Sign Language Interpreter

Chantal Ariu[1], Timothé Van Damme[2], Ividő Fehér[3], Helin Demi[4], Bram Verhaar[5], and Temel Kevin Gur[6]

Vrije Universiteit Amsterdam, De Boelelaan 1105, Netherlands
https://vu.nl/en/education/bachelor/artificial-intelligence

Abstract. Despite the increase in popularity of web conferences, the accessibility features for individuals from the Deaf and Hard of Hearing (DHH) community remain limited. This paper aims to create a comprehensive research plan for the creation of a real-time sign language interpretation tool aimed at bridging the gap between individuals from the DHH community and people not belonging to the community in the domain of web conferences. In order to ensure that this paper is following the standards of up-to-date accessibility research, a comprehensive literature review has been conducted within this paper. The tool incorporates numerous artificial intelligence techniques such as Convolutional Neural Networks (CNNs) for facial expression recognition and Support Vector Machine (SVM) models for hand gesture recognition. This research paper also includes a detailed plan in regard to the system verification process, including both internal and external verification, to ensure that the requirements are met as well as user satisfaction being at a high rate. The success of the development of this tool is significant as the expected results from the system validation process are highly positive, meaning that the tool would significantly improve accessibility standards in the domain of web conferencing for individuals that are part of the DHH community. Recommendations and motivations for future work are also thoroughly discussed since the domain of predictive artificial intelligence being incorporated in tools such as the one presented in this paper has not yet been documented and should be a focus for future research in the domain.

Keywords: Sign-Language Interpretation · Web Conferences · Convolutional Neural Networks · Support Vector Machine

1 Introduction

During video conferences, there are already existing tools that are able to create subtitles in real time, in order for deaf people to follow a conversation where hearing people and deaf people have to communicate together online, through online conferencing platforms for example. A tool like that helps the hearing-impaired to understand what hearing people are saying in real time, however, the opposite direction is not quite considered, resulting in the hearing-impaired having to communicate by text. This creates a lag in the conversations since text as a communication tool is not as real-time as speech or sign. The result of this is that the virtual conversation runs less smoothly and it is more difficult for deaf people to participate in the conversation in real time.

The goal is to create a tool that could be used during virtual conferences (e.g. a plugin), which would be able to interpret sign language using the web camera of the signing person. The tool would create an audible response that translates sign language into audio that is then played during the virtual conversation with minimal delay time for all of the hearing participants to understand. Additionally, it would be possible to integrate the tool seamlessly with existing video conferencing platforms (e.g. WebEx, Zoom).

This tool would be able to help deaf people to communicate effectively and naturally during conversations in real time, avoiding any lags created by texting due to our tool's ability to instantly generate audible responses. The result would be a more smooth and overall more pleasant experience during virtual conversations, which is supported by the fact that deaf people would be given the ability to express themselves naturally, creating a potentially more engaging environment for both deaf and hearing people during the conference.

The AI solution for this tool would involve collecting data and gathering large datasets of people using sign language, specifically video material and corresponding written translations. From then, we would need to collect relevant hand positions and points from the datasets so that the video material can be translated into numerical/ mathematical format. This numerical/ mathematical data would then be used to train machine learning models in a way that certain hand movements and positions will be associated with a correct written interpretation. AI will also be used for helping to predict the sentence, since in order for the real-time aspect of the interpretation to work, AI needs to interpret and generate the signs right as they are performed, rather than having to wait for the full sentence to be signed. Predictive AI already exists for that, however, more training would be needed in terms of applying those algorithms in relation to sign language.

In conclusion, this tool is crucial also due to the number of deaf people in the world. 70 million people in the world are deaf, however, most hearing people do not understand sign language, creating a divide between those groups that harms communication significantly. In order to bridge this divide, our tool could help when communication happens between deaf and hearing people. Possible issues include that every person has a different way of signing, meaning that the program would need to recognize slight variations that might be added when signing. Moreover, multiple sign languages exist which the machine would have to differentiate between and correctly interpret it to the corresponding audio language. However, for the sake of limiting our scope, this tool would start with the development of ASL (American Sign Language) first, due to it containing the largest data sets.

2 Literature Review

2.1 Addressing the Need for a Literature Review

In order to answer and explore the research question of the extent to which image recognition and predictive artificial intelligence can be used to create a real-time sign language interpretation tool for web conferences, a review of the literature that has been published up to this point is necessary in order to ensure the understanding of that prior body of work. This is especially important when discussing the technological feasibility of the proposed tool using the proposed technologies as well as addressing the research question with the most detail possible and, importantly, within the context of the existing literature in the field of sign language recognition and interpretation. Moreover, analyzing prior literature allows for thorough consideration of the performance of benchmarks within the field, allowing the scope of this paper to be within realistic measures.

2.2 Research Methodology

In the following part, the methodology for finding appropriate literature to review will be outlined in order to create protocol of how research was conducted in order to conduct an appropriate literature review for this particular paper. This will insure that the process has high reproducibility as well as credibility due to its documented nature.

Inclusion and Exclusion Criteria In order to conduct a coherent and precise literature review, clear inclusion as well as exclusion criteria must be defined to minimize ambiguity regarding which papers are selected to be included in this literature review. Through inclusion and exclusion criteria, the collection of research stays consistent and reduces bias due to the transparency of having to state clear criteria for inclusion and exclusion.

For this particular research paper which aims to explore the ability to use image recognition and predictive artificial intelligence technologies to create a tool for real-time sign language interpretation, the following inclusion criteria were identified: direct relevance of the literature to the research topic of this paper, in order to ensure a focused literature review; publication date of the research being within the past five to ten years, in order to ensure that the most recent developments and advancements in technology are taken into consideration; types of literature, preferably peer-reviewed journal articles and conference papers to ensure academic rigor within the field of computer science; language, namely English, as this is the common language shared between the people working on this research paper. As opposed to these inclusion criteria, the exclusion criteria defined for the collection of research are the following: irrelevant topics, specifically no research that does not directly relate to the research question or the research topic in general should be included to keep the literature review focused; outdated sources, namely sources published prior to 2013 will be filtered out and excluded as it might not reflect and consider recent advancements in technology that would be detrimental for this research papers' purpose to leave out; research from non-academic sources, in particular sources that do not meet the required standards in academic rigor, such as blog posts, news articles, or promotional materials. The list of these inclusion and exclusion criteria will ensure that the research collected for the literature review aligns with the research objectives of this paper while simultaneously meeting the standards of academic research conducted and published within the field of computer science and engineering.

Identification of Keywords and Databases Following the creation of the inclusion and exclusion criteria, it is possible to move on to the identification of search keywords and databases. The keywords were selected according to the research question, using synonyms and related methods. In order to create a systematic approach to the list of keywords, it would be most efficient to group the keywords based the level of synonymity that they have. The following table includes all of the keywords grouped into four categories: Sign Language Interpretation, Techniques, Accessibility and Application Context, and Specific Technologies.

Table 1. Search Keyword Grouping

Group 1	**Group 2**	**Group 3**	**Group 4**
Sign Langauge Interpretation	Image Recognition	Real-Time Communication Accessibility	Convolutional Neural Networks for Sign Language
Sign Language Translation	Predictive Artificial Intelligence	Accessibility Tools for Webinars	Recurrent Neural Networks for Sign Language
Real-Time Sign Language Translation	Machine Learning for Sign Language	Accessibility in Web Conferencing	Sign Language Recognition Software
AI-Based Sign Language Interpretation	Gesture Recognition for Communication	WebRTC for Real-Time Communication	Natural Language Processing for Sign Language
	Deep Learning for Sign Language	Web Conferencing	

After creating the key words, a selection of databases for the search was created based on the reputation of the database, the credibility, as well as the academic rigor of the research found in those databases. The databases selected for this particular literature review are: ACM Digital Library, IEEE Xplore, Google Scholar, Science Direct, SpringerLink, and Scopus.

Search Strategy After defining all of the necessary criteria and key words, the literature search can be conducted in a coherent manner by combining the key words in specific ways to get numerous results on the multiple databases selected. Table 1 will result in a precise definition of search terms (”strings”) that can be used to find literature combining each keyword from every group. This can be done by using boolean operators such as the ‘and‘ ($\wedge$), ‘or‘ ($\vee$) operators, specifically using the ‘and‘ operator between each group and the ‘or‘ operator between key terms within each group. Naturally, following this strategy will result in duplicates of research papers published on multiple databases, in which case the paper published on the most reputable database will be included in this literature review. Additionally, literature that in itself consists of a literature review will be used through the ”cited by” as well as ”citations” to find further literature potentially relevant to the research topic.

2.3 Findings

In this section, the findings of the literature search will be combined and analyzed in the context of a traditional literature review.

Findings as early as in 2014 show the interest of engineers and scientists in creating human body tracking devices for various reasons. Chen Qian, et al.[1] mentioned that their study, aiming to create a realtime hand tracking system using depth sensors, was ”[...] largely inspired by recent advances in human body tracking”, suggesting the need and demand of such technology almost ten years ago. Chian, et al.[1] specifically created an unprecedented low cost but high efficiency model for accurate and fast hand tracking that could be useful in various areas, including sign language recognition. However, due to the lack in datasets large enough for experiments and testing of their model, the researchers created their own dataset by asking six subjects to perform certain gestures rapidly, resulting in a 400-frame video sequence being recorded for each subject. By today’s standards, the size of this dataset would raise concerns regarding bias and inclusivity of various hand shapes, which is also mentioned by the researchers ”To account for different hand sizes, a global hand model scale is specified for each subject [...] but no further personal adaptation is used.”. Despite these limitations, Chian, et al.[1] have contributed to the advancement in technology for body, specifically hand, tracking devices by introducing a, for its time, efficient and low-cost hand-tracking system. Another focus during that time period within the context of sign language interpretation tools were the analysis of existing and emerging databases. Sahoo, et al.[2] performed a review on various databases for sign language, for example, video material, transcripts and different sign languages and specifically analyzed how they were built, how the data was collected, as well as an overall evaluation of the databases. This review created a strong groundwork and for future work on sign language interpretation since it transparently describes the methodology of the data collection as well as the advantages and limitations of the databases. In more recent years, reviews on the field of sign language interpretation using face recognition have become more specific, since techniques like SVM (supper vector machine), CNN (convolutional neural network), and PCA (principle component analysis) grew in popularity amongst researchers within this field [1]. Modi and Patel [3] conducted a review of recent advancements, examining their performance when faced with challenges such as illumination variation, pose variation, occlusion, expression variation, low resolution, and ageing. The results conclude in a concise manner that all of the recent popular technologies have an accuracy percentage of at least 80%, highlighting the improvement as well as advancement in general, but also domain-specific, technology. The first full systematic literature review followed shortly after and was conducted by Wadhawan et al. [4] on the topic of sign language recognition systems. In this systematic literature review, a comprehensive explanation on how sign language is constructed is included, specifically the difference between one-handed and two-handed, as well as the difference between manual and non manual elements. The latter is particularly important since it poses a direct challenge to face recognition strategies. Non-manual signs do not solely rely on hands to express signs but also on body posture and direction, mouth positions and facial expressions. The interpretation of those particular features in combination with the sign poses a challenge that is still being tackled in today's research. One example of how important the addressing of this challenge is, is outlined in the research paper by Pataca et al. [5], where the main issue addressed is the lack of emotion and expressiveness through text. As discussed extensively before, one of the main motivations for research within the specific scope of this papers' research question is due to the inaccessibility that expressing oneself during text results in. When a DHH (deaf or hard of hearing) individual expresses themself over text exclusively, crucial emotion and tone gets lost. This is why Pataca, et al.[5] proposes a model to capture emotion and use artificial intelligence trained for capturing emotions from verbal communication to then translate it into a more expressive and emotional text message. Although this does not in particular align with this research papers' aim, the technology used for emotion-recognition in the speaking individuals are a great contribution to the development of a tool that aims to interpret facial expressions and emotions from DHH individuals while they are signing, into text. Research relating accessibility features for DHH individuals continued emerging, Bastas, et al.[6] in particular examined other areas which needed more attention in order for increased inclusion of DHH individuals in cultural events. Although this research paper focuses on accessibility for DHH individuals in the context of web conferences, it is important to acknowledge the issue of accessibility for such groups of individual in a broader scope as well, which is why the research conducted by Bastas, et al.[6] is relevant for the purpose of this paper. Bastas, et al. particularly focused on DHH individuals' ability to be part of live theater performances, leading them to develop a solution through synchronized subtitles using automatic speech recognition (ASR), as well as natural language processing (NLP). Both of these techniques are highly relevant for the purpose of this research paper since there is a significant element of speech generation as well as speech interpretation, which could potentially use NLP techniques. Another technique that has proven itself to be very useful in the domain of sign language recognition and interpretation is that of convolutional neural networks (CNNs) which is designed for the processing of structured grid data, making it best applicable in image analysis. Through using specific layers, referred to as convolutional layers, it is able to automatically recognize patterns and learn features from input data, making it a very efficient tool for image recognition in general[3,7,8]. Kumar et al. [7] created a system for 3D sign language recognition which uses color coded topographical descriptors to classify the input data using a two stream CNN architecture. Through specifying different types of inputs, specifically joint distance topographical descriptors (JDTD) and joint angle topographical descriptors (JATD), it is possible to obtain, for a given 3D sign query, a list of class scores as text labels corresponding to particular signs. This allows for precise measurement in efficiency and performance of the model. The specific technique presented in the paper written by Kumar et al. shows an improvement in classifier performance compared to predecessors. These results are further validated by the specific experimental procedure utilizing multiple datasets to test the model. CNNs are a valid technique that has been improving in performance steadily, specifically when applied in the context of sing language interpretation. Rastgoo, et al.[9] came to the conclusion that sign language interpretation models using CNN in 2020 had an accuracy rate of 99.72, meaning it is a highly reliable as well as accurate technique to include when building a tool to interpret sign language. Various other techniques are also mentioned, however the most important technique besides CNNs that is applicable in this context is using vision based models of sign language recognition combined with deep learning. This is particularly relevant considering the advancements in deep learning to-date, as it could directly influence the extent to which it is possible to not only interpret sign language as fast and naturally as possible but also generate a speech response in real time. Additionally, research conducted by Munir, et al. [10] tackles a similar issue as the one aimed to be explored within this paper, specifically by developing a useful and easily usable tool for effortless communication between DHH and non-DHH individuals. The researchers specifically developed an intelligent sign language interpretation system to connect hearing people with the DHH community, ultimately creating a two way correspondance between speech impaired and regular speaking people. Specific steps such as image preprocessing were outlined and what is important to note is that Munir, et al.[10] used a Support Vector Machine (SVM) due to its performance with data that is semi- or unstructured. The researcher were in the end able to create a tool that fosters two-way-communication between DHH and verbal individuals.

2.4 Gaps and Limitations

After summarizing and briefly analyzing ten examples of relevant literature, certain gaps and limitations are important to identify. In particular, considering the datasets used for each of the studies, a major limitation is the representation of different skin colors, especially when considering research that has been done for face recognition using color detection and RGB [7]. This is not the only scenario in which bias is able to be introduced due to datasets not being representative enough of other variations, for example proper lighting needing to be present as well as a stable internet connection for the system to work [7]. As for gaps in research and how this emerging paper will contribute to this already existing, extensive, body of work around sign language interpretation is the real-time interpretation aspect. So far and to our knowledge, the role of predictive artificial intelligence technology in the real-time interpretation of sign language has yet to be defined. For the progression of this paper, it will be one of the main objectives to explore the existing gap in documented literature and report on the findings as well as the feasability of using predictive artificial intelligence in real-time sign language interpretation.

2.5 Conclusion

In conclusion, the technology for sign language interpretation has followed a historical development from early hand-tracking systems to more recent advancements using face recognition and deep learning techniques. This also results in a steadily increasing efficiency and performance, in terms of accuracy, of the models which is significant to the development of a tool of such importance.

It has also been discussed how relevant and important the inclusion and accessibility technologies for the DHH community is to their integration into every-day life with non-DHH people. Various aspects of such accessibility features were discussed, not exclusive to the research question due to its importance, such as sign language recognition, emotion capture and interpretation, and live event accessibility features.

Limitations and gaps found while conducting the literature for this paper were also included since it is important to acknowledge limitations of other research in order to be aware of the possible limitations that this paper will face while emerging. Issues such as datasets that are not representative enough was one of those limitations found in majority of the literature reviewed as well as predictive artificial intelligence technologies not being a part of the development in sign language interpretation in real time.

Successfully building upon prior research while incorporating and considering possible advancements, such as with predictive artificial intelligence technologies for a real time sign language interpretation would result in a positive societal impact in terms of inclusivity of the DHH community as well as general accessibility.

3 Methodology

Based on the aim of this paper, namely to find the extent to which image recognition and predictive artificial intelligence technologies can be used to create a real-time sign language interpretation tool for web conferences, a technical approach will be detailed in the following section.

The main challenge when finding which technologies to use specifically regarding sign language interpretation is the extensive amount of good-performing methods available which show great results for domain specific tasks. For example, convolutional neural networks (CNNs) have shown to outperform many other methods available in the domain of feature extraction [11].

A convolutional neural network uses multiple layers, namely an input layer, which marks the start of the network and the first layer, followed by (usually) multiple convolutional layers, which is responsible for operations such as applying filters or kernels to data from the input layer, usually followed by a ReLU (Rectified Linear Unit) layer, which is an activation function that ensures activation of a node only when a certain value threshold is reached, reducing the number of operations, and therefore, the complexity of the learning, creating a more efficient network. The ReLU layer also introduced non-linearity which is crucial to tackling the challenge of the vanishing gradient when training the neural network [11]. Another hidden layer that usually follows the ReLU layer is the pooling layer, the objective of which being to decrease the spatial complexity of the rectified map, essentially compressing the results of the ReLU layer in order to have more compact data to work with in the following layers. The last layer, namely the output layer, contains a function which is used to predict error and further train the model [11,8].

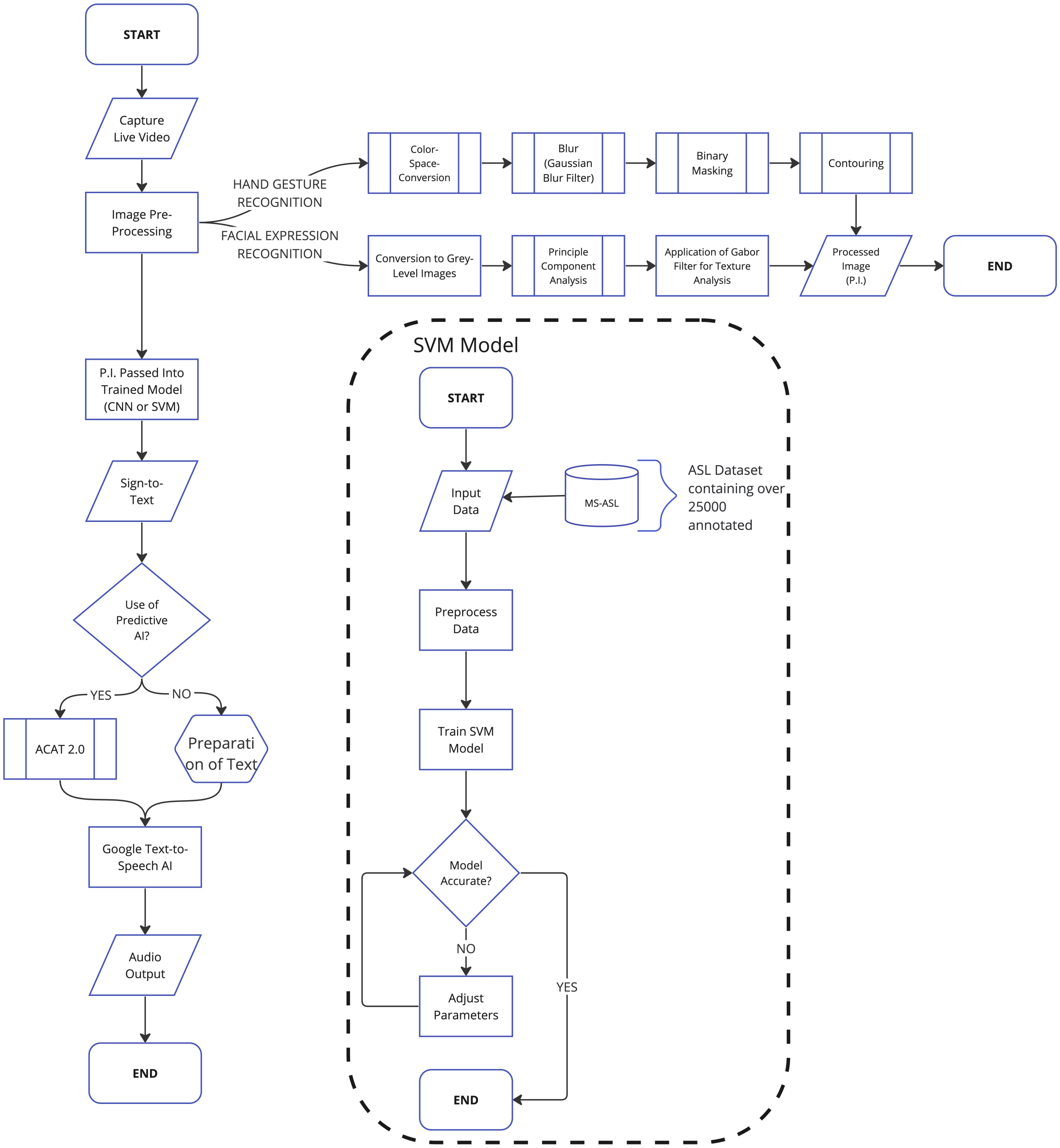
However, convolutional neural networks have shown issues when used in sign language interpretation in previous literature, specifically, "[...] recognition of CNN architectures degrades, when characters have very high interclass similarities." [11]. In order to counter this issue, the proposed solution will not be fully reliant on convolutional neural networks, but rather utilize CNNs for what they perform the best in, while using a different model specialized in the remaining field.

In this paper, CNNs will be used exclusively for facial expression recognition, alongside Local Directional Patterns (LDPs), as opposed to Local Binary Patterns (LBPs), since the LDP variant in combination with CNNs has shown higher performance and accuracy in recognizing facial expressions than LBPs [8]. For recognizing hand gestures, which is the remaining part of developing a sign language interpreter, a Support Vector Machine (SVM) technology will be used, since it has shown success in prior research when it comes to working well with unstructured as well as semi structured data, resulting in a higher accuracy even in signs with high similarities [1].

An SVM model works by receiving input in the form of two data points, from which a closer metric is then returned. If the data points do not have a large distance, the value of the SVM kernel will increase, whereas it would decrease the farther the data points are from each other. In order to avoid high costs in data transformations, similarity metric scores from higher dimensions can easily be found in the SVM kernel [1]. The same researchers already conducted an SVM kernel analysis, which is done by comparing linear, polynomial, Radial Basis Function (RBF), and sigmoid kernels. The results of these comparisons concluded with the RBF kernel at $c = 2.67$ and $\gamma = 5.383$ providing the highest accuracy when tested on the given dataset, which is the reason for why this tool will adopt these measurements when training the model using SVM [1].

Before the input images are given as input to the SVM model, the images are preprocessed using Histograms of Oriented Gradients (HOG), which essentially refers to a technique used to transform raw pixel values into a more meaningful representation of the structure within the image. The technologies chosen for the text-to-speech part of the sign language interpretation tool are similar, whether the user chooses to use predictive artificial intelligence or not, since the Google Text-to-Speech AI provides hundreds of options for natural sounding voices, which they achieved by using various natural language processing deep learning technologies. However, the difference is that if a user chooses not to use predictive artificial intelligence, the TTS AI will activate only when the user has finished signing, as opposed to being ”real time”. Due to the error rate, a user might still prefer the lag over potential mis-prediction, which is why the option exists. In the other scenario, ACAT2.0 could be integrated in order to create a real-time interpretation experience [12]. It is possible to abstract certain parts of the technological approach visually in order to create a flowchart which can be seen in Figure 1. This flowchart is heavily abstracted and provides the function of visualization of the model on a larger scale. In Figure 1, the processes for image pre-processing are elaborated on, differentiated between hand gesture recognition and facial expression recognition. Some choices were made specifically in order to optimize the experience of this tool but also to minimize excessive and unnecessary computation, for example converting the color-space, which refers to converting the original color-space of the image (in this case, RGB) to the YCrCb color-space. This decision is made based on prior findings concluding that face detection was significantly improved when the color space of an image was converted into the YCrCb color space [13]. In the YCrCb color space, luminance (Y) is separated from Chrominance (Cr, Cb), specifically the Cr (Red Difference) component, which represents the difference between red and luminance, and Cb (Blue Difference), which represents the difference between the blue component and the luminance. The advantage this results in is that the separation of luminance and chrominance, consequently allowing luminance to be stored at a higher resolution than chrominance, allows for more efficient image compression.

Fig. 1. Flowchart



In conclusion, this section has detailed the technological methods and AI approaches that this paper will focus on while delving into the question of the extent to which is image recognition and predictive artificial intelligence can be used to create a sign language interpretation tool. A summary of the key technologies can be found in the flowcharts presented in Figure 1, but are as follows: Facial expression recognition will be performed using a trained CNN paired with LDPs whereas hand gesture recognition will be performed using a trained SVM model. The image pre-processing steps are also domain specific, as outlined in the flowchart, due to differences in requirements for top performance depending on the model.

4 Experimental Evaluation

Following the extensive explanation of the methodology planned to be followed for the hypothetical development of this tool, it is important to create a test plan for that system. This section will therefore aim to create a detailed plan for the system verification, also referred to as internal evaluation, system validation, also referred to as external evaluation, as well as for the hypothesis testing. It is very important to note that the nature of this section is highly hypothetical due to the testing not actually being conducted and therefore no actual testing results being recorded within the scope of this paper. It is important to acknowledge the significant limitations that this section entails at the beginning, due to the hypothetical nature of this paper not allowing to conduct actual testing during the development of the research question.

4.1 System Verification

This subsection will focus on the internal evaluation, or also called system verification. The verification of the system is crucial for identifying the functionality of each component by itself. The hypothetical testing process for the system verification will be as follows.

Experimental Setup In order for the hypothetical testing to take place, experimental setup needs to be discussed and well documented in order to be replicable, increasing the credibility of the verification results. For the experimental setup of this testing plan, specific hardware configurations would need to be detailed in order to proceed. Since this system is aimed at a target audience mostly working using laptops or PCs and therefore web cameras, the minimum resolution for the web camera system should be 720p or higher[10] in order to ensure the proper capturing of the sign language gestures. Moreover, the CPU should have at least 8 cores, and the GPU at least 16 cores, which are common standard for most working laptops and PCs. As for the software that will be used for testing, different options are available in terms of commonly used web conferencing software applications that the tests could be performed on, such as Zoom, WebEx, Google Meet, to name a few of the most popular ones at this stage. Due to each of the available software applications used for web conferencing being different and having different methods of integrating plugins or other related add-ons, focusing on the three mentioned software applications for initial testing would be sufficient for the scope of testing the real time sign language interpretation tool. As for other software dependencies that are expected to be used for the purpose of internal verification of the system would be databases. The dataset that would be used for verification purposes cannot be the same one that was used for training the machine learning models in order to simulate an environment during which the tool would be used, but needs to meet several other requirements. The basic requirement is that there is video material with a corresponding interpretation, but also that the video material in the dataset is representative of a diverse population, including, but not limited to, different skin colors, ages, genders, as well as accessories. Moreover, a dataset containing meaningless gestures should be included in order to verify that the system is able to differentiate between meaningful and meaningless gestures.

Verification Methodology For the methodology concerning system verification, it would be best to test components in isolation. For component testing specifically, unit testing would be the most appropriate choice, as it allows for components to be tested in isolation[14,15]. For the purpose of this tool, components such as facial expression recognition, hand gesture recognition, as well as language translation can be tested in isolation. The use of unit testing enables for early detection of issues since they would be written as part of the development process of the tool, therefore being executed early in the process, which would allow issues to be discovered and addressed as early as possible. Due to the very focused nature of unit tests, the verification performed would be very clear since a unit's behavior is compared to expectations of the behavior of the tested unit, allowing for the smallest computation units to be confirmed to work as expected [15]. In order to maximize the efficiency of the testing while also ensuring that this verification process would be as credible as possible, the unit testing would be combined with automation testing in order to allow for end-to-end testing, integration testing, as well as regression testing[16]. These test methods are different from unit tests, since they can focus on a larger aspect, end-to-end testing specifically, is a type of automation tests that, for the scope of this paper, would be used to simulate user interactions between DHH and non DHH people and verify that the tool would work as expected, while integration testing would be used to isolate the testing of different components of the proposed system working together as expected and seamlessly. Lastly, regression testing would be used to ensure that changes made in the system, whether necessary additions or for fixing of bugs, would not introduce new bugs. To go into further detail on how integration testing would be used for the system verification of the real time sign language interpretation tool, specific application areas will be addressed. On a smaller, more specific scale, integration testing would be used to ensure that the sign language interpretation components, namely facial expression recognition and hand gesture recognition work together seamlessly and as expected. However it would also be used to test the compatibility between larger components, such as the image recognition component as a whole, the predictive artificial intelligence module, as well as the audio generation module and whether they work together seamlessly and as expected. Another area in which integration testing would be used is to test the integration of the developed tool with web conferencing platforms, since the tool needs to be integrated with popular web conferencing platforms, such as the ones mentioned above (WebEx, Google Meet, Zoom). Through the integration test on this particular component, one of the challenges, namely the tool working together with the external platform APIs would be addressed. Moreover, in order to expand the tool beyond supporting just American Sign Language, integration testing would be used for cross-language integration in the future work for this tool. Moreover, when adding new support for various other sign languages, integration testing would allow for ensuring that the tool adapts to language-specific nuances while also ensuring that the image recognition and interpretation components match the expectations and requirements of the newly supported sign language.

Expected Metrics In order to properly evaluate tests, metrics have to be predefined. This section will therefore outline the expected metrics that would be used if actual testing for the real time sign language interpretation tool was to be conducted. The hypothetical metrics to be used in this paper are the accuracy of sign language recognition and interpretation, real-time responsiveness, cross-language adaptability, cross-platform integration, and predictive artificial intelligence accuracy. The expected metric for the accuracy of sign language recognition and interpretation, which would be measured as the percentage of correctly recognized and interpreted sign language gestures, is above 90%, which is considered high-performing among current benchmark comparisons[4,7,9,10]. Real-time responsiveness is the metric that would be used to evaluate the tool's ability to provide a real-time sign language interpretation without any major delay, if the predictive artificial intelligence technology is part of the interpretation. The expectations for this aspect vary since no actual testing can be performed. Cross-language adaptability refers to the system's ability to adapt to different sign languages, initially adapting to ASL but showing the potential for extending to other various sign languages. Cross-platform integration ensures the ability for the system to work and be compatible with various web conferencing platforms and due to the heavily documented nature and available APIs for such web conferencing platforms, seamless integration with the most popular service is expected without any significant variation in the performance of the tool. Lastly, predictive artificial intelligence accuracy would measure the accuracy and correctness of the real-time interpretation of the sign language as a percentage. Due to the literature gap found in the use of predictive artificial intelligence combined with sign language interpretation, there are no documented benchmarks that could be used to outline the expected outcome of this metric.

Discussion Although it is possible to create a detailed plan for how testing would take place, limitations are clear in terms of the hypothetical nature of this section. In order to ensure credibility and reliability of the tests in order to ensure proper system verification, tests would need to be performed and adjusted depending on the progress of the development of the tool. Ensuring that the system is verified using the outlined procedures, it is also possible to consider future extensions to the systems, such as personalized gesture additions where a user could add personalized gestures that they would like the tool to use in the future.

4.2 System Validation

This next section will focus on the system validation, also referred to as external evaluation. The validation of a system is one of the most important steps when ensuring that the developed tool will be used as intended and, in the case of this paper, focuses mostly on the usability and user feedback. It is yet again important to note that the nature of this section is also hypothetical, as tests cannot be performed in practice but will rather be detailed and explained in theory.

Validation Methodology In order to ensure validation of the system, the user experience is the most important point of discussion, which is why qualitative research methods are appropriate to incorporate for the sake of collecting valuable feedback and validating the system. For the basic testing of the application in a user centered setting, a number of participants would need to be recruited to participate in the studies. The sampling method most suited for the recruitment of participants would be purposive sampling and convenience sampling. In order to recruit participants with a specific trait, in this case to recruit an individual which is part of the DHH community and uses ASL to communicate, purposive sampling allows for precise recruitment of such individuals[17]. Since the study will include scenarios that aim to recreate the independent use of the system, non DHH participants also need to be recruited, which is what convenience sampling would be used for. Convenience sampling is one of the most widely used non-probability based sampling method which allows for low-cost and abundant participant recruitment due to its focus on availability and accessibility of the possible participant[18]. Using these two sampling techniques, a sufficient sample should be collected. Before the testing could begin, the variables that are aimed to be observed are various independent (IV), as well as dependent variables (DV). IVs, which are the variables that are changed and manipulated during the study are: the user characteristics (IV_1), the level of complexity (IV_2), and the system features (IV_3). The DVs, which are the variables measured as the IVs are manipulated are: the user experience (DV_1), effectiveness (DV_2), usability (DV_3), and perceived accuracy (DV_4). The goal is to validate the system and minimize the effect of possible confounding variables (CVs), which are variables that possibly introduce error and/ or bias in the results of the study.

Once the sample is collected, the participants would be invited to test the application and play out diverse scenarios. The IVs would then be manipulated within various sessions, for example IV$_1$ would be manipulated by having participants of different ages, skin colors, and genders play out mock-conferences with part of the non-DHH participants and repeat the process with IV$_2$ and IV$_3$.

After the completion of the study, all of the participants will be sent surveys assessing the dependent variables with the following layout (Table 2):

Table 2. Survey Questions

Statement	**1 - Strongly Disagree**	**2 - Disagree**	**3 - Neither Agree Nor Disagree**	**4 - Agree**	**5 - Strongly Agree**
1. The sign language interpretation tool provided a satisfying user experience.					
2. My experience with the tool was positive.					
3. I could easily understand the audio sign language interpretations. (Only non DHH)					
4. The tool encouraged effective communication during the call.					
5. It was easy to navigate the tool.					
6. The interface was friendly and intuitive to interact with.					
7. After usage, I am confident that my signs are interpreted correctly.					

The reason for choosing a 5-point Likert scale is due to prior research finding that the response rate and engagement was increased as opposed to 7-, or 10-point Likert scales[19,20,21]. The questions presented in Figure 1 are therefore the questions designed to address the DVs with a rating system, however, there are also important open ended questions that need to be asked to address ratings as well as more general feedback. In order to maximize the efficiency and the quality of the responses, semi-structured interviews would be conducted as the last step of the study, during which open ended questions regarding elaboration of ratings, as well as general feedback questions would be asked. The important closing question, to conclude the success of system validation for this tool would be ”What change would you welcome in order to use this application with every/ most web calls?”

4.3 Hypothesis Testing

As the internal as well as external validation has been discussed, it is now possible to create a hypothetical testing plan for the hypotheses. Earlier in this research paper, hypotheses have been touched upon but not specifically categorized and elaborated on, which is what this section is for.

Hypothesis Formulation This paper will focus on three hypotheses, which are accompanied by a null hypothesis each.

The main research hypothesis (H1) states "Image recognition and predictive artificial intelligence, when using convolutional neural networks (CNNs) for facial expression recognition and support vector machine (SVM) models for hand gesture recognition, are able to be used effectively to create a real-time sign language interpretation tool for web conferences.", while its accompanying null hypothesis (H0) is "Image recognition and predictive artificial intelligence, when using convolutional neural networks (CNNs) for facial expression recognition and support vector machine (SVM) models for hand gesture recognition, are unable to be used effectively to create a real-time sign language interpretation tool for web conferences.".

The second, predictive hypothesis (H2) states "Deaf and hard-of-hearing (DHH) users will experience a more authentic and engaging communication environment when interacting with this real-time sign language interpretation tool, compared to text-based communication.", while its accompanying null hypothesis (H0) is "Deaf and hard-of-hearing (DHH) users will not experience a more authentic and engaging communication environment when interacting with this real-time sign language interpretation tool, compared to text-based communication.".

The alternative research hypothesis (H3, focusing on the accessibility and the inclusivity states "The developed sign language interpretation tool will positively contribute to accessibility for deaf and hard of hearing (DHH) individuals in web conferences and as a result contribute to the inclusivity of the DHH community by bridging the communication gap." while its accompanying null hypothesis (H0) is "The developed sign language interpretation tool will not contribute to accessibility for deaf and hard of hearing (DHH) individuals in web conferences and as a result not contribute to the inclusivity of the DHH community by bridging the communication gap.".

Experimental Design Each hypothesis can be tested using either the values obtained from the system verification or through a statistical analysis from the system validation results. The approach that this paper follows is to reject the null hypotheses by using the hypothetical findings from the internal and external validations which would result in the acceptance of the hypotheses (H1, H2, H3). For H1, the null hypothesis would be rejected if the findings of the system verification metrics meet the expected results in terms of unit and automation testing. The null hypothesis accompanying H2 however could be disproven by applying statistical methods such as ANOVA (analysis of variance)[22] as well as regression analysis [23]. ANOVA would be beneficial in this case since there are multiple IVs that are being manipulated as well as multiple DVs to be measured. Regression analysis would then be more focused on the impact that the IVs have on the DVs and it is also especially useful when multiple variables are involved. The statistical analyses are crucial to rejecting the null hypotheses and accepting the hypotheses stated in this section.

Expected Findings Although the hypothesis testing outlines in this section is hypothetical in nature, if the system verification has the expected metrics as a result and the system validation is performed as outlined and incorporating the feedback throughout the process of the refinement of this tool, it would be possible to conclude that hypotheses H1, H2, as well as H3 are to be accepted whereas all the accompanying null hypotheses would be consequently rejected.

4.4 Conclusion

In conclusion, this section created a detailed plan for the system verification, validation, as well as hypothesis testing for the real-time sign language interpretation tool for web conferences. The main goal is to explore and find answers to the research question that this paper is based around, namely whether image recognition and predictive artificial intelligence can be used to create a real time sign language interpretation tool for web conferences. The experimental evaluation plan outlined in the entirety of this section follows current, state-of-the-art discoveries around artificial intelligence and how it can be used for the development of a sign language interpretation tool. In the future, actual testing should be conducted following the plan outlined in this section and potentially refine it during the process.

5 Discussion

In direct response to our research question of the extent to which image recognition and predictive artificial intelligence can be used to create a real-time sign language interpretation tool for web conferences, this paper has shown that the suggested tool has great potential to reach the goal outlined within the research question. By using Convolutional Neural Networks (CNNs) for facial expression recognition and Support Vector Machine (SVM) models for hand gesture recognition, a tool could be created that bridges the communication gap between individuals of the DHH community in web conferences.

After having created a detailed plan for the development of such a tool, discussions arise regarding the stated hypotheses, the main research question, the interpretation of anticipated results and their significance within the broader scope and context of this field. The results of Hypothesis 1 being accepted suggests that image recognition and predictive artificial intelligence techniques can be effectively used to create this tool, which would be verified using the detailed system verification plan, which would then also support the research question explored within this paper to a great extent.

Hypothesis 2 suggests that the developed tool would improve the DHH users' levels of engagement as well as the authenticity of their communication environment compared to that of more conventional text-based communication. To test this hypothesis, a system validation plan has been created in great detail for future implementation. Said validation plan is predicated on purposive and convenience sampling for the recruitment of participants, surveys, and interviews for communication and is designed to extract valuable feedback which will be necessary for the evaluation of this hypothesis.

And finally, Hypothesis 3 puts emphasis on the tool's potential to expand the accessibility, specifically for individuals within the DHH community, when it comes to online conferences, creating a more inclusive and overall positive experience. The integration of diverse user characteristics, the tool's features, and the complexity levels aim to facilitate a holistic solution.

By emphasizing the systematic verification, user-centric system validation, and statistical analyses a suitable degree of reliability and robustness can be ensured. This allows the methodology within this research paper to be structured in a way that minimizes the impact of unaccounted-for yet unavoidable variability that could affect the results.

The significance of the findings within this paper are profound and if realized, could revolutionize accessibility for individuals from the DHH community. Virtual conversations have the potential to be more engaging, authentic, and inclusive for everyone involved. The aim of expanding this tool to have cross-language adaptability also contributes to the focus on diversity and inclusion that the development of this tool was build upon.

However, the research conducted within this paper relies on certain assumptions, creating limitations as well. One assumption, for example, is that the user of this tool will have a machine (laptop or PC) that meets the advanced hardware and software configurations, which might not always be the case and therefore exclude some users from being able to use the tool. Moreover, seamless integration could only be ensured if providers of existing, popular, web conferencing platforms are willing to collaborate and incorporate such a tool within their platform. This possible limitation also applies to the assumption that it would be possible to use models that have been created by others, such as ACAT2.0 or Google's Text-To-Speech AI.

It is crucial to be aware of the aforementioned limitations and challenges likely to occur during the process for future work. The research presented in this paper remains hypothetical and the resulting absence of actual testing creates another clear limitation. The findings and outlined expected metrics are based on the state-of-art research and theoretical models which means that the specific use-cases could introduce confounding variables that have not been accounted for in the expected metrics outlined in this research paper specifically. The tool's innovative and novel approach to incorporating predictive artificial intelligence techniques in the context of sign language interpretation also means that there is no concrete literature published with metrics that this paper could have used to outline expected metrics.

6 Conclusion

In conclusion, the discussion of the potential of this real-time sign language interpretation tool emphasized the immense impact that this tool could have on individuals that are part of the DHH community within the scope of web conferences. Despite the hypothetical nature of the plan detailed in this research paper, the findings and contents within it could help the actual development and realization of such a tool in the future. As part of concluding this research paper, it is important to mention the suggestions, that accumulated during the process of the creation of this paper, for future studies.

6.1 Future Work

The focus of future studies aiming to deepen the knowledge within the field of research of artificial intelligence techniques and how they can be used for creating accessibility tools, specifically the tool outlined in this paper, should focus on creating tests for the system verification and studies for the system validation of the tool. Further, in-depth, research could also be conducted on the impact and effectiveness of predictive artificial intelligence on real-time sign language interpretation, realizing the hypothetical approach of this paper. In addition to the system validation methods outlined in its respective section within this paper, comparative studies could be conducted once this tool has been realized in order to compare its performance to existing accessibility tools in order to understand its real-world effectiveness and usability. Building upon this, longitudinal case studies may be conducted to delve deeper into the actual long-term impact such a tool would have on the life of an individual that is part of the DHH community within the context of web conferencing. It is certain that future research within this topic is necessary and if realized, could transform the current standards of accessibility completely.

References

1. C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1106–1113, Columbus, OH, USA, 2014.
2. A. Sahoo, G. Mishra, and K. Ravulakollu. Sign language recognition: State of the art. *ARPN Journal of Engineering and Applied Sciences*, 9:116–134, Feb. 2014.
3. P. Modi and S. Patel. A state-of-the-art survey on face recognition methods. *IJCVIP*, 12(1):1–19, 2022.
4. A. Wadhawan and P. Kumar. Sign language recognition systems: A decade systematic literature review. *Arch Computat Methods Eng*, 28:785–813, 2021.

5. C. de L. Pataca, M. Watkins, R. Peiris, S. Lee, and M. Huenerfauth. Visualization of speech prosody and emotion in captions: Accessibility for deaf and hard-of-hearing users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, pages 831–845, Hamburg, Germany, 2023.
6. G. Bastas, M. Kaliakatsos-Papakostas, G. Paraskevopoulos, P. Kaplanoglou, K. Christantonis, C. Tsioustas, D. Mastrogiannopoulos, D. Panga, E. Fotinea, A. Katsamanis, V. Katsouros, K. Diamantaras, and P. Maragos. Towards a dhh accessible theater: Real-time synchronization of subtitles and sign language videos with asr and nlp solutions. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22)*, pages 653–661, Corfu, Greece, 2022.
7. E. K. Kumar, P. V. V. Kishore, M. T. K. Kumar, and D. A. Kumar. 3d sign language recognition with joint distance and angular coded color topographical descriptor on a 2–stream cnn. *Neurocomputing*, 372:40–54, 2020.
8. S. Nadgeri and A. Kumar. A deep learning-based approach to classification of baby sign language images. *IJCVIP*, 12(1):1–18, 2022.
9. R. Rastgoo, K. Kiani, and S. Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.
10. M. B. Munir, F. R. Alam, S. Ishrak, S. Hussain, M. Shalahuddin, and M. N. Islam. A machine learning based sign language interpretation system for communication with deaf-mute people. In *Proceedings of the XXI International Conference on Human-Computer Interaction (Interacción '21)*, pages 4–12, Málaga, Spain, 2021.
11. Q. Xu and N. Zhao. A facial expression recognition algorithm based on cnn and lbp feature. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2020.
12. K. Quazi. Acat 2.0: An ai transformer-based approach to predictive speech generation. Computer science and engineering senior theses, 2023.
13. E. Li and Y. Xu. Face detection based on improved color space of ycbcr. *IOP Conference Series: Materials Science and Engineering*, 439(3):032075, Nov. 2018.
14. L. W. Sze Ee, C. R. Ramachandiran, and R. Logeswaran. Real-time sign language learning system. *Journal of Physics: Conference Series**, 1712(1):012011, 2020.
15. J. van H. Breurkes, F. Gilson, and M. Galster. Overlap between automated unit and acceptance testing – a systematic literature review. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering (EASE '22)*, page 80–89, Gothenburg, Sweden, 2022.
16. N. Islam. A comparative study of automated software testing tools. Master's thesis, St. Cloud State University, 2016.
17. L. A. Palinkas et al. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm Policy Ment Health*, 42(5):533–544, Sep. 2015.
18. M. Elfil and A. Negida. Sampling methods in clinical research: An educational review. *Emerg (Tehran)*, 5(1):e52, Jan. 14, 2017.
19. E. Babakus and G. Mangold. Adapting the servqual scale to hospital services: An empirical investigation. *Health Service Research*, 26:767–780, 1992.
20. S. Sachdev and H. Verma. Relative importance of service quality dimensions: A multisectoral study. *Journal of Service*, 1998.
21. J. G. Dawes. Do data characteristics change according to the number of scale points used? an experiment using 5 point, 7 point and 10 point scales. *International Journal of Market Research*, 51(1), 2008.
22. T. K. Kim. Understanding one-way anova using conceptual figures. *Korean J Anesthesiol*, 70(1):22–26, Feb. 2017.

23. A. Schneider, G. Hommel, and M. Blettner. Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int*, 107(44):776–782, Nov. 2010.
24. P. A. Rodríguez-Correa, A. Valencia-Arias, O. N. Patiño-Toro, Y. Oblitas Díaz, and R. Teodori De la Puente. Benefits and development of assistive technologies for deaf people's communication: A systematic review. *Frontiers in Education*, 8, 2023.
25. World Health Organization. Deafness and hearing loss. Accessed: 14-Sept.-2023.
26. K. Chengeta and S. Viriri. Facial expression recognition using local directional pattern variants and deep learning. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence (ACAI '18)*, pages 31–37, Sanya, China, 2018.
27. A. A. Barbhuiya, R. K. Karsh, and R. Jain. Cnn based feature extraction and classification for sign language. *Multimed Tools Appl*, 80:3051–3069, 2021.