.

2. **Domain Restricted Data**:

   Most of the benchmark datasets present currently are collected from a certain SL media source which is domain-specific. Like the current benchmark dataset for SLT, the RWTH-Phoenix-Weather 2014T dataset (Camgoz et al., 2018b) of German Sign Language, contains videos from the daily weather forecast airings of the German public TV station PHOENIX featuring sign language interpretation. If a model is trained on a domain-specific dataset, then it is possible that it has not generalized well and has a limited vocabulary i.e. vocabulary specific to the domain. Most of the open-source SL sources, like news channels, are domain-specific so, it becomes challenging to develop a dataset that is open-domain.

3. **Lack of Variety in Datasets**:

   In the current datasets, there has been a lack of variety in terms of the number of signers, physical orientation of signers, and camera viewpoints of signers. There has been an average of 10-20 signers across various datasets, with the RWTH-Phoenix-Weather 2014T Dataset having just 9 signers. An increased number of native signers gives a better understanding of sign representation. In SL there are different dialects, this makes variations in signs for the same word. So, it is possible that the same word or phrase can be signed in different ways by different people, or the sign sequence of the same word may differ from one region to another. Therefore, it is better to capture this variation as much as possible by selecting a variety of signers.

   Another aspect that comes under variety is the camera viewpoint from which the signer is captured for the dataset collection. Generally, for a real-time application, it is not necessary that the signer will always be captured from the front by the camera. Currently, more than 85% of the datasets do not have multiple views.

4. **Architecture Usability for Different SL**:

   Recently, the research in SLR/SLT has been increasing. The architectures are capturing various aspects of an SL video sequence. However, after scrutinizing different results from these types of research it is quite apparent that the accuracy results (WER score and BLEU score) are not similar when the same architecture is tried on a different language dataset. Like, for an SLR architecture proposed in Min et al., 2021, on RWTH-Phoenix-Weather 2014 dataset it got 21.2 WER; on Modern Chinese Sign Language (CSL) dataset (Huang et al., 2018) 1.6 WER. Then, for an SLT architecture proposed in Yin and Read, 2020, on RWTH-Phoenix-Weather 2014T dataset it got 22.17 BLEU; on Public DGS Corpus (Hanke et al., 2020) just 3.2 BLEU. Therefore, these results indicate that current architectures are not appropriate for real-world applications, either more data is needed for these models or approaches that are more linguistically sophisticated are required.

5. **Hardware Restriction for Deep Architectures**:

   Another technical gap that is worth mentioning is the limits of hardware for conventional deep learning architectures. It is an important aspect in the light of real-world applications as real-world applications are expected to be robust and swift in delivering outputs. In Davies et al., 2021, the authors present the importance of software maturity for the development of new architectures in deep learning.

4.2. Effectiveness of SignUpCrowd Application

The analysis of the quality of work done using different input types was based on the task accuracy rate and time of task completion. This analysis was done on an overall level as well as sub-task level.

Overall, as shown in Figure 4.2, the mean task accuracy for Sign Language input was 39.12% (SD=19.91); for Text input, it was 43.67% (SD=14.12); and for Click input, it was 49.52% (SD=12.94). We also look at how different input types performed task-wise, shown in Figure 4.1. For the Visual Question Answering (VQA) task, the mean task accuracy for SL input was 44.94%; for Text input, it was 37.85%; and for Click input, it was 39.77%. While for the Tweet Sentiment Analysis (TSA) task,

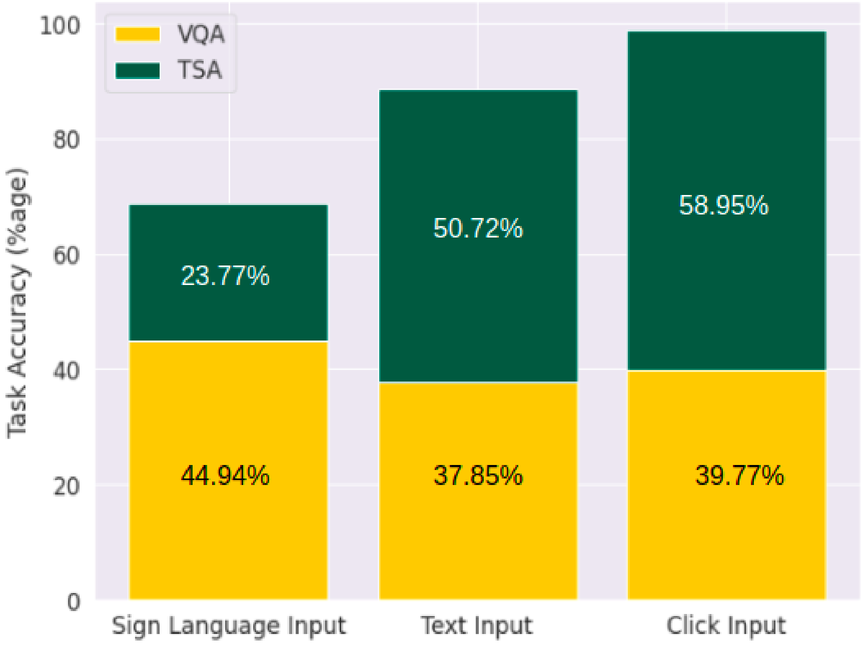


Figure 4.1: Task-wise accuracy comparison between different input types

.

the mean task accuracy for SL input was 23.77%; for Text input, it was 50.72%; and for Click input, it was 58.95%. For a better understanding of the completion time and task accuracy score, we also performed a statistical significance test, as shown in Table 4.3. We listed down 4 main hypotheses which compared SL input with text and click input in terms of task accuracy and completion time. Among the different types of statistical tests available, we chose Wilcoxon Signed-Rank Test as it is a non-parametric test i.e., the population does not follow the normal distribution. It tests the null hypothesis that two related paired samples come from the same distribution. All of the hypotheses mentioned in Table 4.3 are valid except *H04* (No time completion difference between SL and click input types). It means that for the given types of tasks it is efficient to use click input. Although click input is easy and quick for such tasks, it restricts the worker to select from the fixed set of options and that is where SL can be helpful.

Next, we measured the system usability for all three applications using the System Usability Scale (SUS) score Brooke et al., 1996. Research conducted by Bangor showed the range of SUS scores can be seen in Table 1. Using this table, it can be measured whether the application is acceptable or not in terms of usability Bangor et al., 2009. Figure 4.3, shows the result of all input type web applications using SUS. The x-axis line shows each input type from web application with input types such as Sign Language, Text, and Click. The y-axis indicates the SUS mean score. The result of evaluation using SUS of system with SL input got an average of 73.28, text input with a mean score of 70.96, and click input with the highest mean score of 75.92. According to the table designed by Bangor et al., 2009, the value of all designs belongs to the acceptable category which is above 70. In addition to this, there was also a section for feedback and suggestion in the user experience questionnaire. Table 4.2 shows some of the selected user suggestions for all the three input type applications.

After the completion of the task, the workers were asked to fill out a post-task experience form. We divided the questions in the form into three broad categories: Time Enough for Completion,

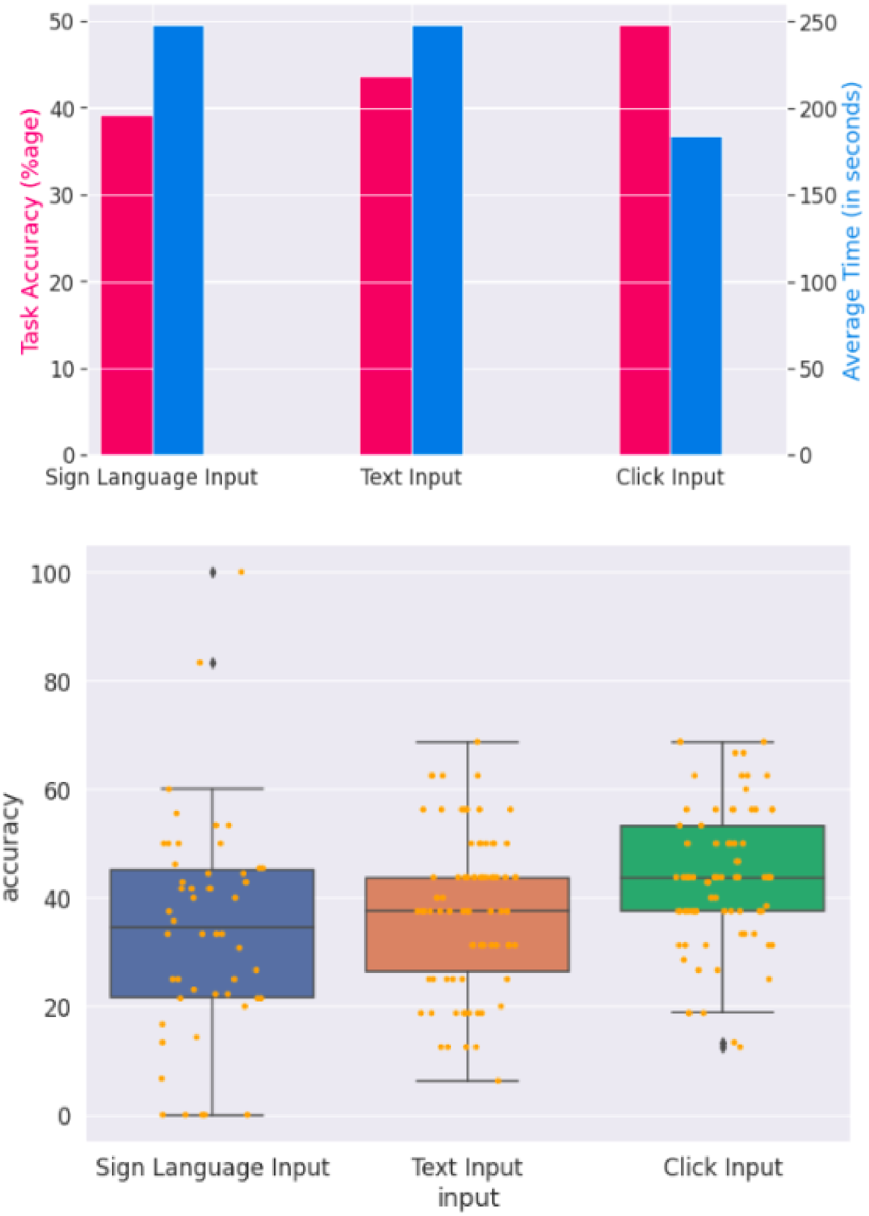


Figure 4.2: Time and Accuracy comparison between different input types

.

SUS score	Interpretation
≤ 50	Not Acceptable
50 - 70	Marginal
≥ 70	Acceptable

Table 4.1: The range of SUS Values by Bangor

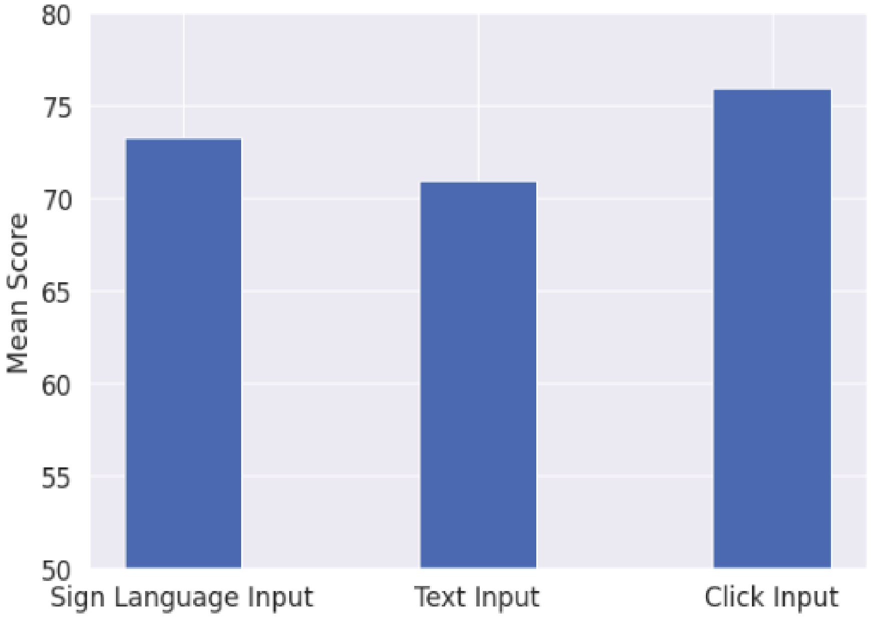


Figure 4.3: SUS Mean Score for different input type application

.

Interface Satisfaction, and Task Preference. Figure 4.4, shows the average ratings for each of the categories for the different input types. The ratings from the survey suggest that the interface available for the tasks was suitable for completion. In terms of task preference, the majority of workers preferred to choose click input for the given tasks (VQA and TSA). Overall, the average rating for choosing click input over text input was 4.3 and for choosing text over click was 2.3. On the other hand, the preference for sign language for the given tasks was 3.0. Among the workers who performed the tasks with SL input, the average rating for 85% workers who did not know sign language was 2.4 whereas the average rating for workers who knew sign language was 3.8.

For people who did not have any knowledge of sign language, there was also a “TRY IT OUT“ section. The survey showed that more than 80% of people utilized this section in the SL input application to make themselves aware of the SL and the application flow. The average rating for the “TRY IT OUT“ section being perceived as helpful was 3.5.

Suggestions and Feedbacks for SignUpCrowd
- “I slowly started getting used to it, but I think a longer and more detailed practice session would be needed.“
- “Next time you could help by maybe giving diagrams of what a yes, no or maybe looks like in sign language.“
- “The webcam was lagging, but overall was a nice studie“
- “The interface was fun and interactive. I enjoyed it.“
- “I think the camera box should be bit bigger“
- “very interesting“
- “The system interpreting was very slow.“
- “the task was not clear enough for me.“
- “There were glitches and several responses were detected.“
- “I struggled with the try it out feature. Using it was complicated, but it is a nice initiative for sign language inclusion.“
- “Overall, the system was fine. I had to do the signs multiple times for it to recognize.“
Suggestions and Feedbacks for Text Input Application
- “No problems; the instructions were clear.“
- “Very hard to understand the language used in the tweets, as made no grammatical sense.“
- “No problem faced“
- “I’d suggest that the text box be big enough for the expected responses and that the box is focused so I don’t have to click on it first to type my response“
Suggestions and Feedbacks for Click Input Application
- “Well usable. The buttoned solution is better than the text.“
- “The graphics for these kind of task could be improved“
- “it was an interesting survey“
- “no not much, everything was nice“

Table 4.2: Selected Post-task User Suggestions for the three applications

	Hypothesis	**Statistical Test**	**Statistic**	**p value**	**Reject?**
H01	No task accuracy difference between SL and text input types	Wilcoxon Signed-Rank Test	469.5	0.3172	No
H02	No task accuracy difference between SL and click input types	Wilcoxon Signed-Rank Test	382.5	0.0350	No
H03	No time completion difference between SL and text input types	Wilcoxon Signed-Rank Test	1447.0	0.3147	No
H04	No time completion difference between SL and click input types	Wilcoxon Signed-Rank Test	685.0	7.3028e-06	Yes

Table 4.3: Statistical Significance Test for comparison in completion time and task accuracy for text and click input types with SL input type

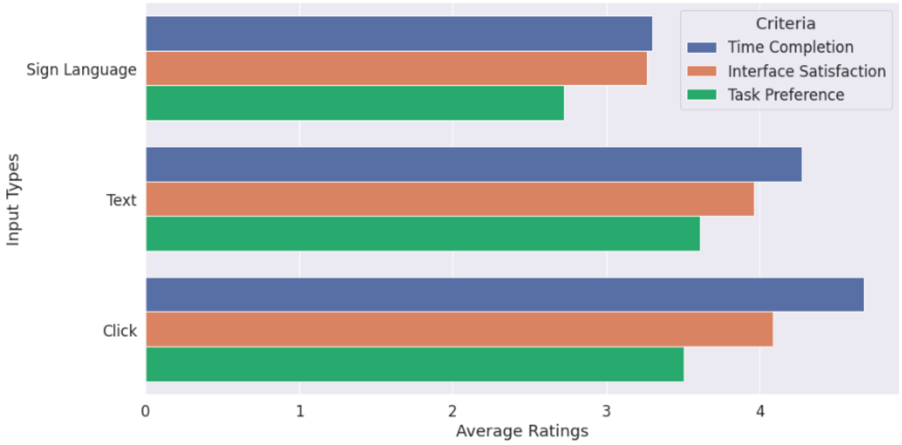


Figure 4.4: Average User Ratings from Post-Task Survey

.

5 Discussion and Implications

In this thesis, we first presented the technical and research gaps in current state-of-the-art architecture for SLR/SLT. We utilized a combination of VAC-CSLR and Transformer for SLT. After experimenting with various SL videos, we listed down the gaps we observed, both technical and research. The gaps mentioned are primarily aimed at identifying the areas where the current architectures and datasets are insufficient for any real-world application. Next, we studied the effectiveness of SignUpCrowd, a microtask crowdsourcing system with sign language as an input modality. Here, our main aim was to check whether there is any difference in task completion time and task accuracy in comparison to other input types like text and click. In addition to this, we also looked at how the crowdworkers experienced the application. Our results indicate that the new input modality for microtask crowdsourcing has comparable results. Moreover, all of the workers who had knowledge of sign language showed interest in the application and showed their preference for completing these microtasks using sign language.

5.1. Challenges for SLR/SLT in Real-time

Our analysis of current architectures for SLR/SLT in a real-time setting involved other architectures as well. However, we chose to utilize the state-of-the-art VAC-CSLR for making conclusions about the gaps for SL input in real-time use. As the VAC-CSLR architecture was openly available and it was trained on a benchmark dataset of RWTH-Phoenix-Weather 2014-T, hence we decided to take advantage of it. We listed down quite a few important gaps in the current research for the use of SL input in real-world applications. Although we believe that all the mentioned gaps are equally important, it is evident that most of them revolve around having a proper distribution of SL data. In our opinion, it is important to have a proper and open dataset that not only solves small specific problems but also considers the bigger goal of achieving decent usability of SL interpretation. This way all the other gaps can also be addressed.

It is important to note that simpler architectures might seem to perform well in a real-time setting due to the small scope of vocabulary. Various research claims to achieve high performance on SL datasets containing alphabets, numbers, or very restricted and limited vocabulary. It can still be seen as a step forward in the direction of increasing participation and inviting the disabled. However, its use will be limited and would not properly achieve the goal of reducing barriers to participation for deaf and mute people.

5.2. Task Quality

An obvious observation from the task accuracy results is that the overall accuracy of the responses across different input types is low. This might be due to the difficult nature of the task at hand. The datasets for the task were filtered, and only the ones with less than or equal to 60% confidence level of response were selected. And as the responses were only assumed correct if they exactly matched the correct label, hence it is possible that the workers could not be completely sure about the answers for the sub-task.

The difference in task accuracy between SL and text input is around 4% while the difference between SL and click input is around 10%. This difference is not very significant. We should also take into account the fact that most of the crowdworkers participating in the SignUpCrowd application had no knowledge of Sign Language. As mentioned in the Participants subsection, there were only 15% participants who had previous knowledge of sign language. So considering this, it seems that the difference in task accuracy is not large. It is expected that people who did not have any knowledge of Sign Language would find it difficult to learn and use SL right away. This fact adds to our argument that there are not many deaf and mute people participating in such microtasks.

Another aspect that was measured during the experiments was the time taken for completion. There was no difference in mean time completion between SL and text input, while the difference between SL and click input was about 64 seconds. From the user experience form, we found out that, for the SL input application, the time to complete the task was enough (average rating of 3.5 out of 5). Hence in this case as well, it can be seen that the overall time taken to complete the microtasks is equivalent to other input types.

5.3. User Satisfaction

The results from the post-task user experience survey show that the ratings for SL input are comparable to other input types. It is clear that the ratings for SL input are on the lower side in comparison, but overall the ratings are still close to ratings for text and/or click input. The time for completion of the task and the interface for the task was, mostly, suitable for all the workers, in general. Also, it is evident from the post-task survey that workers with no knowledge of sign language found it difficult to complete the task. Their preference was more towards the other input types for the given task. However, the workers with sign language knowledge showed a preference for sign language.

5.4. Caveats and Limitations

The overall simplistic nature of the VQA and TSA sub-tasks made it easy to compare different parameters of the task and invite different workers to participate with different input types. The crowdworkers who participated in the experiment were mainly people who did not have any knowledge of SL. This can be perceived as an inaccurate judgment of the effectiveness of SignUpCrowd. As it can be difficult to assess something you are not aware of. To manage this lack of sign language knowledge among workers, we had specific instructions for them to learn about SL and a “TRY IT OUT“ section, where they could try out their sign attempts. We understand that a justified effectiveness evaluation of the SignUpCrowd application would be when the assessment comes from the people who it targets. However, our experiment was performed with the help of the Prolific crowdsourcing platform which did not have an option to select people with sign language knowledge. So, we tried to cover all kinds of participants and their experiences with the task.

Any application which attempts to utilize Sign Language as an input method will have to consider that there will be less flexibility in terms of usability. This is because the worker/user needs to have a device that has a camera and is/can be fixed, otherwise capturing the nuances of Sign Language will become difficult.

Another limitation that is worth pointing out is that microtasks like Content Creation (CC) (e.g.

‘Translate the following content into German’) will need sophisticated architectures that can be applied in a real-time setting, keeping in mind the hardware restrictions for the device.

6
Conclusion and Future Work

In this paper, we studied the gaps in the architectures for SLR/SLT by considering and exploiting existing state-of-the-art architectures. Along with this, we also looked at the introduction of sign language as a new input modality for microtask crowdsourcing by making comparisons with other popular input types like text and click, under similar task settings. We developed three applications for different input types (SL, text, and click) and analyzed crowdworkers responses. The introduction of a new sign language input method would attract a lot of new people to the crowdsourcing landscape. From our analysis, we conclude that the SL input type is not different from other input types like text and click. Although it can be deduced that people with no knowledge of sign language will not prefer to use sign language for performing microtasks, this new input type will provide an opportunity for the deaf and mute to participate in microtasks crowdsourcing. Basically, its introduction will introduce new participants and increase the coverage of the crowdsourcing community. However, our mentioned gaps also suggest that there needs to be more advancement in architectures and datasets to achieve high-level real-world applications.

6.1. Conclusions

We proposed two research questions in Chapter 1. We studied the current state-of-the-art SLR architecture, VAC_CSLR (Min et al., 2021), implemented a system for microtask crowdsourcing with SL input type and conducted a comparison between various input types for microtask crowdsourcing. By analyzing all the information obtained during the study, we propose answers to the two research questions.

Research Question 1:
What are the existing technical and research gaps in the current architectures for Sign Language Recognition/Translation for real-time human interaction?

We utilized the state-of-the-art architecture proposed in Min et al., 2021, VAC_CSLR, to answer RQ1. To employ this architecture for SLT problem, we added a two-layered Transformer to the existing architecture. We answer this research question based on the problem of SLT. For experiments, we utilized different SL videos from different sources (datasets/news channels).

Upon our analysis, we listed the gaps present in the current architectures for a real-world application of SL interpretation. We identified 5 main gaps in the current architectures for SLR/SLT, namely 1) *Limited Number of Datasets Available*, 2) *Domain Restricted Data*, 3) *Lack of Variety in*

Datasets, 4) *Architecture usability for different SL*, and 5) *Hardware Restriction for deep architectures*. We conclude that although the current architectures for SLR/SLT might not be fully equipped for a real-world application for SL interpretation, the progress in terms of dataset and architecture looks promising. As the problem of SLT at hand is difficult, therefore various aspects of SL must be considered to fulfill the difficulty.

Research Question 2:
What is the effectiveness of sign language as an input modality in microtask crowdsourcing?

To answer this research question, we implemented three applications having different input types (SL, text, click) for microtask crowdsourcing. The comparison between different input types helped us understand how different will it be for the workers for task completion. It was important because it helped us understand how good or bad will the introduction of SL is as an input type. There was also a post-task questionnaire for the crowdworkers to understand their experience and preference. The results showed that the introduction of SL as an input type in microtask crowdsourcing would be welcomed by people with SL knowledge. They showed more preference for the SL input type. Comparison with factors like time required for completion, task completion rate, task performance, and interface satisfaction showed that the SL input type was comparable to existing, commonly used, text and click input types.

6.2. Contributions

The contributions of this thesis are:

1. A comprehensive analysis of technical and research gaps in the current architectures for SLR and SLT.
2. SignUpCrowd: A microtask crowdsourcing system with SL input type.
3. A comparative study on how SL input type compares to other input types, like text and click, under the same task setting.

6.3. Future Work

This domain of Sign Language adoption in the technological landscape brings a lot of angles for future work. In our opinion, an interesting future work could be to investigate the provision of creation of a new anonymous dataset. As mentioned in previous chapters, most of the SL datasets are restricted for use so, it can be a promising problem to work towards an anonymizing technique with minimal information loss as facial features are an important aspect of SLR. We also suggest that a future study could investigate increasing the microtasks difficulty and variety, focusing on experimenting with Sign Language speakers. This can also be modified into a setting where the video sequence from the workers is recorded for dataset creation. Finally, several other Sign Language Translation architectures can also be looked upon for utilization in a real-time setting. This can also focus on a different technological landscape such as conversational agents. To motivate the introduction of sign language as a new input modality it is important to start working on such applications, making the best use of currently available architectures.

Bibliography

Abbas, T., Khan, V.-J., Gadiraju, U., Barakova, E., & Markopoulos, P. (2020). Crowd of oz: A crowd-powered social robotics system for stress management. *Sensors*, *20*(2), 569.

Alallah, F., Neshati, A., Sakamoto, Y., Hasan, K., Lank, E., Bunt, A., & Irani, P. (2018). Performer vs. observer: Whose comfort level should we consider when examining the social acceptability of input modalities for head-worn display? *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*. https://doi.org/10.1145/3281505.3281541

Alallah, F., Neshati, A., Sheibani, N., Sakamoto, Y., Bunt, A., Irani, P., & Hasan, K. (2018). Crowd-sourcing vs laboratory-style social acceptability studies? examining the social acceptability of spatial user interactions for head-worn displays. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–7.

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, *4*(3), 114–123.

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., & Grundmann, M. (2019). Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*.

Brashear, H., Starner, T., Lukowicz, P., & Junker, H. (2003). Using multiple sensors for mobile sign language recognition. *SMARTech*.

Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4–7.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018a). Neural sign language translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018b). Neural sign language translation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7784–7793.

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Camgöz, N. C., Kındıroğlu, A. A., Karabüklü, S., Kelepir, M., Özsoy, A. S., & Akarun, L. (2016). Bosphorussign: A turkish sign language recognition corpus in health and finance domains. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1383–1388.

Chai, X., Wang, H., & Chen, X. (2014). The devisign large vocabulary of chinese sign language database and baseline evaluations. *Technical report vipl-tr-14-slr-001. key lab of intelligent information processing of chinese academy of sciences (cas)*. Institute of Computing Technology.

Cihan Camgoz, N., Hadfield, S., Koller, O., & Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Cui, R., Liu, H., & Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7361–7369.

Cui, R., Liu, H., & Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, *21*(7), 1880–1891. https://doi.org/10.1109/TMM.2018.2889563

Davies, M., Labiosa, A., & Sankaralingam, K. (2021). Understanding the limits of conventional hardware architectures for deep-learning. *arXiv preprint arXiv:2112.02204*.

Demartini, G., Difallah, D. E., Gadiraju, U., Catasta, M., et al. (2017). An introduction to hybrid human-machine information systems. *Foundations and Trends® in Web Science*, *7*(1), 1–87.

Draws, T., Rieger, A., Inel, O., Gadiraju, U., & Tintarev, N. (2021). A checklist to combat cognitive biases in crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *9*, 48–59.

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., & Ney, H. (2008). Benchmark databases for video-based automatic sign language recognition. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/pdf/287_paper.pdf

Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., & Ney, H. (2007). Speech recognition techniques for a sign language recognition system. *hand*, *60*, 80.

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., & Giro-i-Nieto, X. (2021). How2sign: A large-scale multimodal dataset for continuous american sign language. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2735–2744.

Farooq, U., Mohd Rahim, M. S., Khan, N. S., Rasheed, S., & Abid, A. (2021). A crowdsourcing-based framework for the development and validation of machine readable parallel corpus for sign languages. *IEEE Access*, *9*, 91788–91806. https://doi.org/10.1109/ACCESS.2021.3091433

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149–1160.

Gadiraju, U., Checco, A., Gupta, N., & Demartini, G. (2017). Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–29.

Gadiraju, U., Kawase, R., & Dietze, S. (2014). A taxonomy of microtasks on the web. *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, 218–223. https://doi.org/10.1145/2631775.2631819

Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 1631–1640.

Gadiraju, U., & Yang, J. (2020). What can crowd computing do for the next generation of ai systems? *CSW@ NeurIPS*, 7–13.

Gadiraju, U., Yang, J., & Bozzon, A. (2017). Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. *Proceedings of the 28th ACM conference on hypertext and social media*, 5–14.

Google-BLEU. (2022). Bleu score interpretation - automl documentation. [Accessed: 2022-07-19].

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hall, W. C., Levin, L. L., & Anderson, M. L. (2017). Language deprivation syndrome: A possible neurodevelopmental disorder with sociocultural origins. *Social psychiatry and psychiatric epidemiology*, *52*(6), 761–776.

Hanke, T., Schulder, M., Konrad, R., & Jahn, E. (2020). Extending the Public DGS Corpus in size and depth. *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Lan-*

guage Community, Technological Challenges and Application Perspectives, 75–82. https://aclanthology.org/2020.signlang-1.12

Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, *2*(7).

Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Hube, C., Fetahu, B., & Gadiraju, U. (2019). Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Humphries, T., Kushalnagar, P., Mathur, G., Napoli, D. J., Padden, C., Rathmann, C., & Smith, S. (2016). Avoiding linguistic neglect of deaf children. *Social Service Review*, *90*(4), 589–619.

Imagawa, K., Lu, S., & Igi, S. (1998). Color-based hands tracking system for sign language recognition. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 462–467.

Joze, H. R. V., & Koller, O. (2018). Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The future of crowd work. *Proceedings of the 2013 conference on Computer supported cooperative work*, 1301–1318.

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, *141*, 108–125.

Koller, O., Zargaran, O., Ney, H., & Bowden, R. (2016). Deep sign: Hybrid cnn-hmm for continuous sign language recognition. *Proceedings of the British Machine Vision Conference 2016*.

Lang, S., Block, M., & Rojas, R. (2012). Sign language recognition using kinect. *International Conference on Artificial Intelligence and Soft Computing*, 394–402.

LI, D., Rodriguez, C., Yu, X., & LI, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. *European conference on computer vision*, 21–37.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Mehdi, S. A., & Khan, Y. N. (2002). Sign language recognition using sensor gloves. *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.*, *5*, 2204–2206.

Min, Y., Hao, A., Chai, X., & Chen, X. (2021). Visual alignment constraint for continuous sign language recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11542–11551.

Murray, J. J., Hall, W. C., & Snoddon, K. (2020). The importance of signed languages for deaf children and their families. *The Hearing Journal*, *73*(3), 30–32.

Nouri, Z., Gadiraju, U., Engels, G., & Wachsmuth, H. (2021). What is unclear? computational assessment of task clarity in crowdsourcing. *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 165–175.

Özdemir, O., Kındıroğlu, A. A., Camgöz, N. C., & Akarun, L. (2020). Bosphorussign22k sign language recognition dataset. *arXiv preprint arXiv:2004.01283*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Riemer Kankkonen, N., Björkstrand, T., Mesch, J., & Börstell, C. (2018). Crowdsourcing for the swedish sign language dictionary. *8th Workshop on the Representation and Processing of Sign Languages, Miyazaki, Japan, 12 May, 2018*, 171–174.

Rosenthal, S., Farra, N., & Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 502–518.

Sridhar, A., Ganesan, R. G., Kumar, P., & Khapra, M. (2020). Include: A large scale dataset for indian sign language recognition. *Proceedings of the 28th acm international conference on multimedia* (pp. 1366–1375). Association for Computing Machinery. https://doi.org/10.1145/3394171.3413528

Starner, T., Weaver, J., & Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(12), 1371–1375. https://doi.org/10.1109/34.735811

Starner, T., & Pentland, A. (1997). Real-time american sign language recognition from video using hidden markov models. *Motion-based recognition* (pp. 227–243). Springer.

Stokoe Jr, W. C. (1960). Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, *10*(1), 3–37.

Tanaka, K., Wakatsuki, D., & Minagawa, H. (2020). A study examining a real-time sign language-to-text interpretation system using crowdsourcing. *International Conference on Computers Helping People with Special Needs*, 186–194.

UN-ISLD. (2021). International sign language day [Accessed: 2022-06-07].

UnUsualVerse. (2020). Infographic: Five things you didn't know about the deaf [Accessed: 2022-04-14].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 58–67.

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., & Alikhani, M. (2021). Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*.

Yin, K., & Read, J. (2020). Better sign language translation with stmc-transformer. *Proceedings of the 28th International Conference on Computational Linguistics*, 5975–5989.

Zhou, H., Zhou, W., Qi, W., Pu, J., & Li, H. (2021). Improving sign language translation with monolingual data by sign back-translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1316–1325.

Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2020). Spatial-temporal multi-cue network for continuous sign language recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 13009–13016.

A

User Experience Questionnaires

A.1. SL Input Experience Questionnaire

The questions were mainly related to three categories: time completion, interface satisfaction, and task preference. Most of the questions followed the Likert scale [1-5], with 1 being Strongly Disagree and 5 being Strongly Agree. Here are the questions for SL input type application:

1. Do I have any knowledge of Sign Language (E.g.: American Sign Language)?

   [multiple choice]: Yes/No

2. If you selected ”Yes” in the previous question, which sign language do you know?

   [open]

3. I used the ”TRY IT OUT” section to get aware of the application.

   [Likert 1-5]

4. I felt comfortable using this system.

   [Likert 1-5]

5. I was able to complete the tasks and scenarios quickly using this system.

   [Likert 1-5]

6. There was enough time to complete the task.

   [Likert 1-5]

7. Overall I am satisfied with how easy it is to use this system.

   [Likert 1-5]

8. The information (such as on-screen messages, task descriptions, and other links) provided with this system was clear.

   [Likert 1-5]

9. The ”TRY IT OUT” section was quite helpful to know the application before starting.

   [Likert 1-5]

10. It was easy to find the information I needed.

    [Likert 1-5]

11. The information was effective in helping me complete the task.

    [Likert 1-5]

12. The time allotted (15 seconds) to finish each sub-task was enough.

    [Likert 1-5]

13. The organization of information on the system screens was clear.

    [Likert 1-5]

14. I would prefer to use Sign Language over typing for such tasks.

    [Likert 1-5]

15. The interface of this system was pleasant.

    [Likert [1-5]

16. The system was able to correctly interpret the signs I made for the sub-tasks.

    [Likert 1-5]

17. Overall, I am satisfied with this system and could use Sign Language to complete any future tasks.

    [Likert 1-5]

18. Do you have any other suggestions / recommendations / problems faced related to the entire task?

    [open]

A.2. Text and Click Input Experience Questionnaire

The questionnaire for text and click input was same. The questions were kept in a similar fashion as for the SL input questionnaire. Some of the questions which were specific to SL input type were replaced/removed. Here are the questions for text and click input type application:

1. I felt comfortable using this system.

   [Likert 1-5]

2. I was able to complete the tasks and scenarios quickly using this system.

   [Likert 1-5]

3. There was enough time to complete the task.

   [Likert 1-5]

4. Overall I am satisfied with how easy it is to use this system.

   [Likert 1-5]

5. The information (such as on-screen messages, task descriptions, and other links) provided with this system was clear.

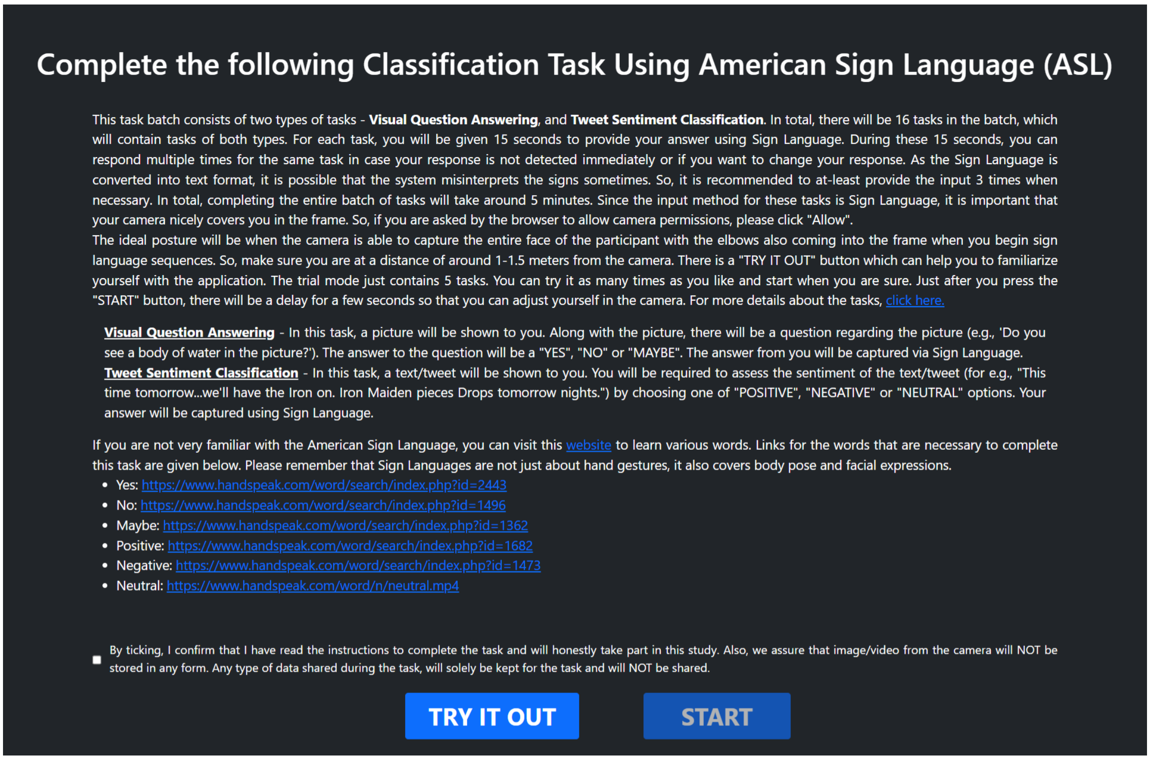
   [Likert 1-5]

6. It was easy to find the information I needed.
   [Likert 1-5]
7. The information was effective in helping me complete the task.
   [Likert 1-5]
8. The organization of information on the system screens was clear.
   [Likert 1-5]
9. I would prefer to type for such tasks.
   [Likert 1-5]
10. The interface of this system was pleasant.
    [Likert [1-5]
11. I would prefer to click buttons for such tasks.
    [Likert 1-5]
12. Overall, I am satisfied with this system.
    [Likert 1-5]
13. Do you have any other suggestions / recommendations / problems faced related to the entire task?
    [open]

B

Web Applications Workflow

B.1. SignUpCrowd: SL input type



Complete the following Classification Task Using American Sign Language (ASL)

This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you will be given 15 seconds to provide your answer using Sign Language. During these 15 seconds, you can respond multiple times for the same task in case your response is not detected immediately or if you want to change your response. As the Sign Language is converted into text format, it is possible that the system misinterprets the signs sometimes. So, it is recommended to at-least provide the input 3 times when necessary. In total, completing the entire batch of tasks will take around 5 minutes. Since the input method for these tasks is Sign Language, it is important that your camera nicely covers you in the frame. So, if you are asked by the browser to allow camera permissions, please click "Allow".
The ideal posture will be when the camera is able to capture the entire face of the participant with the elbows also coming into the frame when you begin sign language sequences. So, make sure you are at a distance of around 1-1.5 meters from the camera. There is a "TRY IT OUT" button which can help you to familiarize yourself with the application. The trial mode just contains 5 tasks. You can try it as many times as you like and start when you are sure. Just after you press the "START" button, there will be a delay for a few seconds so that you can adjust yourself in the camera. For more details about the tasks, click here.

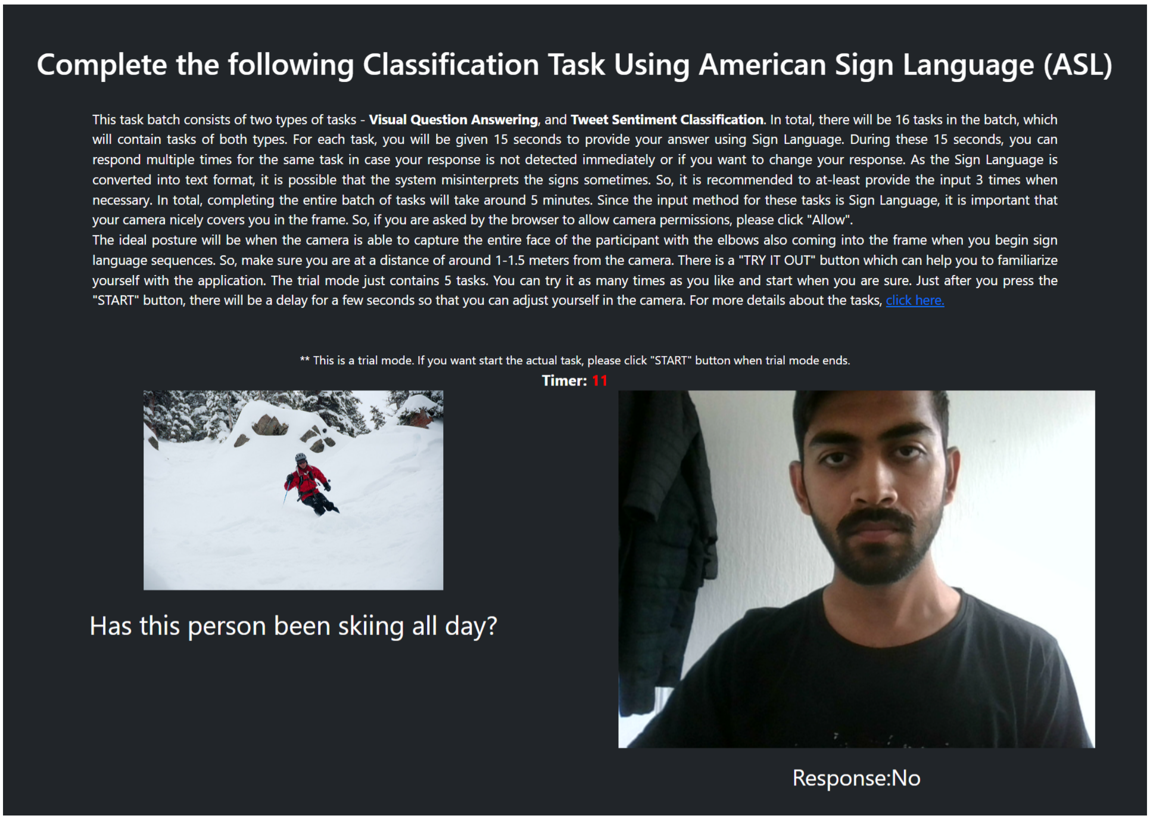
Visual Question Answering - In this task, a picture will be shown to you. Along with the picture, there will be a question regarding the picture (e.g., 'Do you see a body of water in the picture?'). The answer to the question will be a "YES", "NO" or "MAYBE". The answer from you will be captured via Sign Language.
Tweet Sentiment Classification - In this task, a text/tweet will be shown to you. You will be required to assess the sentiment of the text/tweet (for e.g., "This time tomorrow...we'll have the Iron on. Iron Maiden pieces Drops tomorrow nights.") by choosing one of "POSITIVE", "NEGATIVE" or "NEUTRAL" options. Your answer will be captured using Sign Language.

If you are not very familiar with the American Sign Language, you can visit this website to learn various words. Links for the words that are necessary to complete this task are given below. Please remember that Sign Languages are not just about hand gestures, it also covers body pose and facial expressions.

- Yes: https://www.handspeak.com/word/search/index.php?id=2443
- No: https://www.handspeak.com/word/search/index.php?id=1496
- Maybe: https://www.handspeak.com/word/search/index.php?id=1362
- Positive: https://www.handspeak.com/word/search/index.php?id=1682
- Negative: https://www.handspeak.com/word/search/index.php?id=1473
- Neutral: https://www.handspeak.com/word/n/neutral.mp4

□ By ticking, I confirm that I have read the instructions to complete the task and will honestly take part in this study. Also, we assure that image/video from the camera will NOT be stored in any form. Any type of data shared during the task, will solely be kept for the task and will NOT be shared.

Figure B.1: SignUpCrowd home page



Complete the following Classification Task Using American Sign Language (ASL)

This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you will be given 15 seconds to provide your answer using Sign Language. During these 15 seconds, you can respond multiple times for the same task in case your response is not detected immediately or if you want to change your response. As the Sign Language is converted into text format, it is possible that the system misinterprets the signs sometimes. So, it is recommended to at-least provide the input 3 times when necessary. In total, completing the entire batch of tasks will take around 5 minutes. Since the input method for these tasks is Sign Language, it is important that your camera nicely covers you in the frame. So, if you are asked by the browser to allow camera permissions, please click "Allow".
The ideal posture will be when the camera is able to capture the entire face of the participant with the elbows also coming into the frame when you begin sign language sequences. So, make sure you are at a distance of around 1-1.5 meters from the camera. There is a "TRY IT OUT" button which can help you to familiarize yourself with the application. The trial mode just contains 5 tasks. You can try it as many times as you like and start when you are sure. Just after you press the "START" button, there will be a delay for a few seconds so that you can adjust yourself in the camera. For more details about the tasks, click here.

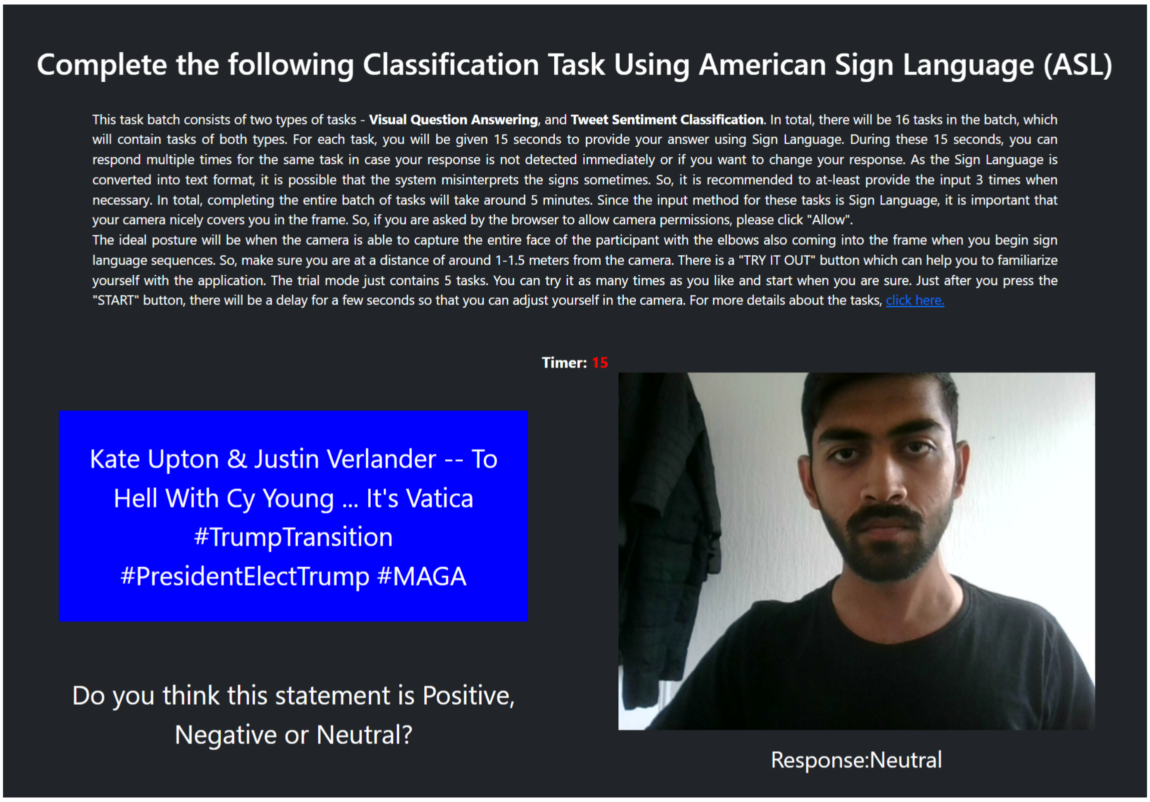
** This is a trial mode. If you want start the actual task, please click "START" button when trial mode ends.

Timer: 11

Has this person been skiing all day?

Response:No

Figure B.2: SignUpCrowd TRY IT OUT section



Complete the following Classification Task Using American Sign Language (ASL)

This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you will be given 15 seconds to provide your answer using Sign Language. During these 15 seconds, you can respond multiple times for the same task in case your response is not detected immediately or if you want to change your response. As the Sign Language is converted into text format, it is possible that the system misinterprets the signs sometimes. So, it is recommended to at-least provide the input 3 times when necessary. In total, completing the entire batch of tasks will take around 5 minutes. Since the input method for these tasks is Sign Language, it is important that your camera nicely covers you in the frame. So, if you are asked by the browser to allow camera permissions, please click "Allow".
The ideal posture will be when the camera is able to capture the entire face of the participant with the elbows also coming into the frame when you begin sign language sequences. So, make sure you are at a distance of around 1-1.5 meters from the camera. There is a "TRY IT OUT" button which can help you to familiarize yourself with the application. The trial mode just contains 5 tasks. You can try it as many times as you like and start when you are sure. Just after you press the "START" button, there will be a delay for a few seconds so that you can adjust yourself in the camera. For more details about the tasks, click here.

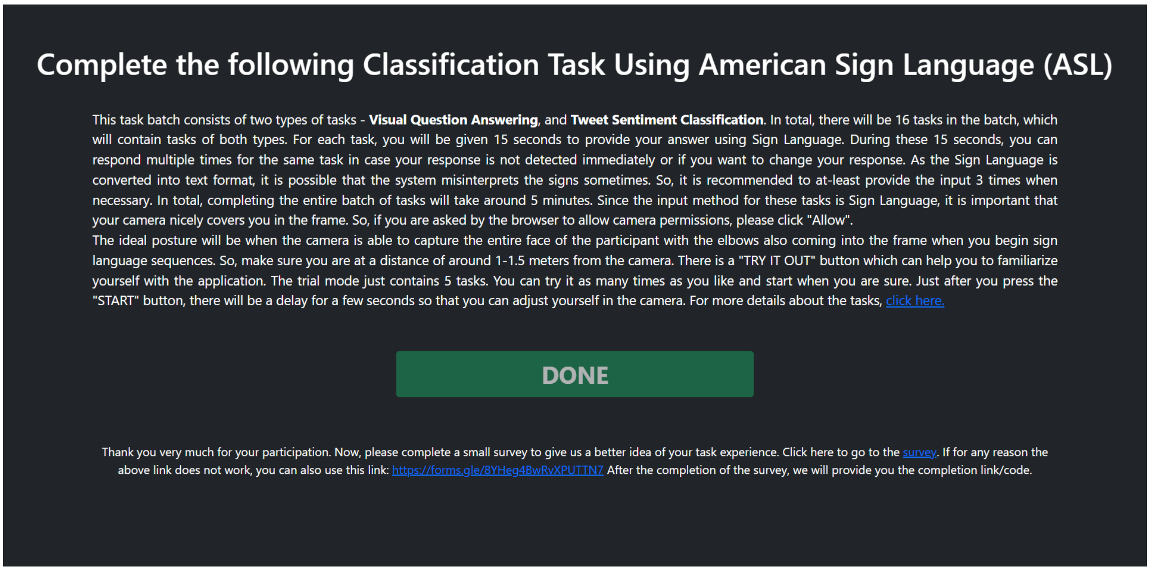
Timer: 15

Kate Upton & Justin Verlander -- To Hell With Cy Young ... It's Vatica #TrumpTransition #PresidentElectTrump #MAGA

Do you think this statement is Positive, Negative or Neutral?

Response:Neutral

Figure B.3: SignUpCrowd Task section



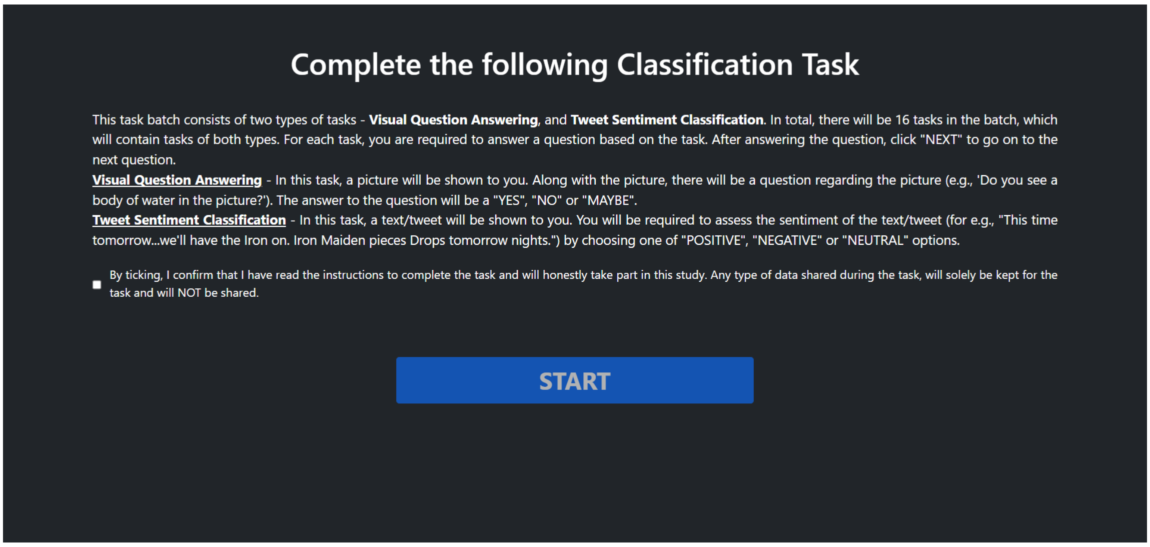
Complete the following Classification Task Using American Sign Language (ASL)

This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you will be given 15 seconds to provide your answer using Sign Language. During these 15 seconds, you can respond multiple times for the same task in case your response is not detected immediately or if you want to change your response. As the Sign Language is converted into text format, it is possible that the system misinterprets the signs sometimes. So, it is recommended to at-least provide the input 3 times when necessary. In total, completing the entire batch of tasks will take around 5 minutes. Since the input method for these tasks is Sign Language, it is important that your camera nicely covers you in the frame. So, if you are asked by the browser to allow camera permissions, please click "Allow".
The ideal posture will be when the camera is able to capture the entire face of the participant with the elbows also coming into the frame when you begin sign language sequences. So, make sure you are at a distance of around 1-1.5 meters from the camera. There is a "TRY IT OUT" button which can help you to familiarize yourself with the application. The trial mode just contains 5 tasks. You can try it as many times as you like and start when you are sure. Just after you press the "START" button, there will be a delay for a few seconds so that you can adjust yourself in the camera. For more details about the tasks, click here.

Thank you very much for your participation. Now, please complete a small survey to give us a better idea of your task experience. Click here to go to the survey. If for any reason the above link does not work, you can also use this link: https://forms.gle/8YHeg4BwRvXPUTTN7 After the completion of the survey, we will provide you the completion link/code.

Figure B.4: SignUpCrowd completion page

B.2. Text and Click input type



Complete the following Classification Task

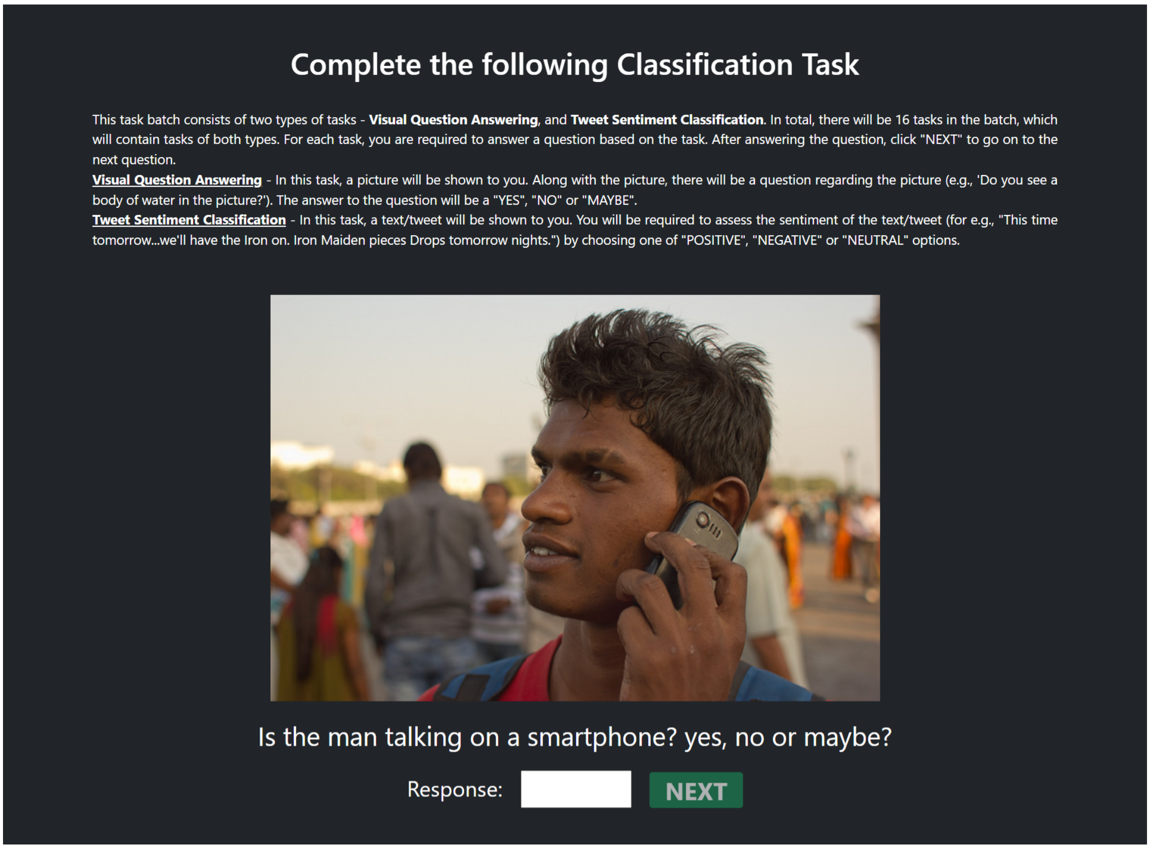
This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you are required to answer a question based on the task. After answering the question, click "NEXT" to go on to the next question.

Visual Question Answering - In this task, a picture will be shown to you. Along with the picture, there will be a question regarding the picture (e.g., 'Do you see a body of water in the picture?'). The answer to the question will be a "YES", "NO" or "MAYBE".

Tweet Sentiment Classification - In this task, a text/tweet will be shown to you. You will be required to assess the sentiment of the text/tweet (for e.g., "This time tomorrow...we'll have the Iron on. Iron Maiden pieces Drops tomorrow nights.") by choosing one of "POSITIVE", "NEGATIVE" or "NEUTRAL" options.

□ By ticking, I confirm that I have read the instructions to complete the task and will honestly take part in this study. Any type of data shared during the task, will solely be kept for the task and will NOT be shared.

Figure B.5: Home page for text and click input web application



Complete the following Classification Task

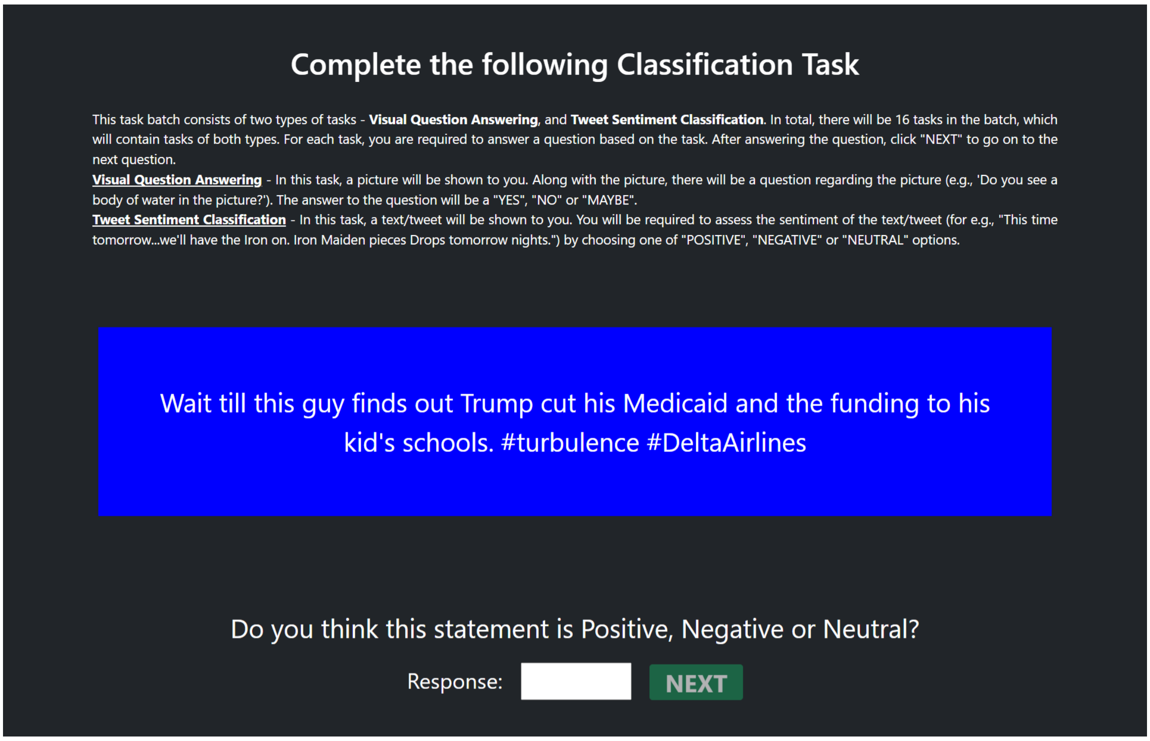
This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you are required to answer a question based on the task. After answering the question, click "NEXT" to go on to the next question.

Visual Question Answering - In this task, a picture will be shown to you. Along with the picture, there will be a question regarding the picture (e.g., 'Do you see a body of water in the picture?'). The answer to the question will be a "YES", "NO" or "MAYBE".

Tweet Sentiment Classification - In this task, a text/tweet will be shown to you. You will be required to assess the sentiment of the text/tweet (for e.g., "This time tomorrow...we'll have the Iron on. Iron Maiden pieces Drops tomorrow nights.") by choosing one of "POSITIVE", "NEGATIVE" or "NEUTRAL" options.

Is the man talking on a smartphone? yes, no or maybe?

Figure B.6: Text input: VQA task



Complete the following Classification Task

This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you are required to answer a question based on the task. After answering the question, click "NEXT" to go on to the next question.

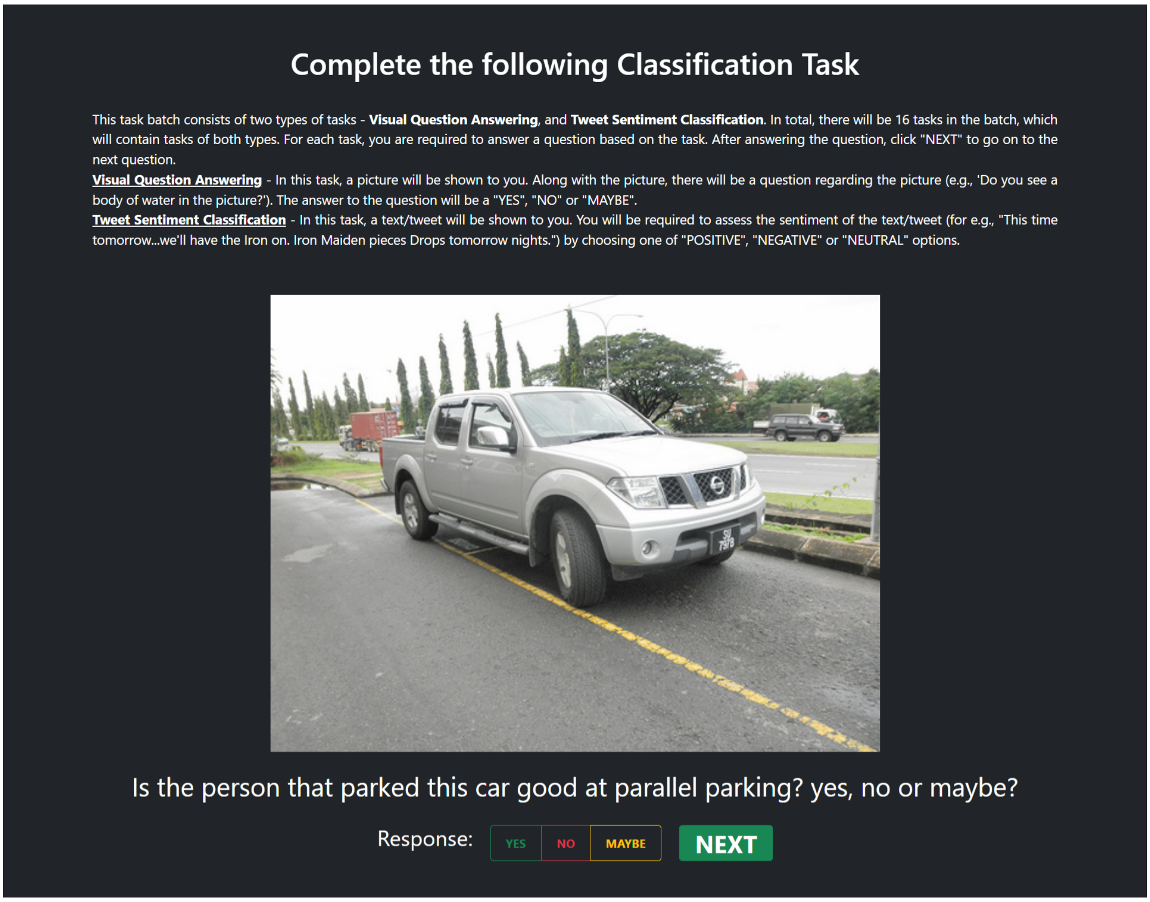
Visual Question Answering - In this task, a picture will be shown to you. Along with the picture, there will be a question regarding the picture (e.g., 'Do you see a body of water in the picture?'). The answer to the question will be a "YES", "NO" or "MAYBE".

Tweet Sentiment Classification - In this task, a text/tweet will be shown to you. You will be required to assess the sentiment of the text/tweet (for e.g., "This time tomorrow...we'll have the Iron on. Iron Maiden pieces Drops tomorrow nights.") by choosing one of "POSITIVE", "NEGATIVE" or "NEUTRAL" options.

Wait till this guy finds out Trump cut his Medicaid and the funding to his kid's schools. #turbulence #DeltaAirlines

Do you think this statement is Positive, Negative or Neutral?

Figure B.7: Text input: TSA task



Complete the following Classification Task

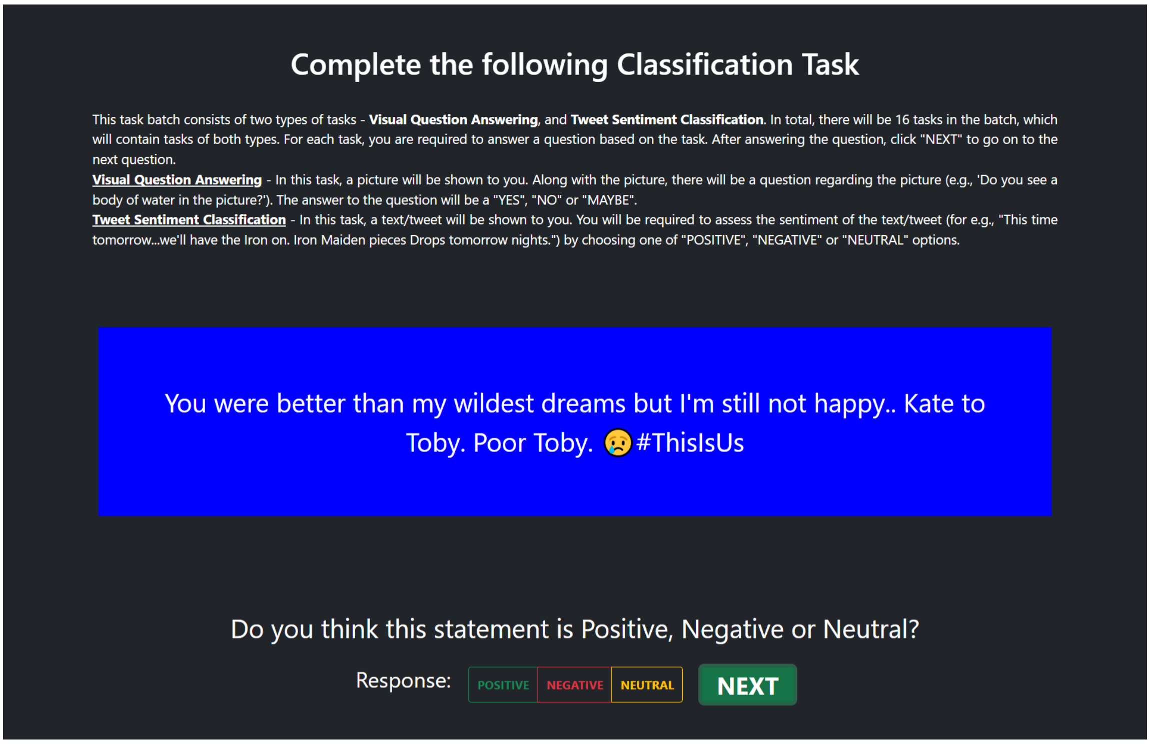
This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you are required to answer a question based on the task. After answering the question, click "NEXT" to go on to the next question.

Visual Question Answering - In this task, a picture will be shown to you. Along with the picture, there will be a question regarding the picture (e.g., 'Do you see a body of water in the picture?'). The answer to the question will be a "YES", "NO" or "MAYBE".

Tweet Sentiment Classification - In this task, a text/tweet will be shown to you. You will be required to assess the sentiment of the text/tweet (for e.g., "This time tomorrow...we'll have the Iron on. Iron Maiden pieces Drops tomorrow nights.") by choosing one of "POSITIVE", "NEGATIVE" or "NEUTRAL" options.

Is the person that parked this car good at parallel parking? yes, no or maybe?

Figure B.8: Click input: VQA task



Complete the following Classification Task

This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you are required to answer a question based on the task. After answering the question, click "NEXT" to go on to the next question.

Visual Question Answering - In this task, a picture will be shown to you. Along with the picture, there will be a question regarding the picture (e.g., 'Do you see a body of water in the picture?'). The answer to the question will be a "YES", "NO" or "MAYBE".

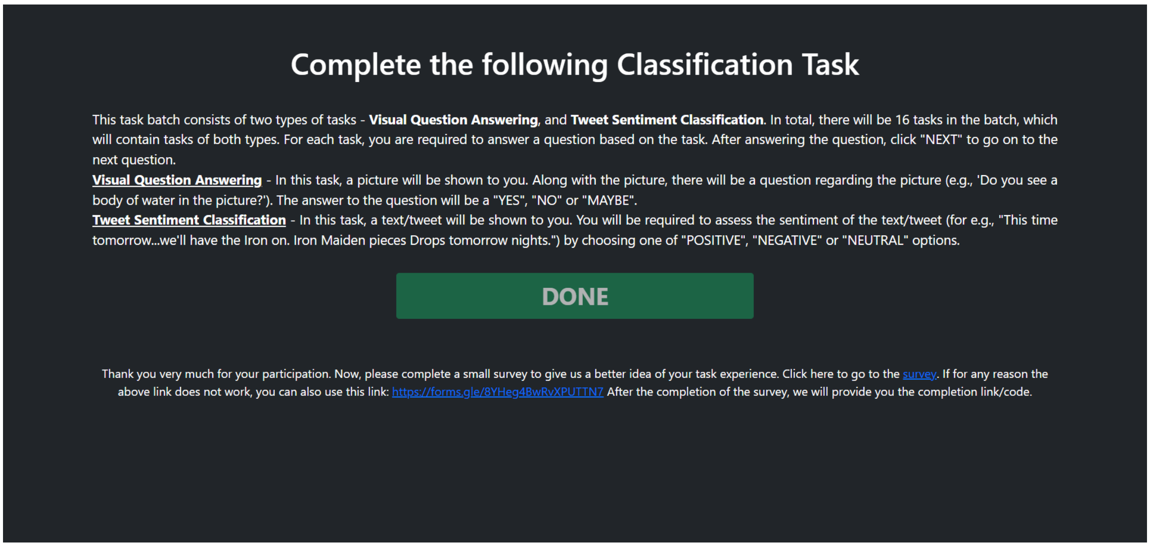
Tweet Sentiment Classification - In this task, a text/tweet will be shown to you. You will be required to assess the sentiment of the text/tweet (for e.g., "This time tomorrow...we'll have the Iron on. Iron Maiden pieces Drops tomorrow nights.") by choosing one of "POSITIVE", "NEGATIVE" or "NEUTRAL" options.

You were better than my wildest dreams but I'm still not happy.. Kate to Toby. Poor Toby. 😟 #ThisIsUs

Do you think this statement is Positive, Negative or Neutral?

Response:

Figure B.9: Click input: TSA task



Complete the following Classification Task

This task batch consists of two types of tasks - **Visual Question Answering**, and **Tweet Sentiment Classification**. In total, there will be 16 tasks in the batch, which will contain tasks of both types. For each task, you are required to answer a question based on the task. After answering the question, click "NEXT" to go on to the next question.

Visual Question Answering - In this task, a picture will be shown to you. Along with the picture, there will be a question regarding the picture (e.g., 'Do you see a body of water in the picture?'). The answer to the question will be a "YES", "NO" or "MAYBE".

Tweet Sentiment Classification - In this task, a text/tweet will be shown to you. You will be required to assess the sentiment of the text/tweet (for e.g., "This time tomorrow...we'll have the Iron on. Iron Maiden pieces Drops tomorrow nights.") by choosing one of "POSITIVE", "NEGATIVE" or "NEUTRAL" options.

Thank you very much for your participation. Now, please complete a small survey to give us a better idea of your task experience. Click here to go to the survey. If for any reason the above link does not work, you can also use this link: https://forms.gle/8YHeg4BwRvXPUTTN7 After the completion of the survey, we will provide you the completion link/code.

Figure B.10: Completion page for text and click input web application