DESIGNING AND EVALUATING A SYSTEM FOR THE EFFECTIVE ANALYSIS OF SIGN LANGUAGE VIDEO CONTENT FOR THE IMPROVEMENT OF VIDEO QUALITY

By

Joseph Moscatiello

B.Sc. in Computer Science, Ryerson University, Toronto, Ontario, Canada 2014

A thesis presented to

Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Science

in the Program of

Computer Science

Toronto, Ontario, Canada, 2014

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

DESIGNING AND EVALUATING A SYSTEM FOR THE EFFECTIVE ANALYSIS OF SIGNED LANGUAGE VIDEO CONTENT FOR THE IMPROVEMENT OF VIDEO QUALITY

Joseph Moscatiello

MSc, Computer Science, Ryerson University, 2014

Abstract

Signed Language communicators use video communication services as they can be used to relay manual communication. One aspect of successful video blogging (vlogging) is being able to communicate a message clearly. Visual clarity is important as manual communication relies on the visual channel exclusively for processing and can become compromised if certain elements in the video, such as the lighting or background, are not set up correctly. A tool, termed the Vlog Analysis and Suggestion Tool (VAST) has been developed to assess Signed Language, talking head style, video and provide feedback to users based on the quality. Quality, in this work, is based on three technical factors: (1) lighting; (2) signing space; and (3) background. Results from a user study on VAST indicate that the tool is easy to use, helpful to users for determining video quality, and the technical factors assessed by the system are important to users.

Acknowledgements

I would like to thank the people without whom I would not have been able to complete this thesis over the last two years at Ryerson University. Note: No students were harmed in the making of this document.

First and foremost, I would like to express my sincere gratitude and appreciation to my thesis supervisor Deborah Fels for her continued support and guidance. Without her none of this could be possible. Thank you Deb!

In addition, I would like to thank my mother and father and my sister and brother for their love, support and encouragement during this lengthy process. As well, I would like to thank my grand parents and my aunt and uncle for providing me with a place to stay, food on the table and for not kicking me out the second I told them I would be living with them for two more years. I would also like to thank my girlfriend Ria Pecora. She has been there through thick and thin and has always been a source of encouragement, IFLY.

I want to also express my sincere gratitude to my close friends: Jonathan Turco, Michael Pouris, Kristian Ott, Chris Tranquada and Paul McLoughlin and whoever else I missed who was there for me. These poor souls sometimes listened to my problems for hours on end. Thank you so much for being there for me.

I could not have completed this document without help from the amazing members at the IMDC lab. Most importantly, Paul Church, Margot Whitfield, Rob Bajko, Leshanne Mori, the lunch time soccer crew and the rest of the labbies. Without them my trips to the gym, coffee breaks, lab outings and simply every day that I spent working at the lab would not have been the same.

Finally, I would like to thank my participants for agreeing to be my guinea pigs in this whole process. I could not have done it without them!

This thesis is dedicated to my Grandfather, Guiseppe Veneziano.

Table of Contents

List of Tables

1 Introduction

There have been a number of innovations in technology, social media and video communications platforms that have created opportunities for individuals to stay connected and share content with one another. Video sharing and social media are now a part of our everyday life. Social media websites, such as Facebook and Youtube, have over 1 billion average users each month and over 100 hours of video content is uploaded each minute to YouTube alone [1]. The benefits of creating and sharing multi-media content are many. The cost of owning a personal computer, tablet or network enabled phone is minimal, interfaces have become easy to use and network access has become cost effective and accessible from virtually anywhere in the world.

Deaf communities have been given the opportunity to take advantage of the visual forms of communication readily available to them and shed the limitations of oral based technology, such as the telephone, which limited them in the past [2]. Video communication has the ability to allow the producer and consumer to feel as if they are interacting with one another in a face-to-face interaction. As well, visual forms of communication allow Signed Language users to interact with the web and with each other using their native language.

Video blogging (vlogging) has become a very popular practice for Deaf individuals who post Signed Language content in the form of video blogs to the internet (vlogs). Vlogs are typically authored by one single person using a talking head style production technique (more information provided in Section 2.5.1) in which the signer positions him/herself in front of a recording device and signs to an invisible audience [3]. Herring [4] found that vlogs are typically created for three main purposes: personal journals, knowledge logs (instructional videos) and as filters (comments on other videos or current events). Signed Language interpreter students also use vlogs for educational purposes to learn and practice a Signed Language.

Signed languages are complex visual structured languages that combine hand gestures, facial expressions, body movements and 3-d space to facilitate, proper, grammatically correct communication [5]. In order to successfully facilitate intelligible Signed Language video communication the quality of the message must be visually clear. Visual clarity can be affected by technical elements, such as lighting and background, along with other more linguistically oriented elements such as grammar and semantic coherency [6]. Viewers may experience comprehension issues if a vlog is not visually clear because of, for example, poor lighting. Guidelines and standards have yet to be created for the proper production of Signed Language video content based on quality. Items such as proper lighting, colour selection for clothing and background and minimum space requirements when recording may help to create a unified standard for vlog creation.

This thesis focuses on the design of the Vlog Analysis and Suggestion Tool (VAST), a simple stand-alone software that assesses and reports the quality of Signed Language video content, in particular talking head style videos. The software also provides recommendations to users for improving their video quality. Common computer vision algorithms, such as an edge detection technique, colour-to-grayscale conversions, a face detection algorithm and frame assessment techniques, are used to evaluate the quality. Quality in this work is based on three technical factors: (1) lighting, (2) background and (3) signing space. Lighting is a combination of luminance and illuminance and can create an aesthetic effect on an area/environment and can help to improve the visual performance of visual tasks. Lighting is important for Signed Language production as under/overexposure may lead to comprehension issues for the viewer if they cannot successfully view the signer or articulate intricacies with the signer's communication [7]. Overexposure (too much lighting) occurs when a loss of detail is experienced due to the high amount of lighting in the image. Detail is effectively "whited out" from the image. Underexposure is caused by low lighting conditions and can reduce the amount of detail visible in the image. Signing space is the amount of 3-d space in front of the signer used to communicate [8]. It affects quality in two different ways. If someone does not record with enough singing space the gestures they create may appear in front of their face, which would make it difficult for viewers to see the facial expressions of the signer. Secondly, if signs are gestured off screen, the viewer will not be able to see what the signer is attempting to communicate and the message may be lost. Background is referred to the background and foreground behind and in front of the signer as they communicate. Background is important as visually noisy elements in the vlog may reduce the visual saliency (reduce focus) of the signer and cause confusion for the viewer if more information is presented on screen [9].

VAST is the first system of its kind to assess Signed Language video based on the combination of lighting, background and signing space and output feedback without attempting to correct quality issues. The literature on enhancing the quality of Signed Language video is growing but currently limited. Most research in the field of Signed Language focuses narrowly on gesture recognition and does not consider other important components of the language.

As part of this research, I wanted to determine if users would find the feedback and recommendations from the system helpful in determining quality and if they found the system useful overall. I also wanted to investigate the impact the factors of lighting, signing space and background had on a user's perception of video quality. With more consumer video technologies and applications being released, with possibilities for higher video resolution and editing tools, would participants still be interested in using a simple tool to assess their created content? VAST is software that can be used to start exploring and identifying vulnerabilities in Signed Language video quality and make the user aware of possible comprehension issues.

To answer the proposed questions, a user study was conducted. The study consisted of five different trials. In each trial the participant was asked to sign the response to a question we asked, while being recorded. Before each vlog was recorded, a technical setting(s) was changed in the room or with the recording device that created various technical scenarios (e.g. different backgrounds, lighting conditions and viewing area of the camera). The technical scenarios were randomized each time to minimize the learning effects. Once their response was recorded we asked the participant to open it up with the VAST system and analyze the vlog for quality. The participant was then asked to fill out a questionnaire based on the feedback given by VAST. A pre-study and post-study questionnaire was also used before and after the study to gather demographic data and overall impressions of the system. The results from the study indicated that the users found feedback on background to be most helpful when determining quality. Participants also stated that signing space and lighting were among the top technical factors participants consider important when determining the quality of a video. In terms of the usability of the system, no participants reported having trouble uploading or assessing their vlog with VAST and participants said that the low processing times added to the experience of the system. The only improvements suggested by participants were to include video playback so that they could see where the quality issues were in their video and to change the colours of the feedback icons to better represent the feedback suggestion.

1.1 Research Questions

The research questions I wish to answer with this research are the following:

1. How can the detection of the technical factors of signing space, lighting and visual clutter be used for assessing video quality of vlogs?

2. What is the usability of the VAST system?

3. What is the level of impact of each technical factor on user perception of video quality?

4. How helpful is the feedback of lighting, signing space and background clutter?

1.2 Contributions of this Thesis

Research has already been conducted to enhance the quality of Signed Language video content automatically. However, much of this past research has focused on quality issues caused by the limitations on sharing, streaming and distribution of large files. Little research has been conducted surrounding user chosen technical quality factors, such as signing space. Furthermore, few systems have been created which inform users of possible errors rather than attempt to correct the issues. I submit that research in the area of quality recommendation systems using the following three technical quality factors is limited: (1) lighting; (2) sign space; (3) visual clutter. The contributions made by this thesis are detailed below.

1. The main contribution of this research is the Vlog Analysis and Suggestion Tool (VAST). The algorithm, which uses known computer vision techniques, analyzes each frame of a user input video and generates feedback on quality based on three technical factors: (1) lighting; (2) signing space; and (3) background. The algorithm identifies different features in the frame, including face and body position, and uses this information to generate scores for the lighting, signing space and background. Once all frames have been analyzed the algorithm uses the generated scores to create feedback and later output this feedback to the user. Feedback is output separately based on the technical factor to provide organization and is represented in textual and visual form. The user study revealed that VAST was usable where it was designed to have a smooth learning curve, allow participants to perform tasks quickly and easily and overcome errors if any were experienced. It also met its primary goal of providing users with helpful information when determining the quality of lighting, background and signing space in their own created vlogs.

2. The second contribution of this thesis is the user study. Few studies have been performed to understand the impacts certain factors have on viewing patterns of participants while viewing Signed Language content. Much research pertaining to Signed Languages focus on gesture and facial recognition and effects of frame and bit rate and do not consider user chosen factors, such as background. This research compares and combines the technical factors of lighting, signing space and background/foreground clutter. The user study found that the chosen factors of lighting, background and signing space had an impact on the viewing patterns of participants. The results showed that participants would be the least likely to watch another vlog if the signer recorded it with a small signing space. Furthermore, participants found the feedback on background to be the most helpful. The methods used in

this study can thus be applied to further studies to identify the impacts of other technical factors and to different populations.

1.3 Thesis Outline

This thesis is broken down into the following 5 chapters:

Chapter 1: Serves as an introduction to the thesis document. This section details the background and motivation to this research, as well as the contributions made.

Chapter 2: Provides an in depth literature review of similar work surrounding this research topic.

Chapter 3: This chapter described the technical aspects of the designed system and the methodology used to evaluate the proposed system.

Chapter 4: Describes the tests used to analyze the results from the user study. It also discusses the results and describes possible reasoning's for the results.

Chapter 5: This chapter presents the conclusion and highlights the major findings in this thesis. It also discusses the limitations faced during the undertaking of this research and suggests some possible future work.

2 Literature Review

This chapter provides a review of concepts related to the VAST system, as well as the motivation and previous research related to this field of work. The creation and construction of the VAST system is performed using concepts and techniques found in multiple relevant fields of Computer Science. There are three major areas of review in this section: (1) the background of ASL and technology use for the Deaf and Hard of Hearing; (2) the technical factors used by the VAST system for vlog analysis; and (3) a review of current automated facial recognition techniques.

It is important to understand the complexity behind the use of Signed Languages and the impact technology has had on the Deaf community. The section that encompasses the background and review information of ASL provides an overview of ASL and a brief history of vlogging and the technologies involved.

The motivation behind the use of the technical factors of lighting, visual clutter and background for the analysis of vlogs is presented in Sections 2.3, 2.4 and 2.5 respectively. These sections include a discussion of the similarities between vlog and television/film production, as well as an insight into the techniques used in the literature to identify and categorize each technical factor.

Lastly, the section on facial recognition techniques provides a brief overview of the current methods available to detect, identify and track faces and objects in images. This section includes a discussion and review of the Viola-Jones facial recognition algorithm, which is used in the VAST system.

2.1 American Sign Language (ASL) and Signed Languages

ASL is a visual-spatial language that originated in the early 19th century through a combination of earlier Signed Languages and is now being used by Deaf communities worldwide [10]. This language is predominantly used in North America by Deaf individuals [11] and has only been accepted as a world language for just over 50 years [6]. Signed languages, such as ASL, have developed over time, as the Deaf communities cannot sufficiently rely on spoken languages for communication. These languages have developed outside of general language processing norms of written and vocalized speech [12]. Signed languages are not directly derived from any other language; however, because Deaf signers learn and grow up in hearing communities they are required to create signs for spoken words [13].

Signed language is communicated through a complex visual structure based on synchronized movements [5]. Signers must coordinate and integrate movements within 2-D/3-D space (depending on communication means), the facial region and spatial memory [14]. A signer is able to use the fine placement of his/her fingers, palm orientation, hand shape and head, hand, arm and body positioning to produce and distinguish signs [11]. The speed of signing (signing speed) and the amount of space used to communicate (signing space) also contribute to the grammar of the language. English words and phrases that do not have a sign equivalent are conveyed using finger spelling. Finger spelling involves the signing of each letter of a word [5]. Items such as proper nouns, including names of locations and people, would be finger spelt.

Facial expressions are necessary for ASL communication as they convey emotion and syntactic structure [15]. The importance of information extraction from facial features and expressions has been documented as far back as the 17th century [16]. Facial expressions such as eye movements, nose wrinkles, eye brow placement and mouth/lip shape all contribute to emotional and grammatical distinction. Keating and Mirus [5] and Michael [17] discuss the processes involved in asking a question. When doing so, a signer's eyebrows will raise, eyes will widen in anticipation, their head/body will lean forward and their hands will motion to depict what type of question it is. A viewer who is only able to see the hand motions of the signer will not be able to determine that a question is in fact being asked [17]. Other than the five informational questions (i.e., who, what, where, when, why), facial expressions can help to deliver topic information, negations and conditionals [17].

Two alternate techniques for communication, from Deaf individuals, include speech reproduction and lip-reading. Speech reproduction is not an uncommon phenomenon, even though profoundly Deaf individuals cannot hear vocalized speech [5]. Lip-reading is a process of communication by which reading the lip movements of the other person is the primary means of understanding conversation. Conrad [13] argues that lip-reading is a required skill and is essential to every-day living because of the bilingual community in which Deaf people are immersed.

2.1.1 ASL Assessment

ASL communication is becoming a popular study in North America [6]. Hundreds of colleges now offer linguistics courses in ASL ranging from beginner to expert levels [6]. It is suggested by Kemp [18] that the increasing demand for ASL study stems from the fact that a learned second language can benefit the user in career advancement and in social activities. There is also a common misconception about the ease of learning the ASL language [18].

Hooper et al. [6] state that the increased demand for ASL has sparked the need for automatic and reliable assessment tools to rate and assess student and amateur performance. The paper written by Hooper [6] describes the process of current educational assessment methods. An instructor must manually watch and assess each student video individually and provide assessment for each segment of the video. Performing assessment for students in a large class often takes several days to complete. Many papers found in the literature focus on gesture recognition, as oppose to individual feature detection, and are often restricted by the limitations of computer vision.

2.2 Deaf Video Communication, Blogging and Vlogging

In the past, the Deaf community has been restricted from the use of oral-based technology, such as the telephone, to communicate over long distances [2]. Newer technologies, including text SMS, email and new web 2.0 technologies, such as real-time video streaming services, have created new opportunities for communication, in general, and specifically for Deaf people. These advancements have helped to bridge the communication gaps between hearing and Deaf persons and have allowed Deaf individuals to become more independent and social [19]. New online Deaf communities have also stemmed from developed technologies. Wilson and Peterson [20] suggest that these new communities encompass families, groups with similar interests, different ethnicities, workplace personnel and may even be the cause of the reduction of physical social events in the Deaf community. Blogs have also played a large role in organizing groups of people into communities online.

Herring [4] defines a blog as *"frequently modified web pages in which dated entries are listed in reverse chronological sequence"* (page 1). Blogging first began around 1997 [4] as text-only communication and became popular as its use did not require any prior web or HTML knowledge. Text-only communication, however, was still not an ideal communication avenue for the Deaf community as the average Deaf individual has a reading capacity that is comparable to a fourth grade level [21]. With the advent of web 2.0 technologies, other forms of media such as graphics, sounds and video, soon began to be included in standard text blogs for, as suggested by Hoem [22], visual expression.

Vlogging is a form of blogging where video, rather than text, is the medium. The creation of vlogs requires no more than a web connection and simple video production technology, found in most PC, smartphone and tablet setups, to use. In a vlog, a user will position himself/herself in front of a camera and microphone and talk or sign to an invisible audience. Vlogs are typically authored by a single individual and are short, roughly, 3-5 minutes clips [23]. Similar to textual blogging, vlogs are used for the purposes of personal expression, public opinion and social commentary [24].

The benefits of vlogging and video technology for the Deaf community are many. Vlogging allows for the preservation and sharing of thoughts and expressions using a Signed Language without the need for textual interaction [25]. Face-to-face interactions are being mimicked online and Signed

Language communication is no longer forced to take place in a close physical proximity [5]. Hibbard and Fels [26] state that the benefits of video communication are comparable to the benefits of the printing press for a hearing audience.

The ubiquitous use of vlog creation for personal expression has opened up many new unexplored areas in Deaf communication. Hibbard and Fels [26] studied the differences of in-person versus online communication through the analysis of ASL vlogs to view the subtle changes signers make when creating online content. Data was collected and analysed for quantitative and qualitative factors to view the differences of items such as topic, sign rate, average length and video quality between vlogs on DeafVideo.tv and Youtube, and between face-to-face interaction and online communication. The results of this study suggest that Signed Language oriented websites (DeafVideo.tv) provide a more comfortable experience for Deaf vloggers. Further evidence to conclude this hypothesis stems from the analysis of signing rate (rate in which sign is produced). Signing rates in videos created for DeafVideo.tv correspond to signing rates of face-to-face communication. In comparison, signing rates in videos originating on YouTube are higher and appear to be in a rushed manner [26].

Since the recording of a vlog is usually informal and can be compared to a personal diary entry [27], the appeal of wearing formal clothing, having a monochromatic background or perfecting all signs in digital space, for example, is not often a part of vlogs. A video, unlike text or photo, allows its producer to feel as though they are interacting directly with the consumer as if in direct face-to-face interaction [24]. Hibbard and Fels [26] discovered that over half of the analyzed vlogs were created at home, in a casual manner. For example, videos were shot in bedroom and living room areas of homes, and strived to create an intimate experience between the producer and consumer of the video content. Constructed vlogging scenarios, such as those in the bedroom however, can be sub-optimal (messy or dark, for instance) making the vlog content/signs difficult to see.

2.2.1 **Video Elements**

Deaf online video communication requires optimal visual settings and legible visual elements for the sharing and viewing of visual content [28]. If a vlog contains distracting elements or is hard to see the viewer may have trouble understanding what the signer is trying to communicate. Fine details in body movements, such as finger spelling and facial expressions, may be increasingly difficult to understand when, for example, there is insufficient lighting.

Recommendations for the production of Signed Language content, for human-controllable factors, have been created to help interpreter students with making comprehensible online video content [25, 26]. These recommendations are enforced in various interpreter-training programs for the purposes of teaching online interpretation [29] but are not officially standardized by any institution or official guidelines. Such recommendations may be helpful for amateur vloggers as well who wish to consider quality when creating Signed Language video; however, these recommendations are simply not accessible to amateur vloggers as they are mainly published in academic papers and through Universities.

The term human-controllable is used in this work to describe a technical setting, such as lighting or signing space, that the user has active control over when creating a vlog. It is important for vloggers to be aware of how these human-controllable elements affect their audience's understanding of their signing and make adjustments if those elements made it difficult for the audience to see and understand what they were signing. VAST assesses a Signed language vlog for three human controllable factors: (1) lighting, (2) background and (3) signing space. I chose these human-controllable factors among others as they are outlined in many works in the literature as being influential to the technical quality of Signed Language video including in works written by Pfau [30], Fels [25], Keating and Mirus [5], Gallaudet University [29, 31]. Lighting causes loss of concentration, eye strain and physical exhaustion [31]. Background can have an effect on communication and visual distraction from objects on screen should be avoided [30]. The full horizontal direction of the signer must be captured [30] and signing space must be fully seen by both communicators (the viewer, if video) to maintain communication [31].

2.2.2 Signed language video compression

Video compression is a technique used to reduce the file size of a video to ensure optimal quality transmission when users access video content [32]. There exists a need to compress large video files for distribution, search and scalability purposes [6]. Limited bandwidth (rate of transfer of bits per unit time (bitrate)), coupled with a low rate of transfer, prevent the seamless transmission of larger, high-resolution video content to some devices [33]. Results of the transmission of un-compressed video include frame-rate issues, video delay and loss of information.

Much work in the area of improving current compression schemes, specifically for Signed Language video, has been done to improve the transmission of Signed Language video content to reduce comprehension issues. One of the main focuses in compression algorithms for Signed Language video is to understand where and on what Deaf people look at during an ASL temporal sequence. Eye tracking gaze experiments were carried out by Muir and Richardson [32, 34] and Agrafiotis et al. [35]. The consensus result among these studies indicates that consistent central focus by the Deaf viewer was maintained in the facial region of the signer during communication. Finger spelling and hand and body movements were viewed with lower resolution peripheral vision [34]. Eye movements away from the face were noted in times when the hands of the signer gestured in front of his/her face and when the signer motioned exaggerated, large, gestures [32]. Viewers also gazed away from the signer's face when they became confused or needed to verify an action Figure 1 displays the set of (x,y) gaze co-ordinates gathered from three separate participants.

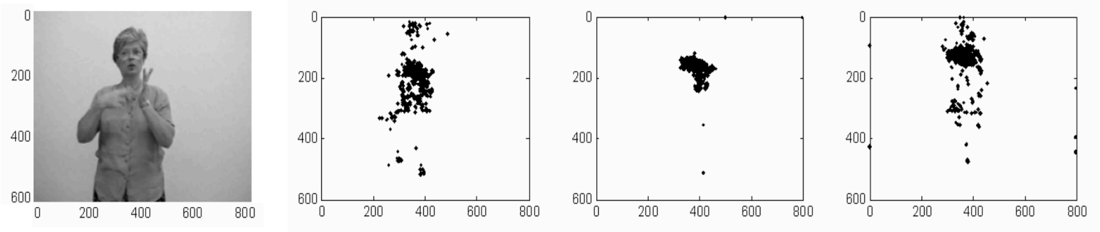


Figure 1: set of (x,y) gaze points of three participants [34]

Muir and Richardson [28] note that the process of visually enhancing all areas in a video frame would be unnecessary for Signed Language video because of the results from their study. As seen with the gaze data in Figure 1, only certain areas of a video frame demand high resolution coding. Muir [32] proposes a technique that allows the frame to have variable image quality, maintain quality in the facial region of the signer and reduce quality everywhere else to reduce file size. The low image quality sections of the frame allow for a reduced bit-rate transfer without the reduction in frame-rate.

Video size is another factor that is considered during the compression process. Previous research in multimedia, which investigated video size for comprehension, did not consider ASL content; however it did find that video size played an important role on comprehension. Hooper et al. [6] examined the effects of video resolution on ASL comprehension. Fifty-one student participants watched multiple 30-second ASL clips and were asked to retell the story of the clip they were shown. The videos were displayed in three different pixel resolutions: 480x360, 320x240, and 240x180. Based on the fluency scores (a score created based on the participant's ability to re-tell the story smoothly, without repetition, and in a way that the researchers have no trouble understanding) generated by the researchers, results indicate that video size does not significantly affect comprehension levels of ASL viewers.

While this thesis does not focus on video transmission, it is important to summarize research on ASL video and view what studies have been performed to enhance the intelligibility of Signed Language video content. Studies, such as the maintenance of central focus during the viewing of vlog content and the effects of video resolution on comprehension, are useful to this thesis work. For example, knowing that Deaf users maintain central focus on the signer's facial region while communicating has influenced the development to place high regard on the facial region when analyzing the video.

2.3 Lighting

Lighting is a term used to describe the general effects of artificial or natural illumination used for the purposes of creating various aesthetic effects and allowing visual tasks to be performed. Lighting has been found to cause emotions such as calmness, hostility and comfort [36-38]. In addition, lighting has been shown to affect physiological and physical functions such as blood pressure, heart rate and body temperature [36] and is also being used to treat various ailments including depression and sleep disorder [39]. Furthermore, poor lighting has been known to cause eye fatigue and loss of concentration [36].

It is first important to understand the concepts related to illumination so that it can be applied to computer vision concepts. Illuminance, not to be confused with luminance, is the amount of incident light that reaches the surface of an object [3]. The illumination intensities on an object will decrease by a factor of $1/d^2$, where d is the distance in feet, as the object moves further away from the light source. This property is known as the inverse square law and holds true if the light source emits the same amount of light in all directions [3]. The illumination angle, intensities and position towards an object will affect the luminance, cast shadows and other viewing factors on that object.

Luminance is a quantitative unit of measure from which brightness is perceived [3]. It can be considered as either light reflected from a surface or light directly being emitted from a source. Reflective luminance is based on the distribution of incident light and the reflective properties of the surface the light interacts with [40]. Since the human eye cannot consistently give an accurate measure for the amount of light in an environment [3], a subjective definition for light is also needed. The reason luminance is important to computer vision is because it determines how bright a human perceives an object in real life and in photographic imagery. The intensity of a grayscale pixel in an image is directly related to the luminance in the real world.

2.3.1 Studies on Performance and Perception of light

A large portion of the early research on lighting effects on humans relates to visual performance. Visual performance is measured by the ability of a participant to complete a specified task correctly under different illumination settings. The illumination conditions in these studies are modified several times over the course of the study to investigate the effects of this change on performance. In one particular performance study for example, Boyce [41] found that as illuminance increases as does the ability of a participant to identify patterns more quickly and accurately on paper using shapes. Boyce [41] also found that a reduction in task difficulty and an increase in participant satisfaction were linked to an increase in illuminance. This lighting study however, did not discuss the results of increasing the illuminance beyond 2000 lux to investigate if higher illuminance levels had the same positive or negative effects on task performance. Since the early work on visual performance, further studies have been applied to retail, education, real estate, computer vision, etc. [36].

The process of exploring human facial recognition performance in images under various illumination conditions was performed by Braje et al. [42]. This study found that illumination levels and directions of light on the face affect the performance of human facial recognition. Participants in this study were asked to identify and match photos of human faces. The training phase allowed the participants to learn the identity of each person in the photo. Once the participants were confident they could remember each identity, a random set of the same photos with different illumination conditions on the faces were shown to the participants for identification. The authors found that poor lighting conditions, cast shadows and different directions of illumination on the images impaired facial recognition and participants took more time accurately identifying the persons in the images.

In a similar study, Erber [38] found that the mean performance of lip reading in Deaf children decreased when frontal illumination was decremented by significant amounts. Moderate changes in the frontal illumination of the speaker decreased intelligibility by 13% and background lighting, which also caused illumination changes in the face, reduced the visual intelligibility by 41%. This study also noted that shadows presented in the lower quadrant of the face produced by viewing angles reduced the performance of lip-reading by more than 21%.

VAST analyzes the lighting conditions using the facial region of the signer in a vlog. The research by Braje et al. [42] and Erber [38] help to stake my claim that the assessment of lighting in the facial region is especially important for signed communication and the identification of low lighting conditions and cast shadows on different parts of the face will affect the comprehension of the user and visual intelligibility. VAST is able to identify cast shadows when they are present in the shot, however shadows were not included in the VAST study as a quality scenario for time purposes.

Many researchers disagree with the performance based approach to lighting research, as they argue that lighting plays a larger role on the impact of human psychology than just on performance alone [43]. Flynn et al. [44] studied the preferred lighting levels of participants and used a three dimensional situational metric to record the participants perceptual judgement in a fixed environment. The three dimensions he used were: (1) bright/dim; (2) uniform/non-uniform; and (3) peripheral/overhead. The situational lighting metrics used by Flynn [44] have been mimicked multiple times in later studies. The findings indicate that not only does light cause perceptual judgements of an area, but it also affects seating selection, posture, facial expressions, viewing patterns and human circulation patterns.

In correspondence with Flynn's research [44], other research has been done to validate and continue progress on preference based lighting research. Butler and Biner [43] investigated the preferences for lighting of participants in different active situations such as eating dinner, interacting with others and reading. This body of work was different from other work on preference-based research, as the researchers simply asked the participants to rate different situational lighting preferences without physically placing the participants in that environment. This study found that interaction with others and reading/writing demanded the highest lighting levels. Two important activities considered when creating vlogs.

There is no consensus among recent literature whether or not conclusions drawn for preferred lighting levels are accurate. Reasons for this are discussed by McCloughan [37] and Veitch [45] and are in part due to the fact that various independent variables exist, such as gender and age, in each study that contribute to varying participant preferences. Knez [46] found that the mood and cognition conclusions drawn by previous research need to be revised based on gender dissimilarities found within the research. A similar body of work performed by and Newsham [45] suggests that race may also play an important part in determining the preferred lighting levels of participants.

Although the studies on performance and perception/preference vary in objective, both conclude that increased illumination levels are important for human performance and perceptual clarity. As ASL is a visual language and many forms of visual reading take part in Deaf communication, the results form the above studies directly relate to the importance of lighting in Signed Language communication [14].

2.3.2 **Lighting in Television production**

The studies on performance and perception of lighting levels have also influenced many domains including techniques used in film and television production. Strategic lighting plays a large role in portraying a scene/scenery in film and can create an important aesthetic effect on an area or object. Lighting is used by many professions including engineers, architects and film producers to contribute to the quality of a design. It is important that the lighting in a room or scene is optimal as poor illumination often results in degradation of quality and user experience [7]. Zettl [3] notes that lighting is important in film production for the following reasons:

> *To help the television camera see well, that is, produce technically optimal pictures*

> *To help the viewer see well – to recognize what things and people look like and where they are in relation to one another and to their immediate environment.*
>
> *To establish for the viewer a specific mood that helps intensify the feeling about the event.* [3] (Page 212).

Image quality is reliant on the illumination condition in an environment [7] and sufficient lighting will significantly affect the ability to identify objects, facial imagery and other important parts of a scene [17]. In relation to Signed Language video, Cuizhu et al. [47] state that a result of inadequate lighting is often under and over exposure, which causes comprehension issues. Best practises for video production include the use of a uniformly distributed light source, avoiding lighting angles that create shadows and avoiding external illumination from behind the signer [38].

2.3.3 **Measurement of Lighting in images**

In regards to photographic photometry, there are many ways to calculate the luminance values of an image, with each technique having advantages and disadvantages [48]. Determination of pixel intensities is significant to the success of multiple computer vision algorithms as lighting levels directly affect the ability of a system to perform specific tasks, such as extract and identify objects from a scene [50-52]. Since there is no standard formula to calculate photographic luminance, the literature presents several ways to determine these values using various colour models. Techniques include colour histogram estimation and equalization [49, 50], measurement through photographic film [48], high dynamic range photography [51] and colour-to-grayscale conversion [7].

The colour-to-grayscale technique converts 3-dimensional (RGB) data into a single grayscale dimension in the range of 0 – 255. Each pixel in a grayscale image represents the light intensity of the corresponding RGB pixel. The most widely used colour space for digital image processing is the sRGB colour space, which originated with the cathode ray tube (CRT) monitor [52] and is mapped based on human colour vision [53]. The goal of a colour-to-grayscale conversion equation is to retain as much information as possible about the original image. Bezryadin et al. [54] detail and compare several formulas used for grayscale conversion; Table 1 displays three of the formulas the authors review in this paper where R, G and B represent the colours red, green and blue, respectively.

Table 1: Colour to Grayscale Formula

Technique Name	Pixel Intensity Calculation
Arithmetic mean	V = (R + G + B) / 3 [54]
Combination of RGB (used in VAST)	Y = $0.299R + 0.587G + 0.114B$ [55]
Max RGB	V = max (R, G, B) [54]

Table 1 outlines three commonly used colour-to-grayscale techniques to determine pixel intensities. The arithmetic mean equation computes an average of the sum of the R, G and B values. The max RGB equation simply takes the highest value between the RGB values. For example, if a pixel's RGB values were: R - 35, G - 189 and B - 22, the equivalent grayscale pixel value would be 189. Finally, the combination of RGB equation computes a weighted sum of the RGB values and is based on human colour vision and the sensitivity of the eye to different colours [53]. This equation is said to give a perceived brightness of the image pixel and is a suggested algorithm by the World Wide Web Consortium for accessibility purposes [54, 55]. Finally, Bezryadin et al. [54] note that the use of any of the colour-to-grayscale equations should be based on the area of application as no colour-to-grayscale equation is superior to the others.

2.3.3.1 Good skin tone based on lighting and skin type

Determining the ideal skin tone for human faces to ensure ideal image quality is not a trivial task as the combination of lighting and race/skin type make this measure difficult to identify. The ideal amount of lighting for a specific skin type may be best solved with heuristics; a heuristic approach is used by Sun [7]. Their procedure involves sampling a large set of celebrity images in order to train their proposed system. With the assumption that photos of celebrities contain optimal skin tones under good lighting conditions, the researchers gathered photos of different celebrities of different races/skin types and placed these photos into five different classes (sets of images) based on skin type (see Figure 2). The researchers surveyed the facial regions in each photo of each set and determined the ideal skin tone of each set.

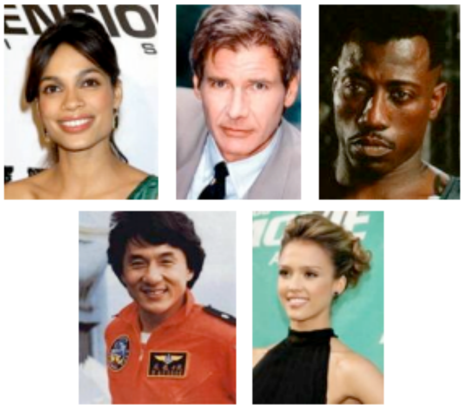


Figure 2: Classes of skin types [7] **(Page 3)**

The method, presented by Sun [7], of determining the ideal skin tone is important to my research as I use a similar technique to determine grayscale ranges for classifying the lighting condition in a video frame. Instead of surveying images of celebrities, I found images of signers under various lighting conditions and of different skin types/races and created models for classifying skin tones based on lighting. For example, I found a set of images of signers with fair skin whose facial region was visually appealing (good lighting). I obtained the average grayscale value of the facial region in each photo of the set and from these averages I determined that the grayscale range for a good skin tone for someone with fair skin was 99 – 130 for example. Since people of different skin types will be using the system, I created three different models (classes) to accommodate for three different skin types: dark skin, tanned skin, fair skin. Please see Section 3.1.5.3 for more information.

2.4 Visual Noise/Clutter

“Visual Noise” [25] is another important contributor to video quality and visual clarity. Visual noise is also known as visual clutter in the literature and is a main contributor to an environment’s visual complexity [9]. Rosenholtz [56] defines visual clutter as the amount of items added to a display that cause some form of excess visual complexity. Visual complexity may cause confusion for the viewer and may prevent them from accessing and reading all of the information easily. There is often not a clear-cut definition of the word clutter however. The literature defines clutter as a scene or real world environment that contains many items/objects [9, 57, 58]. These objects may become obstructed by other objects, densely populated or organized in a specific pattern [59]. Alvarez [57] found that it is also an object’s features, colour and shape orientation that may cause a higher perception of complexity. As well, objects are not the only elements that cause clutter; colours, contrast and textures also contribute to background noise [60] (See Figure 3).

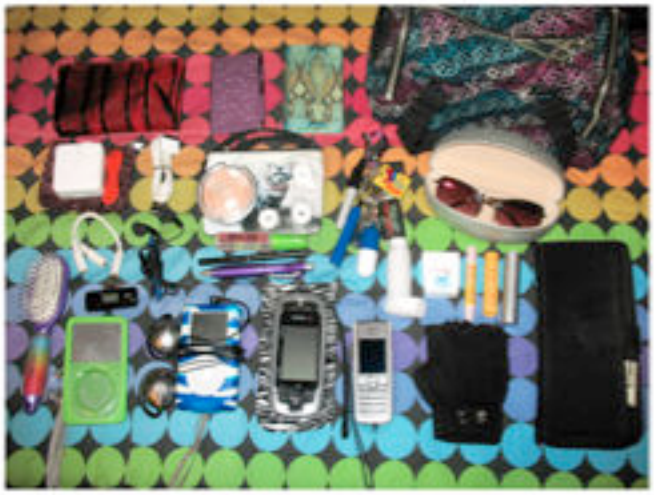


Figure 3: High cluttered image [58] **(Page 6)**

In terms of Signed Languages, certain elements of the background and foreground cause visual complexity and remove focus from the signer [34]. These elements include, windows (incoming light that darkens foreground elements) [38], noisy logos and murals [25], and moving background elements (humans / television sets). Fels et al. [25] note that a detrimental effect of increasing the signing space is that it also increases the probability of introducing an increased number of visually noisy elements into the background. To keep the focus on the signer at all times, recommendations include using a neutral, solid colour backdrop during video communication [29].

Bit depth [6], is another major contributing factor to visual complexity. Hibbard and Fels [26] and [29] state that the chosen combination of clothing, background and skin tone must provide enough contrast to prevent the appearance of meshing of colours in a scene (see Figure 4). A viewer should be able to easily distinguish foreground and background elements, and should have no trouble viewing objects and movements such as motions in the fingers and hands anywhere within the signing frame. Hooper [6] recommends that the number of colours in a shot should not exceed a certain threshold, as an excessive colour range will create a blending between colours. The clothing of a signer should also not contain any wording or visual patterns to prevent further visual complexity.

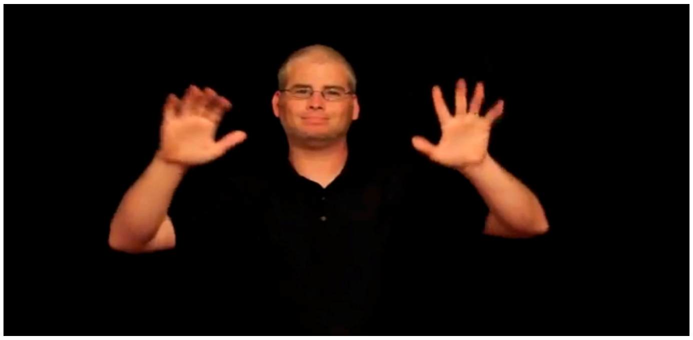


Figure 4: Meshing of colours (clothes/background) [61]

2.4.1 Studies on Performance and Perception

The literature states that clutter can cause many degrading effects on visual performance and perception. These effects include degraded search performance of specific items [56], crowding [9], masking [62], high demands on short term memory [57], and the degradation of performance of creating relationships based on objects in a scene [60]. Masking is a phenomenon by which the impact of an item is lessened because of other objects in a scenery (non-overlapping items) [62]; crowding is the degradation of the recognition of objects by the human peripheral and foveal vision due to crowding of items [62]. These effects should be considered when creating and designing efficient and clean interfaces, and real world applications such as traffic, road layouts and visual search applications [58].

Scenes that are comprised of several different colours, layouts, patterns and objects are also suggested to have significant effects on perceived complexity [62]. Studies attempt to develop a baseline for categorizing visually complex images, as information is limited when determining why people perceive some images as complex and others as not [9].

Oliva [59] attempted to determine the characteristics of environment complexity based on participant perception of different images of real-world scenes (See Figure 5). Their study used images such as kitchens, bedrooms and home entrances that were subjectively judged as visually complex. An environment that is visually complex becomes hard to define, open to misinterpretation and often the amount of information may become overwhelming to the viewer [63]. Participants were asked to group images based on complex versus simple scenes. Findings indicate that participants judged complexity based on colour distribution and variety, organization of the room and amount of objects in the image [59]. For Signed Language communication specifically, more information presented could cause a shift in focus from the signer to other areas of the screen during communication or cause a higher demand on the viewer's short-term memory due to the increased information presented in the shot [57].

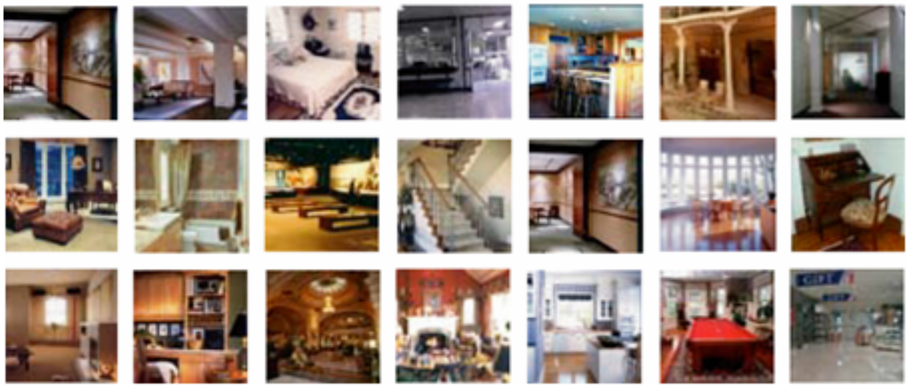


Figure 5: Perceptually complex indoor scenes [59] **(Page 4)**

There is a common belief expressed in the literature that Deaf persons have an enhanced visual ability because of the auditory deprivation they experience. Results from various studies suggest that although signers do perform better with some mental imagery tasks when compared to hearing non-signers, there are no major differences in visual ability between these two groups [64]. Bavelier [64] found the largest difference between the Deaf and hearing populations is that Deaf individuals are more distracted with peripheral imagery than hearing individuals are.

2.4.2 **Edge Detection**

Edge detection is defined by Senthilkumaran [65] as the detection of edges determined by localized changes in the colour intensity in an image. Identification of an edge is likely to indicate a change in depth, material, illumination etc.. Edge detection is able to tell the user important information about the objects contained in the image, such as quantity, object size and shape and structural information, and will reduce the amount of data to be processed in the image. An approximate measure of the amount of clutter in a visual display could have many benefits for applications such as interface design [56], information visualizations systems [60], visual search research [58].

Different techniques to measure clutter include image segmentation, segment areas of the image where most items are found, subband entropy (object redundancy), feature congestion, which attempts to capture the organization of the clutter, and edge detection. The edge detection technique is used to find the quantity of objects in the display and assumes that there are more items in the display if more edges are found [9]. Each algorithm surveyed by Van Den Berg et al. [9] achieves similar results and no algorithm was determined to be better than the other but can be used for different purposes.

The following paragraph will discuss the Canny edge detection technique, a gradient based edge detection technique that is used by the VAST system. A gradient edge detection technique typically follows three main steps to perform detection: (1) filtering is done to reduce noise, (2) detection is done to determine the intensity value of a pixel and it's neighbours and (3) use thresholds to determine which pixels qualify as edges [65]. Common gradient-based detection operators include the sobel, prewitt and roberts edge detectors [66]. Although these edge detection techniques may have also been sufficient for the purposes of the VAST system, as only an edge approximation is needed, Maini [66] notes that these operators are sometimes sensitive to noise and that the Canny edge detector performs better and is less sensitive to noise in most scenarios.

2.4.2.1 Canny Edge Detector

The Canny edge detector, presented by Canny [67] in the 1980's, is a gradient based edge detection algorithm that still rivals many of the newer edge detection techniques presented today [68]. Canny [67] identifies three common criteria that all edge detectors should posses: (1) low error rates (the reduction of the false detection of edges), (2) localized data points (the edges from the original image should match up with the edges identified in the processed image) and (3) single response (only return one edge point for a true edge). The algorithm works in four main steps. The first step is to smooth the image using a Gaussian blur to remove excess noise. The second step is to obtain the image gradient and compute the gradient magnitude and direction for each pixel. The third step is to compare the magnitude of a pixel with it's neighboring pixels to determine if it is an edge. In this step it also performs non-maximum suppression to map only one pixel to an edge.

Finally, two thresholds are used to rid the edge map of non-edge pixels. If a pixel is below the low threshold it is removed, if it is in between the low and high threshold it is labeled a "weak edge" and if the magnitude of the gradient is above the high threshold it was labeled a "strong edge". Only weak edges who neighbours a strong edge will be kept [67].

The main benefits of the Canny detection for the purposes of use in the VAST system are the ability for detection in noisy images, quick processing and the single response restriction. The reduction of non-edge pixels in the edge map and the results of non-maximal suppression will reduce the amount of pixels returned as edges and may prevent the system from classifying an image wrongly because of excess edge pixels.

2.5 Signing Space

Signing space is the amount of three-dimensional space in front of the signer that is used for communication [69]. The use of signing space is particularly important in many Signed Languages [70] as it helps to convey grammatical meaning [14] and is a structuring principle of the language [71]. For example, when dealing with location, Deaf individuals often create a mental imagery of the location they are communicating about and relate objects in the environment to the three-dimensional space in front of them [69]. Similar to location, relationships between time and order, and signing space are created to display time of day or points in a sequence [69].

In online communication the virtual space of an individual often becomes limited [26] by the viewing area of the recording device. Meaning is lost if gestures/signs appear outside of the camera's viewing area, are obstructed by objects or are not relatable in a digital versus in person space. Keating and Mirus [5] and Hibbard and Fels [26] discussed the differences found between face-to-face and online communication. Keating and Mirus [5] showed that participants often implement alternate ways to express signs to improve viewing and comprehension in the digital space. Users were also found to change the delivery of a word entirely to account for the viewing area of the recording device. Similar to the findings of Keating and Mirus, Hibbard and Fels (2011) also found that participants often presented signs in ways that deviated from normal face-to-face interaction. Firstly, the use of the z-axis was used several times to enhance the effect of a word [26]. As well, some signers only showed parts of their face and hands when communicating using vlogs; a technique that limits the ability for full comprehension and one that would often be impossible in a face-to-face scenario.

Hibbard and Fels [26] also discuss three signing space classifications used by signers when creating vlog content: small, medium, large (Figure 6). In a large signing space the viewer is able to see all parts of the signer above the mid-torso (heads, hands, elbows) and there is space in the horizontal for the signer to communicate with (Figure 6a). A medium signing space is similar to a large space except that the elbows of the signer are cut off by the bottom of the screen (Figure 6b). In a small signing space, a signer's hands, head and torso are crammed into the shot (Figure 6c). The use of a small signing space is likely to cause information loss and usually occurs when the signer is positioned too close to the camera or misuses his/her digital signing space.

In this thesis work I substitute the wording of large signing space [26] to normal signing space. I do this so that readers who are unaware or unfamiliar with the various signing space classification may interpret that a normal singing space is the more optimal signing space to use out of the normal, medium and small classifications. For example, in the user study I use two settings of signing space: normal and small. I believe that using the word "normal" instead of the word "large" may be less ambiguous for readers when determining which condition is the better of the two.

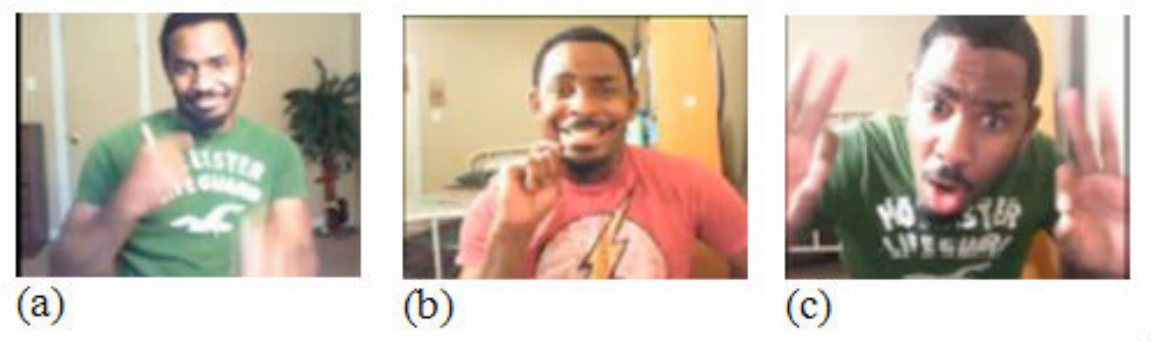


Figure 6: Signing spaces a) normal (optimal setting), b) medium c) small [26] **(Page 62-63)**

2.5.1 **Signing space in relation to television production**

Research from Bauman and Murray [72] found that film/television production and Signed Language are very similar in nature and in the way space is used to structure the production. Film and vlog production are conveyed on multiple levels and communicated to the audience through many elements in a scene. Rasheed [73] suggests that in film, items that include the delivery of lines from actors, the music, lighting, camera movements, scenery etc. are combined to create the optimal production environment. This is also true for vlog production and much of the recommendations for proper vlog creation are taken from the film and television language [72].

Video shots and video production styles in vlog and film/television are commonly used for the purposes of strategically relaying information to the viewer. The 'talking head' production technique has been used to convey information for decades in television news broadcasting [74]. The term 'talking head' refers to the broadcast style where the broadcaster's head and part of his/her torso is displayed on screen while the broadcaster talks to a non-existent audience [3]. This production technique is comparable to the style of production signer's use when creating vlogs. Robertson [74] suggests that the power of the talking head setup stems from the illusion that the television anchor is talking directly to the viewer, or that the viewer is watching an informative conversation taking place between two people. An example image of signer using a 'talking head style production technique is shown in Figure 7.

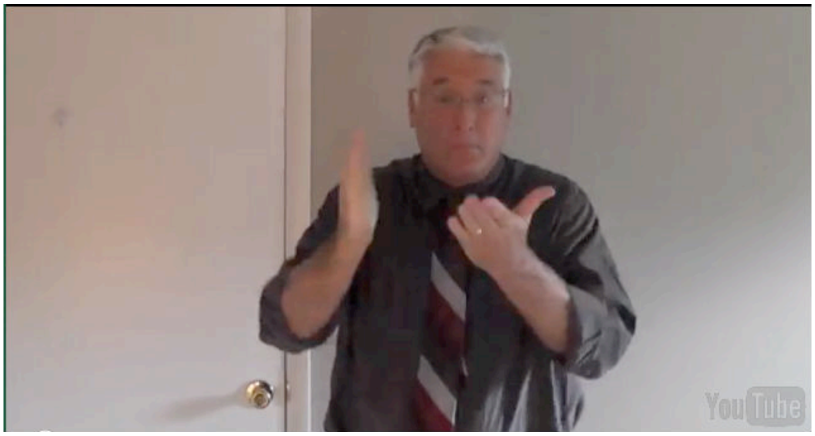


Figure 7 Sample vlog using a talking head production style

2.5.2 **Shot Classification**

Shot classification is a technique described in the literature used to understand what or who is on screen during televised and pre-recorded productions. Uses of shot classification include video lookup, video summarization, determining interesting moments, search etc. The papers reviewed on shot classification in this section are performed using rule-based algorithms applied to a specific domain, such as sports videos. For example, Xu et al. [75] first define three different video shots they are attempting to identify in soccer videos: (1) a global shot where the camera is pointed at the field, (2) a zoom-in shot where the camera is zoomed in on players still playing the game, and (3) a close-up shot where the grass from the field is no longer visible and the player takes up a large portion of the screen (See Figure 8). Using common knowledge that the video will display grass during a global shot or a zoom-in shot, the authors use the colour values of the pixels, among other things, to identify what shot the video is displaying. The authors can conclude that the video is displaying a close-up shot if very few pixels are green in the video and a global shot if a large percentage of pixels are green. Kolekar [76] also uses a soccer video dataset to perform shot classification using colour values, similar to [75], as well as edge density.

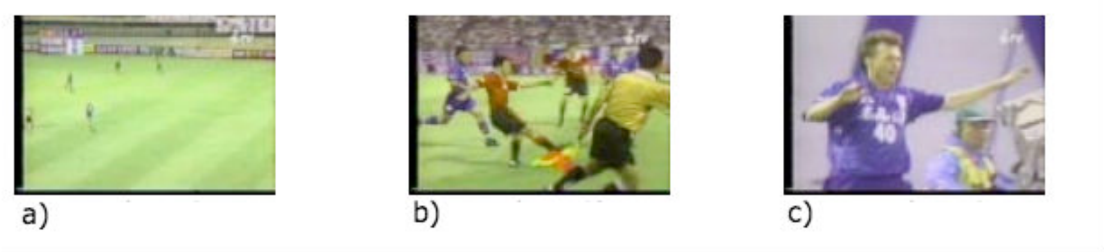


Figure 8: Video shots a) global b) zoom-in c) close-up [75] **(Page 4)**

Shot classification often relies heavily on heuristics; things such as green pixels on the screen, the position of the person in the shot, or the identification of edge pixels is enough information for shot determination in some scenarios. The one drawback to these techniques is that the rules used for certain datasets are not extensible and cannot be used on other video datasets. Shot classification applies to my thesis and specifically to the way I determine signing space and background as I use rule based-algorithms to determine the position of the signer in the shot and the background clutter. For example, if my algorithm determines that the signer's facial region is larger than half of the screen I classify this shot as a small-signing space. I use details about the screen that would only apply to talking head style videos to conclude information about what is going on in the video.

2.6 Facial Recognition

Vast uses a facial detection algorithm to determine the position and size of the facial and body region in each assessed vlog frame. The goal in this section is to summarize different detection algorithms to determine which technique would best benefit VAST.

The key to building aware and intelligent image recognition systems starts with the ability of a system to recognize and identify objects and faces in an image. Face detection has become a popular research topic in recent years as the application for it has become ubiquitous [77]. Over 150 facial detection techniques have been developed in the past 5-10 years [78]. These techniques vary in method and are proposed using various colour models. The goal of any facial recognition technique is to identify and locate faces in images, regardless of the expression, pose, scale and illumination condition of the facial region [79]. Facial recognition is currently being employed for surveillance, computer vision, robotics and smart computing among other applications [80]. Viola [81] notes that facial recognition software should not be put to use until the specific method used yields at most a 1 in 1,000,000 failure rate.

Face recognition is not a trivial task because of the variability presented in every face. Figure 9 displays examples of face images taken from the web [79]. The various illumination conditions, skin tones, poses and gestures of the facial images make it increasingly difficult to accurately identify each face. Items that obstruct the face, including glasses and hair, only compound the problem [82].

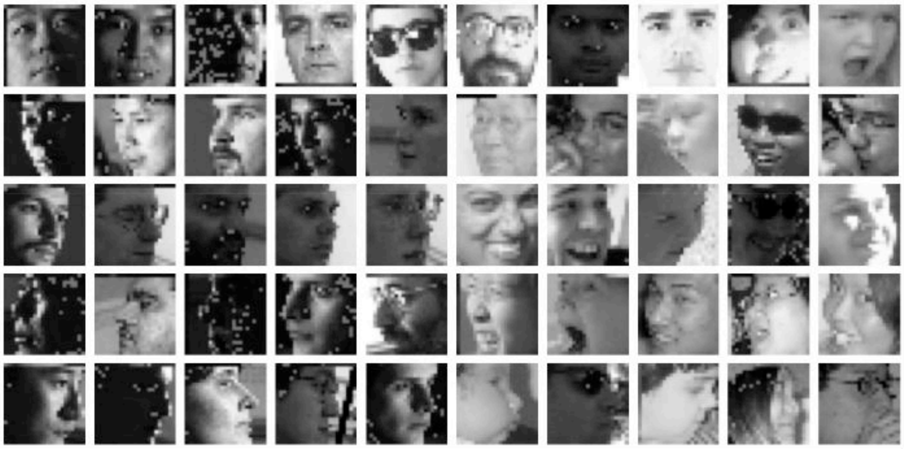


Figure 9: Face images [79] **(Page 2)**

Four approaches to face recognition are discussed in this section: (1) feature based identification [83, 84]; (2) template matching [82]; (3) appearance-based identification [81, 85]; and (4) knowledge-based techniques [86]. Appearance based methods rely heavily on training; template matching techniques use sample face images in an attempt to match the current template with a sub-section of the image. Knowledge-based methods rely on pre-defined definitions and known visual components to classify features. Feature-based approaches identify features common to human faces such as geometric shapes and patterns. These techniques have been applied to areas such as gender and race identification [87], face-relighting [7] and coding scheme improvements [84].

2.6.1 Appearance Based Approaches

Appearance-based approaches to face recognition are accepted in the computer vision community as they have many benefits over typical pixel-by-pixel comparison methods [81]. These methods are ones that survey several thousand images and learn from patterns found in the images to perform detection [79]. Images are often classified into positive and negative categories. Positive image sets contain the object the researcher is attempting to identify and the negative image sets contain images of patterns and other objects that potentially help to create a stronger system. Due to the of advances in computer technology, appearance based methods have flourished in consumer products over the years and many are implemented in popular computer vision libraries. Applications for appearance-based methods are implemented for the detection of human faces/bodies [81], automobiles, pedestrians [85] etc.

2.6.2 Feature Based Approaches

Feature-based approaches use various known information about the human face to perform detection and is used for less complex processing [83, 88]. These approaches tend to be invariant to conditions, such as pose and expression, which cause detection issues in other non-feature based techniques [79]. Known methods include the use of geometric features of the face, edges and colours spaces.

One of the most popular feature-based techniques is skin detection. Skin detection is the identification of skin pixels in an image [88]. This method has very little overhead and is quickly processed on a pixel-by-pixel basis [83]. The main benefits to this technique are that it can be run in real time [52] and is rotation, pose and expression invariant [78]. Some of the main drawbacks are that this technique is dependent on the illumination condition and it is vulnerable to detecting objects whose colour is similar to skin colour.

The YCbCr colour space, composed of one luminance component (Y) and two chroma components (Cb, Cr), is typically used for skin identification and processing, as the chromatic colour planes, with luminance removed, provide an accurate skin model representation [83]. Other colour models including YUV, HSL and TSL are also outlined in the literature as a way to locate skin pixels [52]. Defined ranges in the CbCr (chrominance) space are established for the identification of skin like regions and detection will typically locate these pixels regardless of human race and sample parameters to identify skin pixels are as follows:

$$Skin\ Pixel = Cb \geq 77\ \&Cb \leq 127\ \&\ Cr \geq 133\ \&\ Cr \leq 173 \qquad (1)$$

where Cb is the blue difference colour component and Cr is the red difference colour component [88]. A pixel is classified as skin, if the pixel's colour values fall within the defined ranges. Other algorithms for skin detection include the use of Bayesian classifiers [52], single hypothesis schemes [84] and histogram methods.

Habili [84] applies skin pixel identification to Deaf video content. The algorithm for skin detection and segmentation for facial and hand recognition will convert non-skin pixels to black and will not modify the pixels determined to be skin (see Figure: 10). The areas of the image/video that are not classified as skin are blacked out. This method limits the processing of all areas of an image and focuses on only identified areas and may also provide another technique for compression algorithms [84]. Habili (2001) assumes however, that only supplying the viewer with areas of the screen identified as skin is enough for full comprehension of Deaf vlogs.

Figure: 10 Only showing skin pixels [84] **(Page 11)**

2.6.3 **Template Based Approaches**

Another popular technique for face detection is template matching. In this technique, face templates are compared several times to an image to determine if the template exists within the image [79]. Traditional template based approaches are limited by certain constraints, such as pose and expression, however templates have been developed to account for these scenarios. Developed templates include: correlation, deformable [80] and ratio templates [82].

The comparison of many templates against one image is known as multitemplate detection [89]. Multitemplate detection is required when the chances of a face being located using a single template are low. This form of detection, however, can become exhaustive as the number of templates increase. Template matching algorithms easily fit the need of operations involving less complex detection and would not be suitable during more complex operations such as identifying, for example, 100 different unknown persons in an image [90].

2.6.4 **Knowledge Based Approaches**

Knowledge based approaches are ones that use a pre-defined set of rules to identify faces [80]. These rules are adjusted to guide the search process and are created using human knowledge that is often not easily translatable to computer vision. A popular technique for knowledge-based methods includes the use of mosaic images. A mosaic is a set of images represented at various resolutions [86]. The underlying technique is to convert images into lower resolutions and detect face candidates by extracting patterns reminiscent of human facial features at sub-divided cells of resolutions as low as 2 by 2 and 4 by 4 pixels (see Figure 11). Detection rules are developed for the identification of features of the face and further processing will continue if a candidate is located. One of the main drawbacks of knowledge-based approaches is that since the pre-defined rules aim to capture a normal face, creating the rules to be overly detailed or too general and can lead to many false-positives.

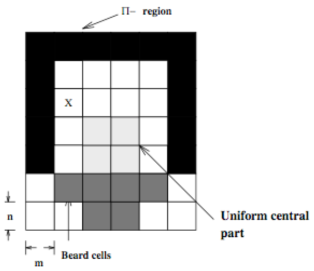


Figure 11: 4 x 4 pixel region [86] **(Page 2538)**

Studies which have been carried out with Signed Languages have focused on the recognition of manual sign [17]. However, as outlined in Section 2.1, sign alone is not the only action needed for the full comprehension of Signed Language communication. The research behind facial recognition is a stepping stone to understanding the distinct motions and emotions in human faces. Using the techniques discussed above, non-manual behaviours such as eyebrow and lip movements could be obtained and tracked for the learning and understanding of Signed Language communication, although, many factors in computer vision still need to be resolved.

3 System Design

This chapter provides a detailed overview of the methods used to construct the Vlog Analysis and Suggestion Tool (VAST), and the study and procedures designed to evaluate the VAST system. The VAST system was created to assess Signed Language vlogs for three technical factors, and provide feedback and suggestions to users for improvement. The three technical factors chosen specifically for Signed Language video/vlogs were: (1) lighting, (2) signing space, (3) background clutter. These factors were selected as they directly affect the comprehension of viewers and are commonly used by Signed Language instructors to teach students how to properly format a video in terms of quality.

To determine overall vlog quality, VAST uses existing image processing techniques and applies them to each frame in the vlog. Once the system assesses each frame, a quality recommendation is generated and output to users. In this chapter, details of the software design of VAST will be discussed.

3.1 Overview of System

The purpose of the VAST system is to analyze Signed Language video content and assess the lighting, signing space and background clutter, and report the result of that assessment to the user. The goal is to provide users with a quick quality assessment of their video and inform them of potential problems certain factor settings may cause. To do so, the system accepts a video file and will begin processing it upon user initiation. Output of the analysis results is provided as text and images. The analysis results not only contain the report of visual quality but also suggestions for improvement. The system was designed with a simple to use interface and the user will only have to perform two actions to receive an assessment of their video. The following images display three screenshots of the system when the user has loaded the video into the system (Figure 12), when the system has begun processing the video (Figure 13), when the system has reported results to the user (Figure 14).

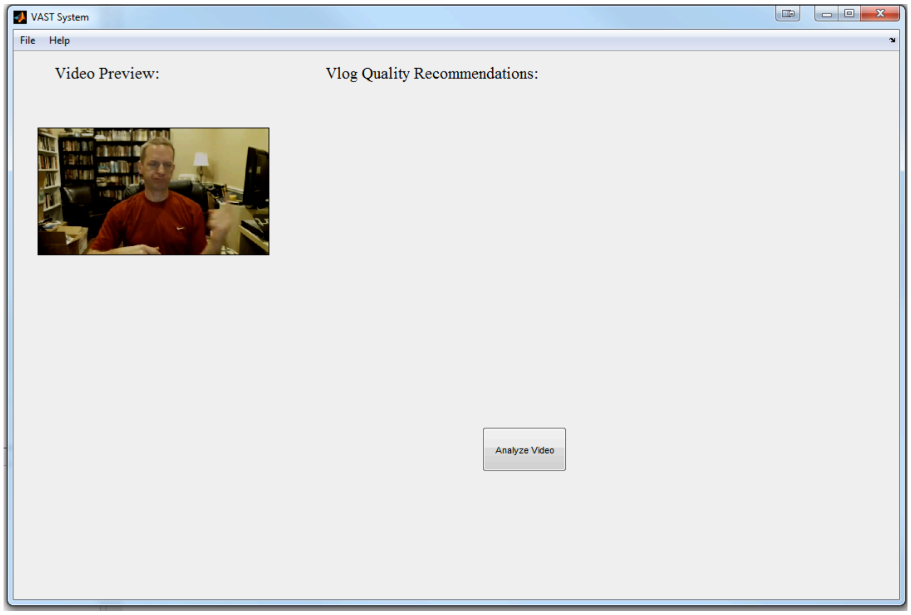
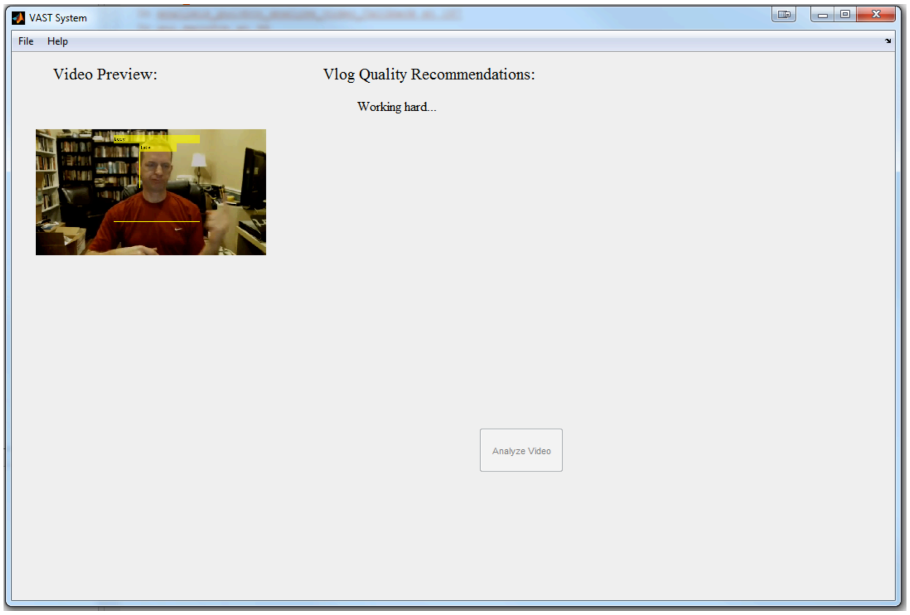
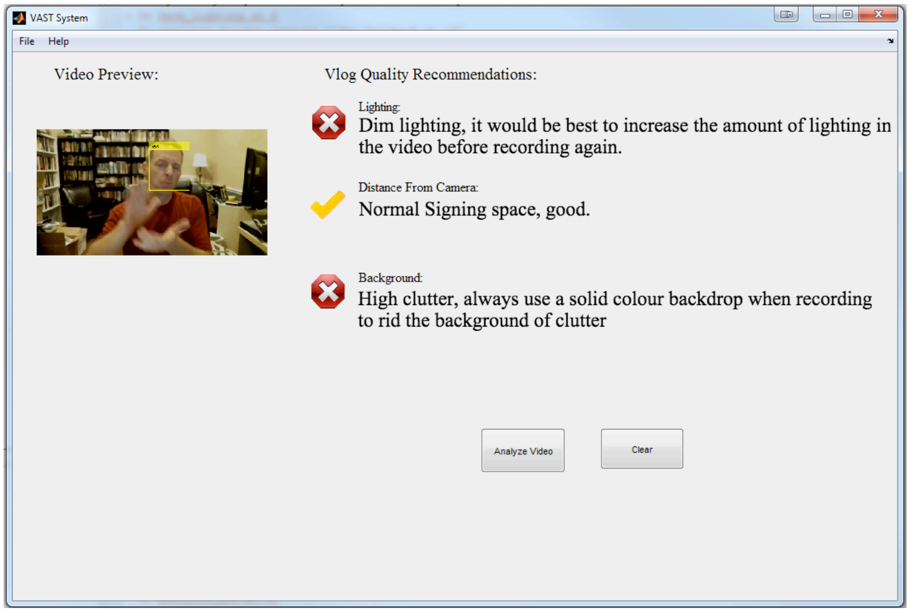


Figure 12 Vlog uploaded to system, waiting for the user to indicate that it should begin analyzing for quality



Working hard...

Figure 13 Screenshot during the image analysis process. Users will be shown which frame is currently being analyzed



Video Preview:

Vlog Quality Recommendations:

✗ Lighting:
Dim lighting, it would be best to increase the amount of lighting in the video before recording again.

✓ Distance From Camera:
Normal Signing space, good.

✗ Background:
High clutter, always use a solid colour backdrop when recording to rid the background of clutter

Figure 14: Analysis complete and recommendations are output to user in textual and visual form

3.1.1 Interface and Flow of actions

The VAST graphical user interface is designed to allow for easy navigation of important functions including loading of a video into the system and initializing analysis of that video by the system. User feedback of the three technical factors is provided separately and consists of an image of the letter x or an image of a checkmark used to indicate whether a recommendation from the system was positive or negative and a text recommendation. At this point in time, feedback in Signed Language is not available although it could be a future addition. Figure 15 shows a workflow of actions the user and system will take to produce the final output.

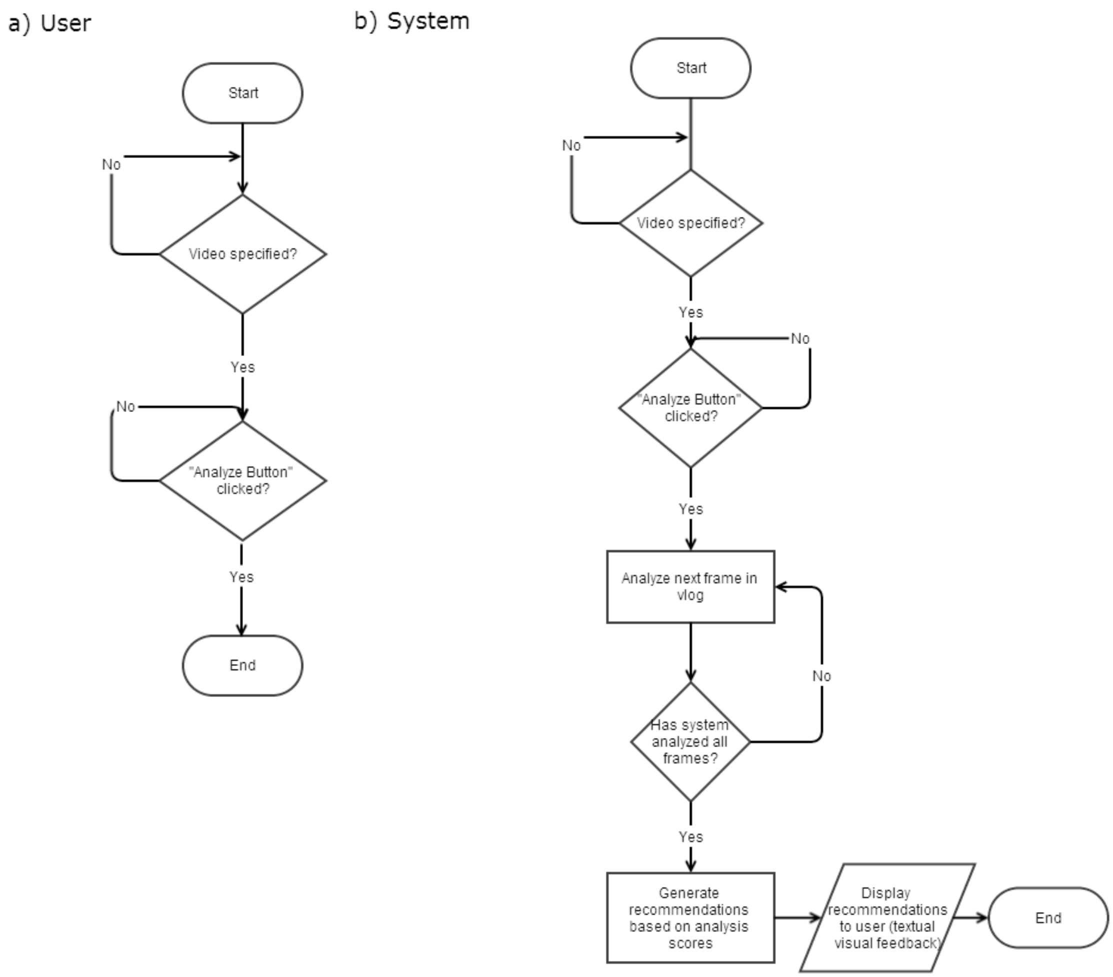

a) User
Start
No
Video specified?
Yes
No
"Analyze Button" clicked?
Yes
End
b) System
Start
No
Video specified?
Yes
No
"Analyze Button" clicked?
Yes
Analyze next frame in vlog
No
Has system analyzed all frames?
Yes
Generate recommendations based on analysis scores
Display recommendations to user (textual visual feedback)
End

Figure 15: Workflow diagrams for user and system

3.1.2 The Viola-Jones Face and Body Detection Algorithm

The Viola-Jones face detection algorithm [81], is an appearance-based method that is comparable in speed and robustness to the best face recognition methods available [79]. This technique has had a strong impact on object and facial recognition in the 2000's [79] and many of the newer facial detection techniques build upon this framework [85]. The method can be run at 15fps and is trained using over 100 million face and non-face images [91]. The algorithm is comprised of three separate functions. These three functions include: AdaBoosting,, the integral image and a cascade structure for the quick discard of non-facial sub-windows [81]. These functions will be discussed in the following paragraphs.

Results from the Viola-Jones algorithm are promising for my research. The average run-time of this algorithm is roughly 0.67 seconds on a 364x288 pixel image and high accuracy for detection is achieved [81]. The run-time is better when compared to other popular facial detection techniques by several times [79]. Due to the cascading phase of the algorithm, only 10 features on average are processed on negative sub-windows [81]. Only the most interesting and important sub-windows are processed by high-level stages. The independent functions presented in the Viola-Jones framework seem to work well together to create an accurate and efficient system. Sample results from face detection are displayed in Figure 16.

I use this algorithm for face and body detection. I did not modify the algorithm in any way and use it as a black box system. Although the Viola Jones algorithm is primarily used as a face detection algorithm, it can be trained to detect many different objects. The Image Processing Toolbox in Matlab (the system that I used to develop VAST) also provided a way to detect the body of an individual in an image.

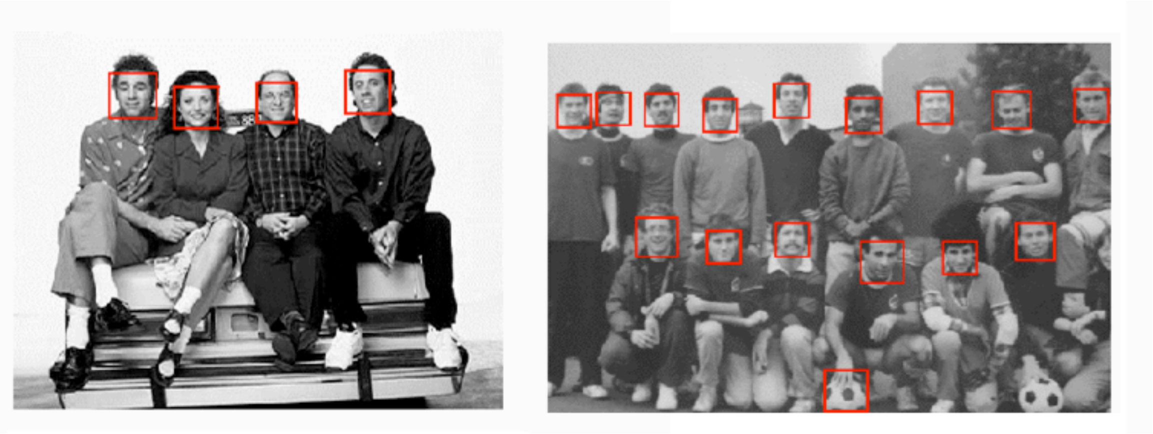


Figure 16 Results from Viola-Jones face detection [81] **(Page 152)**

3.1.2.1 Integral Image

The integral image, also known as a summed area table, is a conventional way to calculate the sum of the pixel values of an area in an image, with the addition of only four pixel values. The introduction of this method saves computation times and avoids the re-calculation of sums during every area calculation. For this to work, each pixel is re-calculated using the summation and subtraction of the surrounding pixels in one single pass. The main benefit to using the integral image is that finding the area of any rectangle in an image can be run in constant time.

3.1.2.2 Cascading

The Viola-Jones algorithm first extracts rectangular features from 24 by 24 pixel sub-windows of the image to analyze patterns in the pixel values. A rectangular feature is simply a rectangle placed on the image and the value computed for that feature is the sum of pixels under that rectangle. Each image is converted to grayscale so that values of each pixel are bound between 0 and 255. There are three types of rectangular features used in the algorithm: a two-rectangle feature, a three-rectangle feature and a four-rectangle feature (Figure 17 A-D) [81]. These features are computed in each sub-window of the image at different sizes. Each sub-window will produce over 160,000 different features computations. The values obtained for each feature are later used in determining interesting facial features in the image. Training for the identification and understanding of facial features is discussed in Section 3.1.2.3.

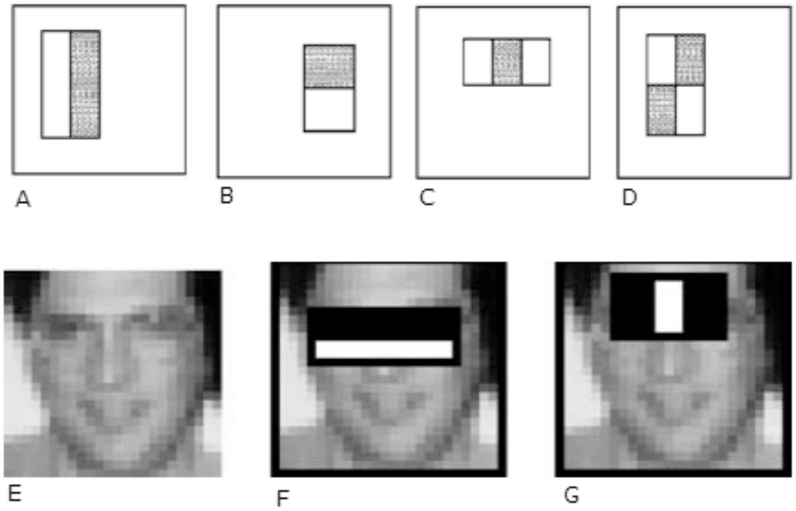


Figure 17 Haar-like Features [81]. **(A-B) Two-rectangle, (C) Three-rectangle (D) Four-rectangle (E-G) Features applied to 24x24 pixel face image**

Once all features are computed, they are taken through a step of classifier cascading [81]. This step significantly reduces the processing time of the system as it eliminates a large amount of non-useable sub-windows based on their computed features. Sub-windows consisting of parts of the background, for example, will be discarded early in the cascading process, as the feature values extracted from them will not be determined to be significant.

3.1.2.3 AdaBoosting

The learning phase of the algorithm is performed using a technique called AdaBoost, which is broken down into several rounds known as boosting rounds. This phase is performed before running the detection on an image and is a way to train the algorithm to detect faces. Thousands of face and non-face images are used to build a linear combination of "weak classifiers" (features which on their own cannot reliably determine where a face is in the image), which combine to create a strong classifier. The goal of AdaBoost, in each round, is to determine the single best feature that separates a face from a non-face. This is determined based on the lowest error rate of all of the features using a specified threshold. This threshold is chosen such that the fewest number of face images are misclassified.

Each time a feature is selected in a boosting round, it is added to a linear combination and termed a weak classifier. The reason for the weak classifier name is that on its own, the single feature cannot determine a face. Only the joint combination of weak classifiers, termed strong classifiers, is reliable for accurate face detection.

The linear combination created with the boosting rounds is further broken down into smaller aggregates of weak classifiers, simply termed classifiers or cascading stages. The creation of multiple classifiers is done to ensure that if a sub-window fails one classifier, processing time is not wasted testing it with subsequent classifiers [81]. These classifier stages are created using desired false-positive and false-negative rates. As the level of classifiers increase, the difficulty and false positive rates do as well. Depending on the classifier stage, there are a variable number of weak classifiers that are combined to create the classifier.

3.1.3 **Analysis of Quality Factors**

Once the vlog has been opened with the VAST system and the user has chosen to process it, VAST will begin to analyze frames in the vlog using three functions which determine the following:

1) Pixel intensity values in the facial region of the signer (Grayscale conversion algorithm)

2) Amount of edge pixels around the signer in the frame (Canny edge detection)

3) A relational measure of the signer's facial and body region compared to the screen width and height

To begin the analysis process, the Viola-Jones recognition technique is used to identify the facial and upper region of the signer's body in each frame. This algorithm allows extraction of a sub-image of the current frame and is able to spatially locate and identify positions of different features of the signer including the x, y coordinates of the facial region. Spatial information about the signer is beneficial when determining the quality score for each frame for the following reasons: (1) only the signers facial region is needed when determining the lighting in the shot, (2) knowing where the face and body are located relative to the video frame will help to determine how much signing space is used by the signer and (3) knowing the location of the signer in the frame will prevent the algorithm from including the signer when determining background clutter. Sample facial region segments extracted from various vlogs using the algorithm are displayed in Figure 18.

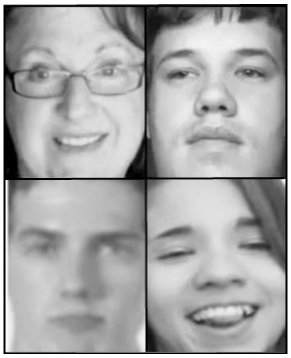


Figure 18: Facial segments identified using Viola-Jones algorithm

3.1.3.1 Assumptions

When assessing videos in VAST the following assumptions were made that:

1. The signer in the vlog is not continuously signing in front of his or her face.

2. The signer in the video does not have any facial obstructions, such as a large amount of facial hair, or is wearing any clothing that may cover the face.

3. No signer will be positioned so far away from the recording device that even though they are in the shot, viewers cannot view their gestures or facial expressions.

4. The Viola Jones algorithm will detect the facial region and body of the signer correctly.

3.1.3.2 Lighting

Lighting is comprised of two separate elements: luminance and illuminance. The illuminance is the amount of incidence light falling on an object and the luminance is the amount of light radiating from an object. There is a direct mapping between the light intensities in an image and real world luminance. When an image is over or under exposed, problems when viewing the Signed Language content may arise. It is important that the viewer is able to see the facial region and all parts of the signer to ensure that the message is communicated clearly. The photometric luminance of the facial region of the signer in the vlog is determined using a colour-to-grayscale measure [7]. A colour-to-grayscale measure is a technique used to identify the pixel intensities or the luminance in an image. In VAST, only the facial region of the signer is used to assess lighting; having good quality lighting of the face is very important in the understanding of Signed Language [38]. In addition, surveying the whole image would result in inaccuracies, as lighter and darker coloured objects in the foreground/background of a video would cause the average grayscale value to increase or decrease but has no relationship to the actual visibility of the signer.

To calculate the grayscale value of a pixel, the luminance value is determined. This value represents the light intensity of a pixel and falls between the ranges of black and white (0-255 respectively). The luminance calculation:

$$Y = 0.2989 * R + 0.5870 * G + 0.1140 * B \tag{2}$$

for a pixel, where R, G and B represent the red, green and blue values of the current pixel [54].

One drawback to using the Viola-Jones algorithm for illumination assessment is that this algorithm does not segment around the contours of the human face. Often the individual's hair and small portions of the background are returned with the facial region image segment. To account for this, 3% of the overall width and height of the image segment is clipped from all sides of the sub-image to ensure that the sub-image contains mainly skin pixels.

It is also important to account for lighter and darker skin tones. Although an environment may have optimal lighting, the result from a light intensity assessment on an individual's face with a darker skin tone may be different from a light intensity assessment on a facial region with a lighter skin tone. Because of this, three different categorization values (values that determine the analysis score of the image) had to be created to categorize the lighting of each frame in the assessment process to accommodate for this requirement. The person using the system must select which skin tone category they fall into.

3.1.3.3 Signing Space

Signing space is the amount of 3D space in front of the signer used for communication [69]. In video, signing space often becomes limited and is bounded by what the recording device is able to capture. The user is tasked with creating signs in the viewing area of the camera without having their hands or parts of their body going off screen, which would cause signs and gestures to be lost. Signing space may also affect how much of the signer's face is visible. Signers who do not use enough signing space often gesture in front of their face and block necessary facial expressions needed for communication. Facial expressions contribute to the grammar of the language and are important to ensure valid communication.

An approximation of signing space is determined using known spatial and positional values of the signer's facial region and body within the frame. To compute signing space the system requires the following information: (1) the approximate distance in meters between the signer and the recording device; (2) and how much of the signer's body is visible on screen (e.g., can the viewer only see the signer's head in the shot or is the signer's shoulders and torso also visible). General information about the video, such as height and width of the video frame is also required.

To calculate signing space, first, the ratio of the width of the video frame and the width of the signer's identified facial region in pixels is used to obtain an estimate in percentage of how far the signer is positioned from the recording device. The percentage value is determined

Distance from camera (%) = ((head width in pixels) / (video frame width in pixels)) * 100 (3)

where 'head width in pixels' is the width of the region the Viola Jones algorithm returns when a face is detected in the image. The distance from camera percentage value will decrease as the distance of the signer from the recording device increases.

The ideal signing space for vlogs is one where the entire head is in view (i.e., the top of the head is visible), the torso above the belly button is visible and the arms and hands can be seen when the arms/elbow form about a 120 degree angle away from the body (see Figure 19a). The signing space that is being used by the signer is determined using the obtained position of the signer's facial region relative to the frame height and is illustrated using equation 4:

Vertical Space = (lowest point of facial region) / (screen height) * 100 (4)

The vertical space is the percentage of the distance from the lowest point of the signer's face to the bottom of the frame. The higher the percentage, the closer the facial region is to the bottom of the frame. If the signer's detected facial region resides above the mid-point of the y-axis and below the top of the screen, I assume that all or some of the signer's torso is visible in the frame. This however, depends on how high above the mid-point the lowest part of their identified facial region is and the distance they are determined to be from the recording device. If the highest point of the signer's facial region is detected underneath the mid-point of the y-axis and they are positioned in close proximity to the recording device, it is concluded that their head and at most their shoulders are visible to the viewer.

Figure 19a and Figure 19b display two scenarios of signing space, one a normal signing space (Figure 19a) and the other a small signing space (Figure 19b). In Figure 19 the signer's head is above the mid-point of y-axis and the width of his facial region is small compared to the width of the frame; it will be concluded that he is using a normal signing space. In Figure 19b, the lower half of the signer's facial region is below the y-axis and the determined width of the facial region is large in comparison to the width of the screen; it will be concluded that the signer is not using enough signing space.

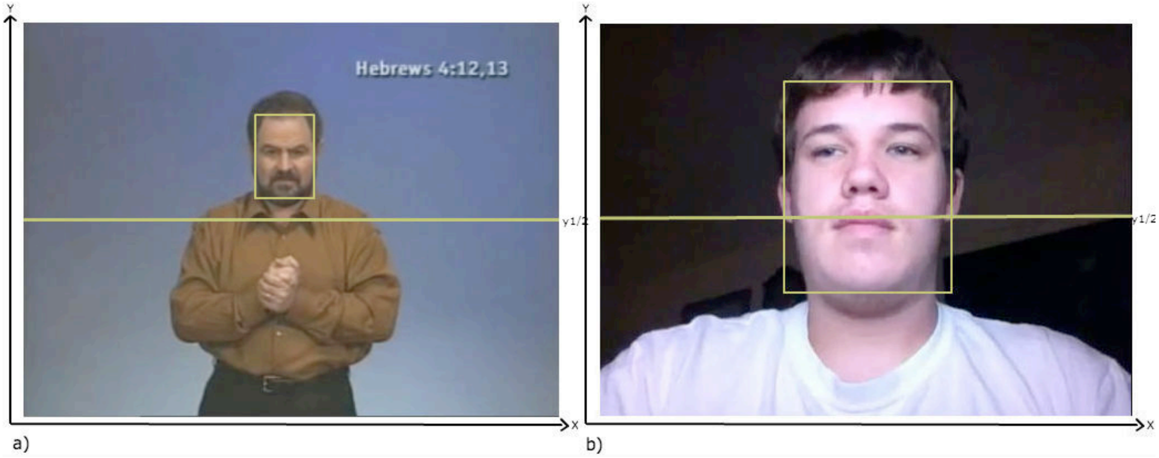


Figure 19: Normal signing space; b) small signing space

3.1.3.4 Clutter

Clutter is defined as the number of objects/items in a visual display. A scene becomes more cluttered as the number of items in it is increased [9]. The goal of this research in determining clutter is to obtain a general understanding of the number of items in the background/foreground of the scene to determine how cluttered it is. This research uses the Canny edge detector [67] to generate an edge map for the image frame. A discussion on how this detector works is located in Section 2.4.2.1 of this document.

The second step in this process is to remove the signer from the resulting image (see Figure 20 b). The main reason for this is that the signer should not be considered as clutter, as they are the most important element in the vlog. Eliminating the signer from the frame will remove the possibility of obtaining edge pixels where the signer is positioned. To remove the signer from the frame, the Viola-Jones algorithm is used on the original image to identify the pixel regions associated with the head and torso of the signer. The Viola Jones algorithm automatically outputs this information when it is run on an image. These pixels are set to 0 in the resulting image from edge detection. Figure 20 displays a vlog frame prior to it being adjusted and the resulting image after running edge-detection and removing the signer identified pixels.

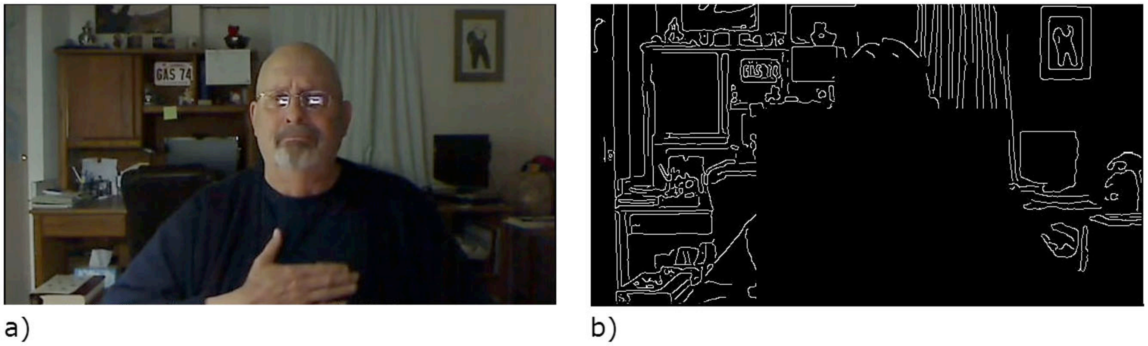


Figure 20: a) Without edge detection applied; b) Image after running edge detection

The final step of this process is to assess the percentage of edge pixels that exist within the image frame. To perform this action, the following equation is used:

$$\text{Total usable pixels} = \text{frame\_pixels} - (\text{Head\_pixels} + \text{Body\_Pixels}) \tag{5}$$

where the frame_pixels is the width multiplied by the height of the frame and the head_pixels and body_pixels variables are the number of pixels found for the face and body areas of the signer in the image frame, determined by the Viola Jones algorithm. Finally, the percentage of clutter in the current frame is determined by the following equation:

$$(\% \text{ of clutter}) = (\text{total\_edge\_pixels} / \text{Total usable pixels}) * 100 \tag{6}$$

where the total_edge_pixels variable is the number of edges identified by the Canny edge detection. The head and body pixels represent the number of pixels returned from the Viola Jones algorithm when it is run on the current frame.

3.1.4 **Frame analysis and array storage**

Each time a frame is analyzed for the grayscale average, signer positional information and percentage of edge pixels, the results are categorized and stored into three separate arrays (see Figure 21). These arrays are created with size n, where n is the number of frames in the video. It is important to note here that each node value is independent from the node value before and after it.

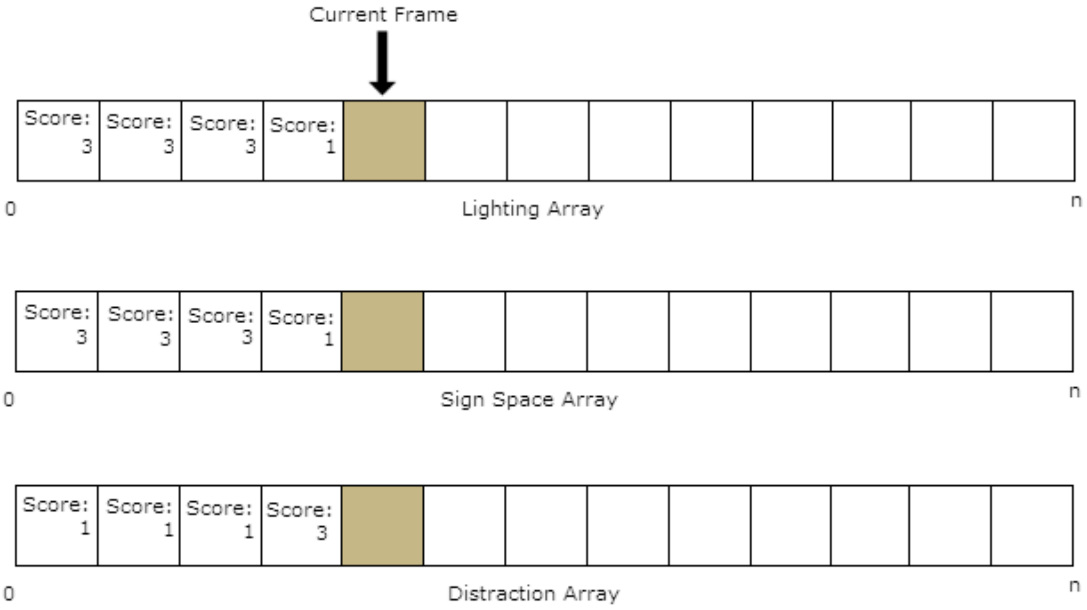


Figure 21: Individual frame scores and arrays used to store vlog quality totals

The values stored in each array node are not the raw percentages calculated but rather a score ranging from 1-3 for signing space and background and 1-5 for lighting. The scoring strategy is outlined in Sections 3.1.5.1 (lighting), 3.1.5.2 (signing space) and 3.1.5.3 (background).

3.1.5 **Scoring strategy for each factor**

After a frame is analyzed for a specific factor it will receive a score for that factor. Individual frame scores for each factor are stored until all the frames are analyzed. The mode is then used to determine the feedback given to the user about the quality of the factor. The scoring thresholds are determined using a collected library of reference screenshots/images from vlogs found on DeafVideo.TV and YouTube.com. The images used to create scores for the three factors were screenshots taken from amateur vlogs found on DeafVideo.tv.

3.1.5.1 Lighting Assessment

Lighting levels were scored in a range from 1-5 where 1 was “too dark” and 5 was “too bright.” To create scoring values for lighting, three sets of 25 images were used. In set 1, each image contained an individual whose face was not well lit and it appeared that the frame had been taken from a video that was recorded in an environment with moderately dark lighting (images were underexposed). Images from set 2 contained individuals whose facial region was well lit and it appeared as if the signer had recorded their vlog with a sufficient amount of lighting (facial expressions could be seen clearly). Images from set 3 were overexposed and the facial region of the signer was considered moderately bright.

The facial regions in the images were then segmented and the pixel intensity average of each image segment was calculated. The values obtained from each image calculation were then used to determine the luminance threshold values to categorize future images, analysis scores and feedback (see Table 2,

Table 3 and Table 4). The values obtained in each set were compared with values obtained from the other sets related sets to determine thresholds. For example, the values calculated from images with good lighting were compared with the values of the sets from moderately bright lighting and moderately dark lighting. As an example, the highest value in the good lighting set was averaged with the lowest value in the moderately bright set to create the threshold between these two categories.

Table 2: Lighting Ranges with associated feedback for fair skin

Result	Analysis Score	Luminance intensity thresholds (0 – 255)
Too Dark	1	< 64
Moderately dark	2	>= 64 & < 110
Good Lighting	3	>= 110 & < 146
Moderately bright	4	>= 146 & < 192
Too Bright	5	>= 192

Table 3: Lighting Ranges with associated feedback for tanned skin

Result	Analysis Score	Luminance intensity thresholds (0 – 255)
Too Dark	1	< 50
Moderately dark	2	>= 50 & < 78
Good Lighting	3	>= 78 & < 136
Moderately bright	4	>= 136 & < 192
Too Bright	5	>= 192

Table 4: Lighting Ranges with associated feedback for dark skin

Result	Analysis Score	Luminance intensity thresholds (0 – 255)
Too Dark	1	< 45
Moderately dark	2	>= 45 & < 72
Good Lighting	3	>= 72 & < 140
Moderately bright	4	>= 140 & < 192
Too Bright	5	>= 192

The processes above to determine the lighting thresholds were performed three separate times for three different skin types: fair skin, tanned skin and dark skin. Because the assessment of lighting on a fair skinned facial region would be different than the assessment of lighting on a dark skinned facial region for example, three different models to classify lighting were created. During the study however, only the fair skinned model was used as no participants with tanned or dark skin applied to do the study.

3.1.5.2 Signing Space Assessment

Signing space was assessed using a combination of the two values discussed in Section 3.1.3.3: (1) the approximate distance of the signer from the recording device and (2) the vertical positional information of the signer. The values generated from these two measures are combined to classify a video frame into one of three signing spaces: small signing space, medium signing space, normal signing space. The threshold values for classification were determined using three sets of 25 images each. Set 1 contained images of various signers using a normal signing space (head and torso visible).

Set 2 contained images of signers using a medium signing space (only the head and upper torso are visible). Finally, set 3 contained images of signers using a small signing space (only the face and at most the upper shoulders are visible). Equation (3) and equation (4) were then run with each image of each set and the results were used to generate the threshold values for each classification.

Table 5 displays the threshold values created to score the combination of outputs of equation (3) and equation (4). The 'Width Value' column in the table represents the threshold values from equation (3) (approx. distance of signer from recording device) and the 'Height Value' column displays threshold values for equation (4) (vertical position of signer).

Table 5: Signing space threshold ranges

Result	Analysis Score	Width Value (%)	Height Value (%)
Normal Signing Space	3	<= 16%	<= 50%
Normal Signing Space	3	<=23%	<=30%
Medium Signing Space	2	<=13	> 65% & <= 80%
Medium Signing Space	2	<=23%	> 51% & <= 65%
Small Signing Space	1	> 23%	<= 65%
Small Signing Space	1	else	

In Table 5, there are two different scenario combinations that can result in the same classification (small, medium and normal signing space). For example, the two scenarios that result in a normal signing space are: (1) if a signer is far enough away from the recording device that it captures the whole or most of the signer (see Figure 22a) or (2) if the signer's head is high enough in the frame that the camera is able to capture as low as their navel and their head is not being cut off by the top of the frame (see Figure 22b).

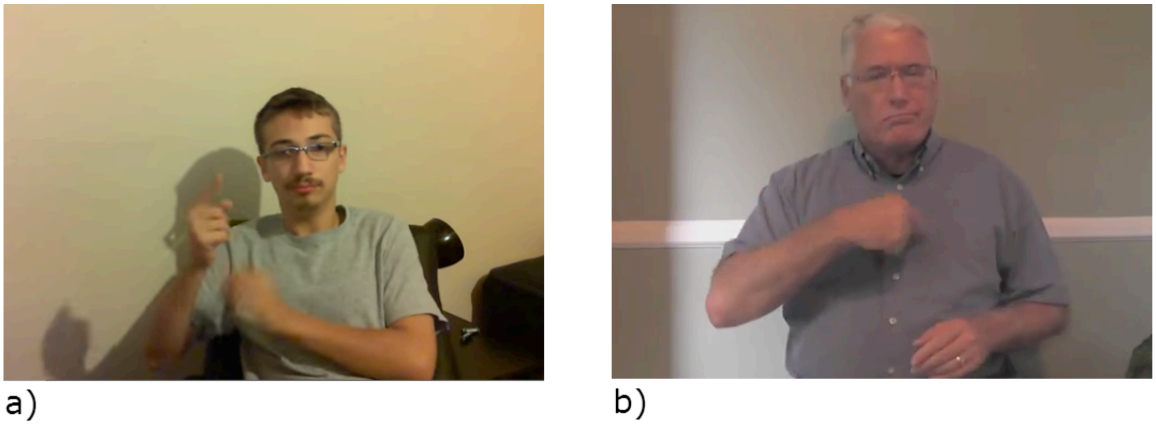


Figure 22: a) far distance from camera b) signer position high in frame

3.1.5.3 Clutter Assessment

Clutter in the VAST system is assessed using the Viola-Jones recognition algorithm combined with an edge detection technique to identify the number of objects in the background of the vlog. My research assumes a relationship between the number of edge pixels found and the number of items in the shot. The system was created to categorize an image into one of three different categories of background clutter: (1) no clutter; (2) medium clutter; and (3) high clutter.

If the signer uses a monochrome backdrop (e.g., wall, curtain) for their vlog and the facial recognition technique accurately detects the face and body, very few edge pixels (under 1% typically) should be found by the system and the image will be scored a 1 or having 'no clutter'. Objects in the background, such as a wall edge or doorframe, will add to the number of edge pixels found by the system but do not contribute to the visual complexity of the background. The 'medium clutter' classification is used to classify images where a monochrome backdrop is not used but there are not many objects in the vlog that could potentially distract the viewer from important content. In this scenario, the system will typically find about 1% - 6% of edge pixels in the frame. Finally, the system will classify an image as 'high clutter' if it finds many edge pixels due to an abundance of objects being present (typically greater than 6%). The thresholds used in determining the amount of clutter are displayed in Table 6. The ranges in the 'Edge Pixels' column are used to score the resulting values of Equation (6).

Table 6: Suggestion output based on percentage of edge pixels versus frame pixels

Result	Analysis Score	Edge Pixels
No Clutter	1	<= 1%
Medium Clutter	2	> 1% & <= 6%
High Clutter	3	> 6%

To determine the threshold values for the edge pixel classifications, three sets of 25 images were used. These images were screenshots from videos created by Signed Language vloggers. Set 1 contained 25 images of a signer communicating in front of a solid monochrome background (e.g., wall, curtains). Set 2 contained images of a signer with few objects behind him/her and the third set contained a signer with many items (e.g., books, boxes, logos and posters) behind him/her while communicating. Values from Equation (5) (usable frame pixels) and Equation (6) (total percentage of edge pixels) were calculated for each image in each set and the results were used to generate the threshold values for each score.

3.1.5.4 Final Score

The final recommendation for each technical factor is then determined by taking the statistical mode of the arrays generated for the assessed vlog. The statistical mode is used to determine the value that appears most often in the given array. The mode is used because the values that are obtained each time the system analyzes a frame are independent of each other.

The value generated by the mode in each array is passed to three individual functions which will then return a string with the feedback and recommendation for the particular factor and will set either a checkmark image or an x image beside each recommendation based on whether the factor setting is the optimal one or not.

3.1.6 Implementation Details

3.1.6.1 MATLAB

The VAST software was constructed with a functional programming language called MATLAB (version R2013b 32-bit). MATLAB is a mathematical programming environment used for the processing of mathematical functions and complex equations. Many external libraries have been built for MATLAB that allow for external functionality beyond what MATLAB offers. The built-in functions and external libraries fit the needs of my research. MATLAB also contained a GUI Builder (GUIDE) that allowed for the integration of GUI elements and multimedia objects such as images and videos.

The Image Processing Toolbox is an external MATLAB library that is used to perform many of the functions found in VAST. The toolbox provides access to image processing algorithms, such as image noise reduction and feature detection. VAST uses this Toolbox specifically for image segmentation, determining image intensities, face and body detection and edge detection.

4 Evaluation

This chapter provides the results for and analysis of the user study of VAST. Statistically significant ($p<0.05$) results are presented first followed by descriptive statistics and examples. The results are discussed with respect to the research question posed in Section 1.1. The total number of participants (N) for each statistical test is 15.

4.1 User Study

A user study was developed to evaluate the VAST system for usability, factors chosen for video quality assessment and to determine what impact the feedback had on user's reactions to video quality. In this study, participants were asked to create five vlogs and analyze them with the VAST system.

The study was held in a room where the state of each quality factor could be controlled and adjusted. When the user enters the room the cameras were already set up and markers were placed on the floor showing the participant where they were required to stand during the study.

Fifteen participants were used in the VAST study. Based on previous research done with mathematical models this number is more than enough to evaluate the usability of the system [92]. Nielson [92] researched the cost benefit trade-off for the amount of users needed to identify usability problems in small, medium and large systems. Nielson [92] found that four users were needed to identify 75% of usability problems for a small-scale system; size of the interface, number of expected users and severity of error determined scale size. Many more users would be needed to identify the remaining 25% percent of usability issues.

4.1.1 **Study Procedure**

The user study is approved by the Ryerson Ethics board (see Appendix **A** for approval letter). When the participant first arrives for the study he/she is asked to sign a consent form to participate in the study. After signing the consent form, the study video camera is turned on and the participant is asked to complete a pre-study questionnaire that collects demographic data such as gender, age and educational information, as well as vlogging viewing and creation habits (see Appendix **5.2.3** for pre-study questionnaire).

Upon completing the pre-study questionnaire, the participant completes five separate trials. In each trial participants are asked to answer a specific question (see Section 4.1.1.1) in ASL while being videotaped by a second video camera. Before each question is asked, the settings of the technical factors are changed to create various scenarios (e.g., different backgrounds, lighting conditions and viewing area of camera) to evaluate. There are seven different quality settings that are manipulated in the study: two background settings (cluttered, no clutter), two signing space settings (normal, small) and three lighting settings (bright, normal, dark). A "cluttered" background consists of a background containing items such as boxes, books and equipment; an uncluttered background uses a monochromatic (blue screen) wall. For the "small" signing space the field of view of the camera only recorded the signer's face and upper torso (close-up shot); for the recommended normal signing space the recording device captured just below the waist of the signer and just above the signer's head with enough space in the horizontal plane so that the camera could capture signs that required the signer's arms to be extended outward. For the "bright" lighting setting, the average illuminance on the facial region of the signer was above 800 lux, between 300 and 400 lux for the "normal" setting and approximately 10 - 50 lux for the "dark" setting.

Five combinations of factor scenarios were chosen for evaluation. The settings for four of the five trials were chosen to emphasize one technical factor (i.e., lighting, background, signing space) and in the final trial the user was asked to manipulate the technical factor settings to create "a good quality vlog." The user was given the opportunity to manipulate the factor settings themselves so we could view what factors the users would take the time to fix and which they would not. In the first four trials the setting of the factor being emphasized was set to a non-optimal level (dark/bright lighting, cluttered background and small signing space) while the other two factor's settings were adjusted to good levels (e.g. good lighting, normal signing space, uncluttered background). The order of the trials was randomized to minimize learning effects. A sample set of five trials for a study is illustrated in Table 7.

Table 7: Sample scenarios for a study. The settings that are not ideal in each trial are bolded.

	Lighting	Clutter	Sign Space
Trial #	3 possible conditions	2 possible conditions	2 possible conditions
1.	Good lighting (approx. 450 lux)	Monochrome background	**Small Sign Space**
2.	Good lighting (approx. 450 lux)	**High Clutter**	Normal Sign Space (good)
3.	**Moderately Dark** **(Approx. 15-25 lux)**	No Clutter (good)	Normal Sign Space (good)
4.	**Moderately Bright** **(Approx. 1000-1200 lux)**	No Clutter (good)	Normal Sign Space (good)
5.	**Moderately Dark** **(Approx. 15-25 lux)**	**High Clutter**	**Small Sign Space**

After a participant finished recording each vlog the video recording is uploaded to a desktop computer and the participant is asked to open the video recording with the VAST system and use the system to analyze it for technical quality. Next, the participant is asked to provide their opinion via a seven-question questionnaire on whether the feedback and recommendations generated by the VAST system is helpful in determining the overall technical quality of the video and if they would use this video or re-shoot it based on a given task (e.g. share it with friends, submit it to a professor for evaluation. Appendix 5.2.4 contains the vlog assessment questionnaire.

Once all of the trials are complete the participant is asked to complete a post-study questionnaire that collects summary opinions on the ease of use of the VAST system, the overall helpfulness of the advice provided and whether the technical factors are important for assessing the quality of a Signed Language vlog (see Appendix 5.2.5 for the post-study questionnaire).

To ensure that there were three unique lighting scenarios (one good and two sub-optimal), a light meter was used to measure an approximate amount of incident light in lux that was falling on the participant's facial region during each lighting scenario. Measuring the lighting levels was done prior to the studies taking place and the exact lighting scenarios were mimicked during the studies. Between 350 and 750 lux is sufficient for a working environment where reading and other visual tasks take place as recommended by Occupational Health and Safety Guidelines [93] and the Engineering Toolbox [94] For lighting levels in working areas that do not require visual tasks for long periods of time, 150 lux and below is satisfactory. Finally, lighting levels above 1000 lux are typically reserved for tasks that do not require high contrast. In my study, the moderately dark lighting scenario was measured at approximately 14 – 25 lux, the good lighting scenario measured 450 lux, and the bright scenario had a lighting level of 1000 – 1200 lux.

4.1.1.1 Questions Participants were asked to sign

The questions asked for each trial are the following:

1. What is the name of the school you attend, in what city is it located and which program are you currently enrolled in?
2. Why did you choose to enroll in the program that you did?
3. Tell me about one thing you enjoy about school and why?
4. Tell me about one thing you dislike about school and why?
5. If you could choose any other profession, what would it be and why.

The order of the questions asked to the participants for each vlog was never changed because the response to the question nor the content of the answer was important to the study. These questions were chosen solely because we anticipated that the response would be simple enough that everyone could answer and because we anticipated a response to take about approx. 15 – 20 seconds. Asking simple questions ensured that the participants did not have to think too much about how to sign their response and in turn ensured that the study did not go over the one-hour limit.

4.1.2 **Participants**

The pre-study questionnaire was used to gather demographic data such as participant age, gender and year of study. The questionnaire was also used to determine how often participants use vlogging technology to create, view and share content online and what scenarios they thought were ideal for vlog creation.

Fifteen people (14 female, 1 male) participated in the VAST evaluation study to determine the usability and usefulness of VAST and the three factors used for vlog assessment. The age range of the participants was 18 – 44. Each participant was enrolled in an ASL-English interpreter-training program in Canada and was between the first and third years of study.

All participants reported that they used a computer every day for various purposes. Two of the fifteen participants stated that they created vlogs every day; five participants reported that they create vlogs every 2-3 days; four participants said that they create vlogs once a week; 3 participants stated that they created vlogs only once a month; one participant said that they had never created vlogs. Of the 14 participants that said that they created vlogs periodically, thirteen said that they typically created vlogs for coursework and one participant reported that they created vlogs for personal reasons.

The pre-study questionnaire also asked participants to specify how often they view Signed Language vlogs created by other people. Five participants said that they viewed signed content every day; six stated that they watched vlogs every 2-3 days; one person said that they watch vlog content once a week and only one person stated that they watch vlog content created by someone else once a month.

The final question in the pre-study questionnaire asked the participants to rate their level of agreement with statements pertaining to vlog quality and scenarios that may affect the quality of a video. Participants rated their agreement on a scale from "1 Strongly Disagree" to "5 Strongly Agree". **Table 8** provides a summary of the responses.

Table 8: Mean and std. deviation of user responses on vlog creation

Statement	Mean	Std. Dev
Direct sunlight is the best lighting for vlog creation	2.27	0.59
The lighting in this room is not ideal for vlog creation	1.87	0.99
Mood lighting should be used when creating vlog content	2.07	0.88
A solid colour painted wall with no items on it is an ideal background for vlog creation	4.40	1.06
An office environment would not be an ideal background to use when creating vlog content	3.76	0.98
Producing vlog content in front a window with incoming light is best for vlog creation	1.73	0.96
Adjusting the camera so that your face and a small space around your head is in the frame is best for vlog creation	2.33	1.29
Adjusting the camera so that it is able to view your body but not your face is best for vlog creation	1.27	0.46
Adjusting the camera so that your head and body are in the video frame is not ideal for vlog creation	1.20	0.41

4.1.3 **Analysis of Data**

The trial question data was analyzed using repeated measures ANOVA in order to determine whether there were significant differences between the various factor settings (dim/ good/ bright lighting, small/ normal signing space, no clutter/cluttered background). Paired t-tests were used post hoc to find evaluate and differences in the pairs that arose from the repeated measures analysis.

For the post-study questionnaire results chi-square goodness-of-fit tests were used to compare the results of the Likert scale questions to chance. Since there were only 15 participants used in the study and all Likert scale questions were asked on a 5 point scale, the Likert scales were compressed from five-point scales to three-point scales due to the lack of data needed for this test as recommended by [95].

4.1.4 Hardware

4.1.4.1 Recording

A Sony 12 mega pixel video camera is used in the study to record the user created vlogs. The videos are recorded in standard definition at a 4:3 aspect ratio and no special features, such as low lux lighting correction, were used. The camera was set up at waist level from the signer and the zoom feature on the camera was used to recreate the large and small signing spaces.

4.1.4.2 PC

The specifications of the computer used for the study are outlined in Table 9:

Table 9: Specifications of the computer used for the study

Component	Specific
Operating System	Windows 7 Enterprise (64 bit)
Graphics Card	Nvidia GeForce GTX 660Ti
Central Processing Unit	Intel(R) Core(TM)2 Quad CPU Q6600 @ 2.40Ghz
RAM	16 GB RAM

4.2 Results

4.2.1 Between Study Questionnaire Results

The between study portion of the user study consisted of individuals creating five different vlogs. Each time a vlog was created it was uploaded to a computer and analyzed with VAST. Each vlog was recorded with varying quality (i.e., bright/dark/optimal lighting, cluttered/not cluttered, background and normal small signing space). The following paragraphs outline the tests used to analyze the data collected from the questionnaires participants filled out when they read each recommendation.

A repeated measures ANOVA and paired-samples t-tests were used to analyze the between study results. A repeated measures ANOVA allows the researcher to determine if there are statistically significant differences between related means and are used in tests where the same individuals are tested for all conditions of the study. One benefit of this test is that it requires less participants to run [96]. In the case of this research, I used this measure to determine if there existed any statistical significance in the means between the factor settings (bright/dark/optimal lighting, cluttered/not clutter background, normal small signing space) based on a proposed question.

An assumption of the repeated measures ANOVA is known as sphericity. Sphericity is the assumption that the variances of the differences between condition pairs are similar [96]. In other words, sphericity tests to see if the relationships between the pairs of conditions are the same. SPSS (IBM's statistical analysis software [97]), the software I use for the analysis of results, uses Mauchly's test of sphericity to test this assumption when running repeated measures ANOVA statistical tests. Violation of sphericity may result in the detection of statistical significance when in fact there is not [98]. If in fact this assumption is violated, Field [96] recommends using a Huynh-Feldt correction to obtain a proper F-ratio (the F-ratio can tell us if two variances are equal). The correction is applied to the degrees of freedom in the F-ratio. Furthermore, when sphericity is violated, the Huynh-Felt correction will raise the degrees of freedom in the F-ratio and from this it can tell us if there is statistical significance or not between the variances of pairs.

The paired-samples t-test is used to determine if the mean difference between two conditions are significantly greater than 0 [96]. This test is not run in this research if the repeated-measures ANOVA tells us that there is no significance between factor pairs. This test applies to my study as it is typically run when the same individuals are tested on different conditions. I used this test because it could tell me between which pairs there was significance and between which pairs there was not. For example, this test could tell me if the mean difference for the bright lighting condition and the dark lighting condition was significant or not.

4.2.1.1 Helpfulness of Technical Factor Recommendations

A repeated measures ANOVA was run on the data to examine the differences in helpfulness of the VAST recommendations based on the 3 technical factors and their settings. The dependent variable for this measure is helpfulness. According to Mauchly's test of sphericity, the assumption of sphericity was violated ($\chi2(20) = 34.69$, $p = .026$) and so a Huynh-Feldt correction (epsilon value = 0.833) was used. Statistical significance was found in the helpfulness of the recommendations [$F(5.0, 69.94) = 2.70$, $p = 0.03$].

Paired t-tests were then carried out to compare the differences in helpfulness between the technical factor settings pairs. Results from the t-tests showed significance between the signing space and background ($p = 0.28$). None of the other pairs was significant. Participants stated that receiving good feedback on a normal signing space setting (M=3.40, SD=1.06) was significantly less useful than receiving feedback and a recommendation for improvement on a video with a cluttered background (M=4.47, SD=0.64). The descriptive statistics are provided in Table 10.

Table 10: Descriptive statistics for helpfulness of each quality factor. This data was collected after participants created a vlog and received feedback from the system based on the quality of their video. Participants were asked to rate the helpfulness of each recommendation on a 5-point Likert scale.

Factor # - Factor Setting	Mean Helpfulness	Std. Deviation
1- Light Dark	4.20	0.68
2- Light Bright	4.13	0.74
3- Light Good	4.00	1.07
4- Background Cluttered	4.47	0.64
5- Background Good	4.07	0.70
6- Space Small	4.20	0.86
7- Space Normal	3.40	1.06

4.2.1.2 Likelihood of Watching

A repeated measures ANOVA was carried out on the likelihood of a participant watching another vlog if it had similar visual quality as the vlog created during that trial. Mauchly's test of sphericity indicated that the assumption of sphericity was not violated ($p > 0.05$) and so no correction was used. Statistical significance was found ($p < 0.05$) in the likelihood of re-watching another vlog based on if it had the same quality as the vlog created in the trials [$F(3.14, 56.00) = 21.59$, $p = 0.00$].

Paired t-tests were then run between the factor settings pairs. Statistical significance was found between the background and signing space ($p = 0.018$). Participants indicated that they were more likely to watch another vlog that had clutter compared to a vlog where the signer used a small signing space. The descriptive statistics are provided in Table 11.

Table 11: Descriptive statistics for likelihood of each quality factor

Factor # - Factor Setting	Mean Likelihood	Std. Deviation
1- Light Dark	2.47	1.46
2- Light Bright	2.73	1.10
3- Background Cluttered	3.27	1.16
4- Signing Space Small	1.93	1.16
5- Quality Good	5.00	0.00

4.2.1.3 Likelihood of Submitting to Professor

To examine the likelihood of re-creating a video before submitting it to a professor for evaluation based on the quality, a repeated measures ANOVA was used. According to Mauchly's test of sphericity, the assumption of sphericity was violated ($p < 0.05$) and therefore a Huynh-Feldt correction (epsilon value = 0.785) was used. Statistical significance was found in the likelihood of re-submission [$F(3.14, 43.95) = 21.60$, $p = 0.00$].

Paired t-tests were then run on the factor settings pairs to determine differences in the specific technical factors. Significant differences were found between the non-ideal factor settings (factors 1-4) and the good quality setting (factor 5) (See Table 12). Participants stated that they were likely to re-record a vlog before submitting it to a professor if it contained any issues with quality. The descriptive statistics are provided in Table 12.

Table 12: Descriptive statistics for likelihood of submitting vlog to professor

Factor # - Factor Setting	Mean Likelihood	Std. Deviation
1-Light Dark	4.40	1.24
2- Light Bright	4.40	0.99
3-Background Cluttered	4.33	1.05
4- Signing Space Small	4.67	0.82
5- Quality Good	1.47	1.25

4.2.1.4 Likelihood of Submitting to Video Sharing Website

A repeated measures ANOVA was run on the data to determine the likelihood of a participant re-recording a vlog before submitting it to a popular video sharing website because of the quality of that vlog. Mauchly's test of sphericity indicated that the assumption of sphericity had not been violated (p

> 0.05) and no correction was used. Statistical significance was found in the likelihood of re-submitting a vlog based on quality [F(3.10, 43.44) = 16.55, p = 0.000].

Paired t-tests showed that there were differences between each non-optimal setting (factors 1-4) and the "quality good" setting (factor 5). Participants were likely to re-record a vlog to submit to a popular sharing website if the quality of that vlog was not optimal. The descriptive statistics from this test are provided in Table 13.

Table 13: Descriptive statistics for likelihood of submitting vlog to video sharing website

Factor # - Factor Setting	Mean Likelihood	Std. Deviation
1- Light Dark	4.07	1.03
2- Light Bright	3.73	1.36
3-Background Cluttered	4.40	0.74
4- Signing Space Small	4.40	0.97
5- Quality Good	1.80	1.66

4.2.1.5 Comparing Submission Types

Paired t-tests were used to determine if participants were more likely to submit their vlog to a video sharing website rather than to a professor or vice versa based on a setting of one of the three technical factors. For example if the vlog contained a cluttered background, was the participant still likely to upload the video to YouTube.com but not submit it to their professor for evaluation.

The t-tests results showed that there was a significant difference in likelihood of submitting a vlog when it experienced a fault with lighting. Participants were more likely to re-record their vlog before submitting it to a professor (M = 4.40, SD = 1.24) than online to a public video sharing website (M = 4.07, SD = 1.03) if the vlog was not recorded with sufficient lighting. No other significant values were found when comparing factor settings pairs. Descriptive statistics, which display values from Table 12 and Table 13, are provided in Table 14.

Table 14: Descriptive statistics for comparison of sharing

Pair #	Factor Setting	Type	Mean	Std. Deviation
1	Lighting dark	Professor	4.40	1.24
		Website	4.07	1.03
2	Lighting bright	Professor	4.40	0.99
		Website	3.73	1.34
3	Background cluttered	Professor	4.33	1.05
		Website	4.40	0.74

4	Signing space small	Professor	4.67	0.82
		Website	4.40	0.99
5	Good Quality	Professor	1.47	1.25
		Website	1.80	1.66

4.2.2 **Post-Study**

4.2.2.1 Levels of Difficulty

Participants were asked to rate the level difficulty of performing tasks using the VAST system from “1 – Very difficult” to “5 – Very Easy”. A goodness-of-fit chi-square test was carried out for the questions related to task difficulty to compare the participant responses with chance. Originally the question was asked on a 5-point Likert scale but due to the small sample size the values were compressed to a 3-point Likert scale [95] ranging from “1 - Difficult” to “3 - Easy.” The goodness-of-fit chi-square test could not be run with questions 1 and 3 because after compressing the data to a 3-point Likert scale all selections from participants became “3 - easy“. The goodness-of-fit chi-square test requires a minimum of two groups and since all answers were the same this test could not be run for these questions.

The chi-square test for difficulty of generating vlog feedback (question 2) was statistically significant ($\chi_2(2)=11.27$, p = 0.00); the observed and expected frequencies were not the same. More participants found that generating feedback was “easy” (14/15) compared to “Neutral” (1/15) and “Difficult” (0/15). The descriptive statistics are displayed in Table 15.

Table 15: Descriptive statistics for difficulty of tasks

Question #/ Task	Mean	SD	Chi-Square	P-Value
1 - Import Vlog into system	5.00	0.00	-	-
2 - Generate vlog feedback	4.80	0.56	11.27	0.00
3 - Read information provided by VAST	4.60	0.51	-	-

4.2.2.2 Conditions for watching a vlog

Participants were asked to rate their interest in watching eleven different vlogs with different quality settings such as a vlog with dim lighting or a cluttered background. Participants rated each statement from “1 - Would not watch vlog” to “5 - Would definitely watch vlog”. The mean ratings, standard deviations and the corresponding eleven statements are provided in

Table 16.

Table 16: Descriptive statistics for watching a vlog based on a filming scenario

Statement	Mean	Std. Dev.
1. The vlog has a blue coloured wall in the background.	4.87	0.52
2. The vlog is recorded where you can see the face and shoulders of the signer.	3.20	1.42
3. It is recorded with optimal lighting levels	4.87	0.35
4. There is a lot of clutter in background of the signer	2.53	1.19
5. The light in the vlog is too bright	2.13	1.25
6. The vlog is recorded where all you can see is the signers face	1.53	1.06
7. It is recorded with dim lighting	2.20	1.21
8. The vlog is recorded where you can see the face and body of the signer	4.93	0.26
9. It is recorded with bright lighting	3.27	1.49
10. No clutter is found in background of signer	5.00	0.00
11. It is recorded with dark lighting	2.07	1.28

4.2.2.3 Important Quality factors

A frequency analysis was used to determine which quality factors the participants thought were the most important when deciding quality. The results are shown in Figure 23.

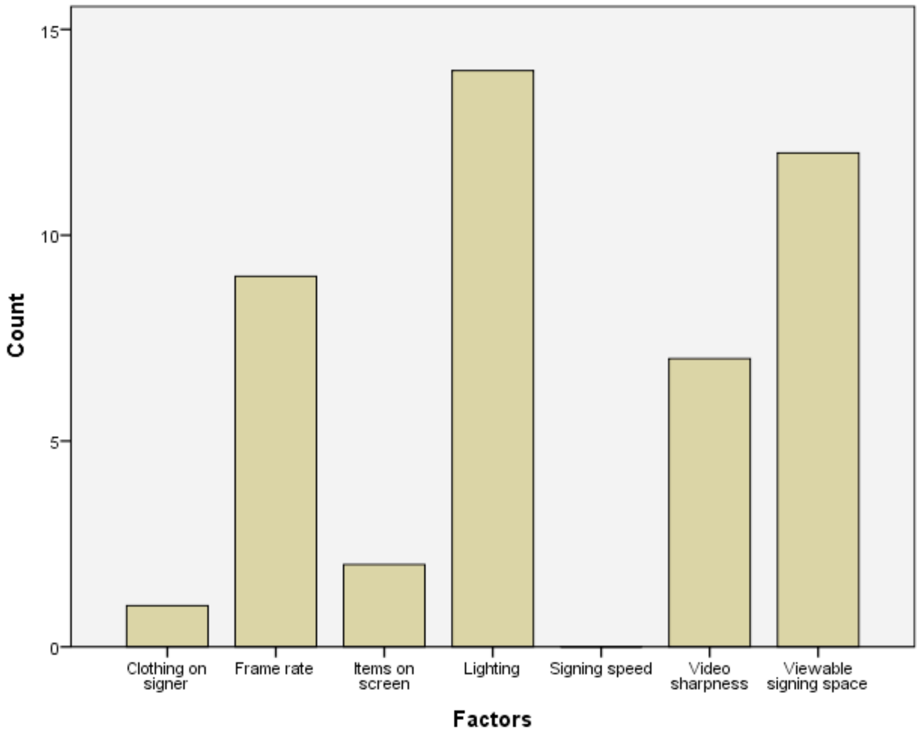


Figure 23: Frequency chart for important factors when choosing quality

4.2.2.4 Effects System had on Participant

Participants were asked to read eight sentences about the possible effects of using the VAST system and indicate which sentences were agreeable. Participants were not limited to how many sentences they could check off and the order of the sentences was randomized so that there was no evident pattern between negative and positive effects. A frequency analysis was used to view how many participants agreed on each effect.

Table 17 displays the descriptive statistics for number of participants who agreed with each statement and Figure 24 displays these results in a bar graph.

Table 17: Descriptive statistics for the list of possible effects of VAST

Effect	Number of respondents in agreement
1. It would help me to improve the quality of my future vlogs	9
2. I learned about making vlogs with good quality	7
3. It helped me to understand the impact of certain technical factors on viewers	2
4. It made me pay attention to vlog quality	8
5. It took too much time	0
6. I was bored	0
7. There was nothing new	1
8. I was confused by the recommendations	0

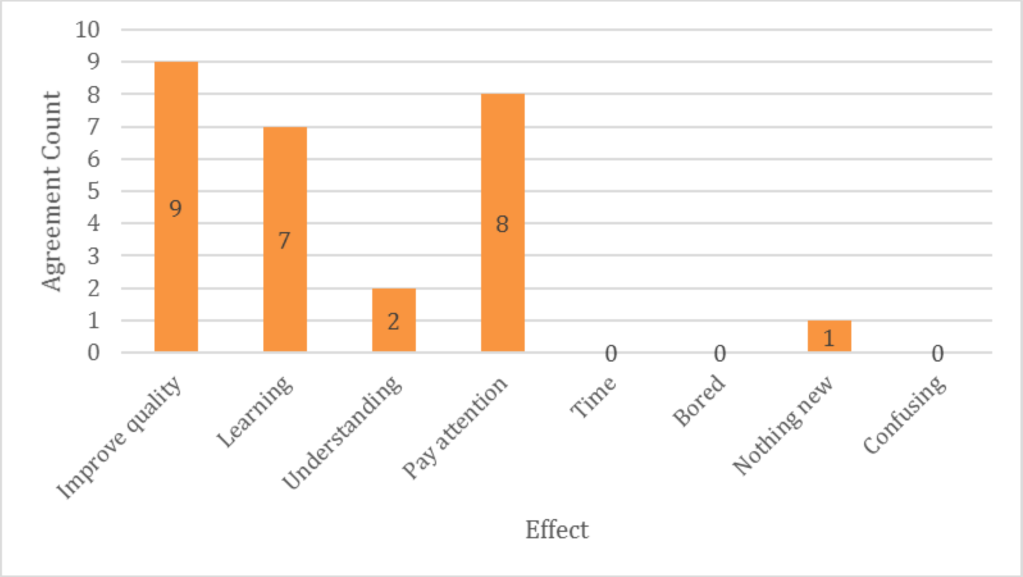


Figure 24: Frequency chart for effects of VAST

4.2.2.5 Open Ended Questions

To analyze the statements made by participants in the open-ended questions of the post-study questionnaire, a thematic analysis [99] was performed on the answers to the following questions:

1. Rate the level of difficulty of reading the information provided by the VAST system? Why? (Likert scale question and explanation)

2. What did you like most about the VAST system?

3. What did you dislike most about the VAST system?

The themes developed for the questions and their associated definitions are shown in Table 18. Comments made by participants were categorized into the themes by two independent raters. An Intra-Class Correlation (ICC) coefficient was then generated for each question. All generated ICC values were greater than 0.820. This indicates that the raters were in strong agreement. The number of occurrences for each theme is displayed in

Table 19.

Table 18: Themes and definitions used for thematic analysis

Theme	Definition / Examples
Processing speed	Recommendations from VAST were generated in a short amount of time *"What I liked most about the VAST system was how quick it was to critique the quality of the vlog."*
Ease of use	User friendly, easy to use; tasks completed without confusion; clean/ simple design; icons help to determine feedback quality *"the system is extremely user friendly and easy to understand."*
Feedback information	Easy to understand; enough written to understand problem; wording clear / concise *"The information was very clear and concise. Simple to understand."* *"Sometimes language can be difficult to understand"*

Table 19: Number of occurrences for each theme for positive and negative occurrences

Theme	Number of Positive Occurrences	Number of Negative Occurrences	Total Comments
Processing speed	4	0	4
Ease of use	15	4	19
Feedback information	10	7	17

4.3 Discussion

The following section is a discussion of the study results in relation to the research questions. Questions 2 – 4 will be addressed in this section. The research questions are restated here as a reminder to the reader.

3. How can the detection of the technical factors of signing space, lighting and visual clutter be used for assessing video quality of vlogs?

4. What is the usability of the VAST system?

5. What is the level of impact of each technical factor on user perception of video quality?

6. How helpful is the feedback of lighting, signing space and background clutter?

4.3.1 Research Question 2: Impact of feedback

This section aims to answer the second research question: What is the impact of feedback of lighting, signing space and background/foreground clutter on quality and re-recordings of user vlogs? The discussion in this section will deliver my claim for the usefulness of the feedback given by VAST for the improvement of quality of user created vlogs.

4.3.1.1 Factor helpfulness

Examining the results from the between study questionnaires on helpfulness of feedback for each factor, it seems that participants found the feedback to be helpful when determining vlog quality. Participants found that a recommendation from the system to improve the background when it was cluttered was most helpful (M = 4.47, SD = 0.640 where 5 or very helpful was the maximum score on the Likert scale). The least helpful recommendation was when VAST correctly reported that the signer was using a normal signing space (M = 3.4, SD = 1.056).

Participants overall found the system to be most helpful when it reported that there was a quality issue with the input vlog. The mean helpfulness score when the system reported a fault with quality was 4.25 compared to 3.82 when the system did not find any faults with the quality of the input vlog. A reason for this finding could be that participants may have found the system less useful when it told them their video quality was good as they could have avoided using the system all together. For example, when the system reported that the background did not need to be fixed one participant commented saying, "*The background was already clutter-free. I could tell by seeing it but recommendations helps to verify*".

It was originally hypothesized that the helpfulness scores would be highest when the visual quality of the vlog was poor or difficult to determine. It would seem that this hypothesis was born out by the results and that having feedback about the background clutter was most helpful. This result could be related to which factors are easier to notice or are most obvious to determine. Problems with lighting and signing space may be more noticeable than background clutter as these factors directly affect how much of the signer is visible. Background and foreground clutter may not directly influence the visibility of a signer and therefore the factor may be less noticeable and more difficult to determine an optimal setting for it.

Another reason why the feedback on background clutter may have been useful to participants is because there is so many different backgrounds a signer can communicate in front of and it may be more difficult to know which backgrounds are acceptable and which are not. Participants may have found VAST useful in assisting them in confirming the quality of this factor. As one participant stated, "*It is sometimes hard to determine which backgrounds are good because you see so many different ones".*

4.3.1.2 Vlog re-creation

Participants were asked if they were likely to re-create their vlog before submitting it to either a professor for evaluation or a popular video sharing website based on the feedback given from VAST. These questions were intended to answer the question of whether the recommendations supplied by the system would influence the participant's decision to create new content based on the quality of their old content.

Participants rated the likelihood of recreating their vlog on a scale from "1 Not at all likely" to "5 Very likely." Results indicate that, in both scenarios, participants would be most likely to recreate their vlog if the system found that they were not using a normal signing space. The mean likelihood of a participant to re-create a vlog was 4.67 (SD = 0.82) for submission to a professor and 4.4 (SD = 0.99) for submission to a video sharing website. A participant was the least likely to re-create a vlog for submission to a professor if the background was cluttered (M = 4.33, SD = 1.05) and the least likely to change their vlog if there was too much lighting (M = 3.73, SD = 1.335) when submitting to a popular video sharing website.

Only when a video contained bright lighting and was being submitted to a popular video sharing website were participants not more than likely (over 4.0 "likely to re-create") to re-create their vlog based on the recommendation to do so. This may indicate that participants have a high regard for the feedback VAST delivers and trust the system enough to re-create their vlog if it so tells them to do it. Comments featured below support this claim.

"to get that framing JUST right is a challenge. So, to have the system tell you that you did go out of your frame instead of having to watch your entire video can be very helpful."

"I would not have thought about the shadows in the background. This was helpful. I would re-create because of this"

The results above are consistent with the results from other sections in that signing space was found to have the most impact on a user's perception of quality. There was little difference in the mean likelihood scores of vlog re-creation if the signer did not record with a normal signing space between the two submission scenarios. This could indicate that regardless of where the vlog is being submitted, signing space has the highest impact on a user's perception of video quality.

4.3.1.3 Reading the feedback

In the post study questionnaire I asked the participants to rate the difficulty of reading the information provided by the VAST system. Users were also asked to explain the reasoning for their selection in a one or more short sentences. Thematic analysis was used to analyze the textual responses of the participants. The overall participant response to feedback information was relatively positive. A total of 17 occurrences were noted for the *feedback information* theme; ten positive occurrences and seven negative occurrences. Of the ten positive occurrences, seven participants noted that the wording was clear and concise and easy to understand. Another three participants said that there was enough written for each recommendation to understand the feedback given for each factor. Two participants thought that the wording was unclear and could be changed to provide a better understanding of the vlog issues. Five participants stated that more information could have been provided with the recommendations so that the problems could be better understood. Example comments for feedback information are displayed below:

"The information was very clear and concise. Simple to understand."

"Everything was straightforward … All of the analysis blurbs were to the point with no fluff."

"I think a more in depth recommendation would have been more helpful."

> *“They are very easy to understand, but I would appreciate a bit more information about what the problem is."*

Only three occurrences were noted where participants said that the system helped them understand the impact certain technical factors had on overall quality. Although this result is not a favourable one it is not surprising as the system did not provide justification for each assessment. The goal of the system was to provide three short and concise sentences as a summary to the quality of the vlog based on the three technical factors. The reason there was no further justification with the feedback information was because I did not want the users to be tasked with reading large amounts of text to determine the vlog quality. I also avoided creating large amounts of text due to the target audience (Signed Language users) and knowing that many individuals who use the system in the future may not be native English speakers. Information about the negative effects the three technical factors can have was provided in the “About Factors” section that was accessible through a menu item at the top of the screen. No users however navigated to this section of the system during the study nor were they instructed to do so. This section was created only as a means for further explanation if any participant asked about a specific factor.

4.3.2 Research Question 3: Impact of technical factors

The goal of this section is to discuss the research question: What is the level of impact of each technical factor on user perception of vlog quality? The discussion in this section will help to confirm my claim that the factors chosen for assessment are important to users and do have an impact on their perception of video quality. It will also identity which factor(s) have the most impact on user perception.

4.3.2.1 Likelihood of watching vlog

Participants were asked to rate the likelihood of watching a vlog if it had similar visual quality as the videos created during their trials. Ratings ranged from “1 - Not at all likely” to “5 - Very likely.” Participants indicated that they were unlikely to watch a vlog if they found the videos to have any issues with the lighting, background or signing space. Results show that participants were the least likely to watch a vlog containing a small signing space (M = 1.93, SD = 1.16). This result was followed by a vlog with moderately dark lighting (M = 2.47, SD = 1.46), a vlog containing moderately bright lighting (M = 2.73, SD = 1.10) and finally a vlog containing a cluttered background (M = 3.27, SD = 1.16). As expected, all participants indicated that they would watch a vlog if it had no flaws with the lighting, signing space and background (M = 5.00, SD = 0.00).

It seems that although participants found the system most helpful when it told them that the background in their vlog was cluttered, they were still more likely to watch a vlog with a cluttered background than a vlog with moderately dark/bright lighting or a small signing space. A possible reason for this outcome could be that signing with a cluttered backdrop has become a common practice among vloggers. In fact, vlogs are used as personal diary entries and so the various backdrops could be used to add personal expression to the vlog [24]. Signers may be interested to know how the system would assess the different backgrounds but it does not mean that clutter would deter them from watching another vlog more than lighting or singing space would. Hibbard and Fels [26] support the claim that although vloggers have a choice to create signed content in front of a monotone coloured background many choose not too. I suspect this result either because of personal expression or because signers do not know the possible effects of introducing a cluttered environment into Signed Language video; the detrimental effects of a small singing space or bad lighting may be more obvious to determine.

Participants commented on vlog backgrounds in the open ended questions in the between study questionnaires. Some of the comments support the claim that visually busy backgrounds are common to see and are more acceptable when compared to a vlog with quality issues which directly affects the signer's legibility. For example, participants said:

> *"It's not ideal but watching someone sign in front of a cluttered background is more common than you would think. In addition, between a cluttered background, dim or too bright lighting or too close/too far camera - cluttered background is the easiest to view and comprehend"*

> *"Most of us do not have a beautifully painted dark wall waiting for us in our home, so to know if our background is acceptable or too cluttered for our viewing audience is extremely important."*

It was originally hypothesized that lighting would influence a user's decision to watch a vlog more than clutter and signing space because of its known impact on reducing intelligibility [33]. However, it seems that not being able to view the areas around the face and mid-section of the signer (signing space) had a higher influence on a user's decision to watch the content than lighting did. The mean likelihood of a participant to re-watch a video if it did not have optimal lighting was 2.73 (SD = 0.28) for a video with high brightness and 2.47 (SD = 0.38) for a video with dim lighting. The mean likelihood of re-watching a vlog containing a small signing space had a mean score of 1.93 (SD = 0.30). A possible reason for this result could be that on many video displays, including laptop and computer monitors, the lighting levels can be adjusted to increase or decrease the brightness if needed. Participants may have thought that brightness issues could be resolved by making adjustments on their own display independent of the lighting levels in the actual video. This would not be the case for signing space, which is irreversible once the video has been recorded. Lighting may affect a user's comprehension but signs may be still readable even if the lighting is not good. If signs occur outside of the viewing frame of the recording device, the signs cannot be seen or understood regardless of any user adjustments.

4.3.2.2 Factors affecting "good" quality

In the post study questionnaire participants were asked to select their top three choices of technical factors they considered important when deciding vlog quality. Lighting was chosen most by participants (14/15), followed by viewable signing space (12/15) and frame rate (9/15). Signing speed (0/15) and the "other" choice (0/15) were not selected by any participant. Of the choices that were selected by at least one participant, clothing worn by the signer (1/15) and number of items on screen (2/15) were the least selected.

It was originally hypothesized that the background (number of items on screen) would be a top choice of participants for factors when considering quality. However, participants chose frame rate more often than they did background. Hooper [100] found that the comprehension scores of participants when re-telling a story increased as the frame rate increased. Participants scored significantly lower when they were shown a vlog at 6fps compared to 12fps and 18fps. Hooper also found that these results were based on the participant's signing experience. Signers ranked at an intermediate level scored better than signers who were ranked at a beginner level for each frame rate setting. Since the participants in the VAST study were mainly beginners, the results found by Hooper may explain why frame rate would be a more important factor when determining quality than the number of items on screen. In a future study, the opinions of participants on VAST and on the impact of the three technical factors on quality could be compared and contrasted between groups of participants with different experience levels. This would help to determine if a participant's perception of quality based on lighting, signing space and clutter changes because of experience level.

A second possibility as to why the *number of items on screen at one time* option may have been chosen by so few participants could be that the survey option was too vague and more explanation was needed. It was never explicitly stated to the participant that there was the number of items in the foreground/background of the vlog was synonymous with background clutter. The participants may have interpreted the survey option as something other than background. In future it would be best to correct this by explicitly stating this relationship to the participant or by simply keeping the wording consistent to avoid any confusion.

4.3.3 **Research Question 3: Usability of the system**

4.3.3.1 Usability

Participants were asked to rate the difficulty of performing the following tasks using the VAST system to determine the systems overall usability and difficulties participants experienced:

1. Import vlog into VAST

2. Generate vlog feedback and recommendations

3. Read information provided by VAST

It seems as though participants did not have a difficult time using the system. The mean difficulty scores for questions 1 – 3 were all above 4.0 (easy). Every participant rated importing a vlog into the system a 5.0 (very easy), followed by generating feedback (M = 4.8, SD = 0.56) and reading information provided by the system (M = 4.6, SD = 0.51). Participants also stated, in the open-ended questions, that the system did not have a steep learning curve and overall VAST was very easy to use. Sample comments from participants are seen below:

"It is very easy to use and it did not take a long time for me to learn how to use it."

"the system is extremely user friendly and easy to understand."

No participants commented that they had any difficulty using the system to open their created vlog file and analyze it for technical quality. These results may indicate that the usability of VAST is such that it is easy for users to accomplish tasks efficiently and without error. The ease of using the VAST system may be a result of the minimalistic design and low number of steps needed in order to load a video file into the system and generate feedback. Lidwell [101] states that, "Design should minimize performance load to the greatest degree possible. Reduce cognitive load by reducing unnecessary information from displays … reducing unnecessary steps in tasks". Lidwell [101] also writes that a design must contribute to the self-esteem of a user. With this in mind, I designed VAST so that Signed Language users did not feel like the design was a barrier because of the amount of text in the visual display. The amount of text was taken into account when designing the system and although Signed Language help videos were not included with the system during the study, a future consideration would be to include ASL videos that deliver the same information the textual feedback does.

Only 17% of comments were made about the interface of the VAST system; one positive occurrence and four negative occurrences. All but one comment made about the interface pertained to the icons used to help convey the textual recommendations. These icons were either an 'x' icon to indicate that there was a problem with the specific technical factor or a checkmark icon to indicate that there was no problem with the specific factor setting. One participant stated that the icons helped him/her determine the feedback quality of the factor settings. Three participants said that although the icons did help to determine quality, changing the colour of the checkmark icon would help the system better convey the outcome of the recommendation. No participants commented on the placement of any elements on the screen.

When asked what participants dis-liked most about the VAST system one participant said that they would have liked the system to provide a real time playback of their video. They also noted that this feature would have significantly helped them understand the VAST recommendations better as they could identify the problems VAST pointed out. This participant went on to say that having this feature would prevent them from having to open up another software to replay their video.

It seems that most people thought that using the VAST system was a mainly positive experience. Participants thought that VAST would help them most to improve the quality of future vlogs (9/15). The system also assisted participants in pay attention to the quality factors in their vlog (8/15) and taught them about making good quality vlogs (7/15). No participant thought that the system made them bored (0/15), were confused by the recommendations (0/15) or that the system took too much time to use (0/15). Only one (1/15) participant stated that the system did not offer anything new.

One concern I had when creating the system was that people would find it slow and time consuming as the analysis does not work in real-time during the video recording process. This concern did not materialize as an issue for participants; no participants said that the system was boring or took too much time to operate. The responses to open-ended questions on this topic also revealed that participants found the processing time of the system to be relatively fast. This result may have also contributed to the ease of use and simplicity of the system that was reported in other sections of this chapter. Sample comments pertaining to the speed of the system are shown below:

> *"It's fast. It literally takes seconds to process the video. This is great, if the software took a long time to process each video I likely wouldn't be patient enough to use it."*

> *"What I liked most about the VAST system was how quick it was to critique the quality of the vlog."*

The quick processing time is, in part, due to the fact that the videos analyzed in the study were relatively short in duration (under 30 seconds in length). Depending on the video size and the number of frames analyzed the processing time may be longer resulting in longer wait times.

4.4 Limitations

4.4.1 Limited participants

Originally I had intended to have hearing and deaf participants in the study to compare and contrast the opinions of the groups about the system. However, due to time limitations and scheduling difficulties with Signed Language interpreter students I only recruited interpreter students for the study. Having both groups would have provided more insight and a more robust result. In future, a similar study using VAST can be performed with deaf vloggers and the results of that study can be compared to results of this study to view the difference in opinions and preferences of either group.

In addition, I only recruited 15 participants, which limited the statistical analyses that could be carried out, and the generalizability of the results. Any conclusions arising from the data can only be tentative until further study with more participants is completed.

4.4.2 Limited demographic

Three different lighting models were created to assess the lighting levels assuming that there would be participants with different skin types. However, this was not the case as only participants with fair skin participated in the study. Therefore, I was not able to evaluate the lighting conditions for individuals with dark and tanned skin because of this limitation with participant demographics.

4.4.3 **Novelty effect**

The first study that was conducted for VAST was a one-time, short duration study. Participants did not work with VAST over multiple sessions so I was unable to either predict or measure the long-term acceptability or usefulness of VAST. The results reported in this thesis could therefore be a novelty effect rather than a true representation of the usefulness of VAST in helping people create better quality vlogs. However, comments from the open-ended questions did reveal that participants thought that they learned more about creating good quality vlogs during the study. A next phase in this research could then be to determine whether the use of VAST over time allows users to improve the quality of their vlogs. It would also be useful to determine whether using VAST had any impact on vloggers initial designs and preparation for their vlog entries to avoid having quality errors in the technical factors. Finally, investigating viewer's reactions to vloggers improvements could also aid in determine the impact of VAST on the quality of vlogs.

4.4.4 **Biometric Data**

Biometric data such as eye movement was not collected during my study. Eye tracking may have been able to provide data, such as if the user did fixate their attention to the cluttered background when viewing the vlog. Muir [32] found that users mainly fixated on the facial region of the signer when viewing Signed Language content but she did not study this outcome when the technical quality factors in the room were changed. Results from my study could have provided a better insight into the patters of eye gaze when viewing Signed Language content at varying quality.

4.4.5 **Re-recording options**

In the study a user was asked if they would re-record their video, based on the recommendations given by VAST, before submitting it to a professor or online. I originally had planned to allow users to re-record any of their videos in the study as many times as desired. The reason for re-creating a video would be to improve on its quality based on the feedback given from VAST. I hypothesized that the total number of times a user re-recorded their video could have indicated the importance of the specific factors to users rather than just asking people what they thought it would be. The study was approved to be one hour long so that a user would not feel fatigued from creating the videos. Extending that time by adding re-recording options would have likely extended that time.

5 Conclusions and Future Work

5.1 Conclusion

Vlogging has become a more commonplace activity as a new opportunity for communication and expression. Since Signed Languages are visual and do not have a written form, the Deaf community who speaks Signed Languages have seen vlogging as a new method to communicate with others without the need for text. One of the main drawbacks to vlogging for Signed Language communication, in real time and for pre-recorded content, is that it is vulnerable to video quality issues. If the content cannot be seen (rather than heard) by the viewer/audience, then it is not understandable and the intentions and meaning of that vlog are lost. Little research has been performed on ways to improve the experience of online multi-media use for Signed Language users. Furthermore, few resources exist on best practices for creating optimal quality video and many resources that do exist are not readily accessible to the general public.

A system called VAST was developed to create a more accessible experience for individuals who create, access and share Signed Language video content. This system is the first of its kind to assess the quality of talking head style videos based on the combination of lighting, signing space and background factors, and recommend improvements to the settings of these three factors. The system uses text and visual feedback to inform the user of possible quality issues and offers solutions to fix these problems. A user study was conducted on the VAST system to determine its usefulness and the impact the three factors had on a user's viewing experience. Results from this study indicate that participants found the recommendations from the system to be very helpful in determining quality.

The VAST study revealed that signing space had the highest impact on user perception of video quality. This was determined based on the results of the questions asked about the viewing and sharing of vlog content. Participants reported that they were the least likely to watch a vlog or submit their own vlog to others if a small signing space was used. Twelve out of fifteen participants ranked singing space to be among the most important technical factors when deciding quality. Participants also reported that signing space was important to them because a loss of viewable space may mean that important signs and user facial expression (critical information to signed communication) may not be seen by the viewer.

Lighting has also been shown to have an impact on a user's perception of video quality. The lighting factor was chosen more times (13/15) than any other factor when participants were asked to choose their top choices for factors to consider when determining quality. Participants also stated that they were not very likely to watch a vlog or share it with others if it experienced poor lighting. Participants were asked to comment on why lighting was important to them when creating vlogs. Participants noted that lighting contributes to the visual clarity of the vlog and viewing content without proper lighting may cause eyestrain and frustration.

Although participants reported that the VAST system provided the most help in determining the background setting of their vlog, users were still more likely to watch a video with a cluttered background than a video with faults with the other two technical factors. Only two out of a possible fifteen participants said the number of items on screen, which is correlated with background clutter, was one of the more important factors when determining vlog quality. It may be that since signers may be accustomed to seeing backgrounds with visual clutter, such as a messy bedroom or kitchen, they may regard clutter as a technical factor that does not contribute to the overall quality of the vlog as much as lighting or signing space do. In addition, it may be that the lighting and signing space factors are easier to perceive as they directly affect how much of the signer is visible. The background may be less noticeable as it does not directly affect the visibility of the signer.

The result that only two participants chose background as an important factor to consider when creating a vlog surprised me. I anticipated that the background would play a larger role in the perception of video quality as many Signed Language papers do conclude that including many items in the background may cause visual distraction for viewers, Pfau [30] for example. The use of a monochrome background is still a pre-requisite considered by Gallaudet University [29], which is a highly respected institution for deaf studies [5]. Therefore it is still useful to include the assessment of background into the system for the time being.

My study also provided an evaluation of the usability and functionality of VAST with Signed Language interpreter students. It was found that the interface of the VAST system was easy to read and easily understood, and that accomplishing the core tasks of the system was elementary. Participants noted that there was not a high learning curve to the system and that the simple design contributed to the overall ease of use. Participants also noted that the quick processing time of the system contributed to a positive overall experience of using the system. Few comments were made by participants for improvements of the system. Of the comments that were made, participants noted that the amount of information provided by VAST could be increased to ensure that each recommendation from the system was understood clearly. Participants also suggested that changing the colour of the feedback icons might help better represent the outcome of the recommendation. For example, one participant commented, *"the yellow checks were a bit confusing. Yellow is usually associated with "caution". I think green would have been more clear"*. No difficulties when performing the core tasks of the system were noted.

In summary, the results of the user study into the feasibility of a system that assesses and critiques Signed Language video content suggest that VAST is useful to users for determining overall vlog quality. The purpose of the study is not only to determine if a system that only reports issues and does not attempt to fix them is useful but also if the factors VAST analyses are important to users. The study results reveal that the factors assessed by the VAST system are important to users and do have an impact on a user's perception of vlog quality. Participants report that the system makes them attend to vlog quality and also that it would help them improve the quality of their future vlogs. Overall, VAST has been shown to provide valid and understandable feedback about three important technical factors

5.2 Future Work

5.2.1 User Evaluation

1. One of the most frequently asked questions by participants during the study is why we were not recruiting deaf individuals to be participants. A future research suggestion is to run a similar study with deaf vloggers and to gather their views of the VAST system. A comparison between hearing interpreter students and Deaf vloggers could be made to view the differences in opinions on the helpfulness of the VAST feedback and to view if amateur vloggers care about the quality of their vlogs enough to use the system. Signed Language interpreters are forced to pay attention to the quality of vlogs but amateur vloggers have no commitment to the quality of their vlog. As well, it would be interesting to compare the opinions of the interface between the two participant groups. The hearing interpreter students did not comment on the textual feedback and barely commented on the interface of VAST. Since Deaf individuals will be a primary user group of the system it would be feasible to get their input of the interface and to gather their opinions on how the feedback is presented (text and icons).

5.2.2 Technical Suggestions

1. VAST, as it was built for this thesis, assesses a vlog in its entirety. However, for later iterations of the software it may be feasible to provide feedback on different parts of a video if the quality settings are changed while the video is being recorded. For example, if the user begins recording a video with low lighting and then half-way through the recording he/she increases the lighting, the system should provide feedback for both lighting scenarios. Timing information may provide a more robust and interactive experience for users who create longer duration videos and want to test out and/or learn about quality settings by trying out different settings as they record.

2. VAST is developed as a proof-of-concept system designed for my Master's thesis to begin to address the need for Signed Language assessment tools. The user study has shown that this tool is useful for users to determine vlog quality and results indicate that if given the opportunity users may actually use this tool in the comfort of their own home. Future work could include the commercialization of this project to bring this system into the houses of individuals who create Signed Language video content and in turn help them to create better quality video content.

3. To simplify the process of using VAST, a technical suggestion would be to include a recording feature in the software. This would prevent users from having to use external software to record the vlog and then upload it into the system if they quickly want to create a vlog and have it assessed in one step.

4. A suggestion for better accessibility for Deaf users is to provide Signed Language equivalents of the assessment descriptions/feedback provided to users. Due to time limitations and that the system is only a proof-of-concept, definitions of neither functions nor suggestions are translated to ASL.

5. Some participants noted that it would be helpful if they could replay the video they uploaded so that after they received the recommendations from the system they could see how it applies to their vlog. This technical suggestion could work together with suggestion #1. If a video player is provided in VAST, recommendations can be placed visually on a timeline so that users can see where in the video does not have optimal quality.

6. Although the VAST system is a standalone application, future work could consider integrating this technology into vlog hosting websites. The tool could provide meta-data information for the uploaded vlogs and could be run in the background without the need for user input or initiation. The feedback could be modified to numerical ratings for easy viewing from other users of the site. A generated rating would provide a new filter for the searching of vlogs and

remedy the known issue for the lack of meta-data information users provide with their uploaded videos. The user who uploaded the video would still receive feedback in his or her own personal section of the site (text and icons) and would be given the option to re-upload the video or leave it as is.

Appendix

A. Ethics Approval Letter

REB 2013-289

Project Title: VAST: Vlog Analysis and Suggestion Tool

Dear Joseph Moscatiello,

The Research Ethics Board has completed the review of your submission. Your research project is now approved for a one year period as of Oct 23, 2013.The approval letter is attached in Adobe Acrobat (PDF) format.

Congratulations and best of luck with the project.

Please note that this approval is for one year only and will expire on October 23, 2014. Shortly before the expiry date a request to complete an annual report will be automatically sent to you. Completion of the annual report takes only a few minutes, enables the collection of information required by federal guidelines and when processed will allow the protocol to remain active for another year.

Please quote your REB file number (REB 2013-289) on future correspondence.

If you have any questions regarding your submission or the review process, please do not hesitate to get in touch with the Research Ethics Board (contact information below).

No research involving humans shall begin without the prior approval of the Research Ethics Board.

Record respecting or associated with a research ethics application submitted to Ryerson University.

Yours sincerely,

Toni Fletcher

Research Ethics Coordinator on behalf of

Lynn Lavallée, Ph.D.

Chair, Research Ethics Board

Associate Professor

Ryerson University EPH-241

350 Victoria St., Toronto, ON

(416)979-5000 ext. 4791

lavallee@ryerson.ca

rebchair@ryerson.ca

http://www.ryerson.ca/research

Toni Fletcher, MA

Research Ethics Co-Ordinator

Office of Research Services

Ryerson University

(416)979-5000 ext. 7112

toni.fletcher@ryerson.ca

http://www.ryerson.ca/research

B. Study Consent Form

CONSENT AGREEMENT SUMMARY

1. You will have the opportunity to participate in an evaluation of our vlog assessment and suggestion tool.
2. The researchers are interested in your experience and opinion of software that can analyse the technical quality of vlogs.
3. Agenda: Consent form, pre-questionnaire, sign a video recorded response to 5 different questions with a questionnaire after each response, post questionnaire
4. Participation is voluntary and you can stop at any time.
5. Everything you say and do will remain confidential.

Consent Agreement

Principal Investigator: Joseph Moscatiello, Ryerson University

(416) 979-5000 ext. 2523 or jmoscati@ryerson.ca

Faculty Supervisor: Deborah Fels, Ph.D., P.Eng. Ryerson University

(416) 979-5000 ext. 7619 or dfels@ryerson.ca

Joseph Moscatiello is a Ryerson University student in the School of Computer Science currently studying towards his Master degree in Computer Science. His supervisor is Deborah Fels. This research is for Joseph's master's thesis, which is a requirement for his graduation.

Project Title: VAST: VLOG ANALAYSIS AND SUGGESTION TOOL

You are being asked to participate in a research study. Before you give your consent to be a volunteer, it is important that you read the following information and ask as many questions as necessary to be sure you understand what you will be asked to do.

Purpose of the Study: The Vlog (video blog) Analysis and Suggestion Tool (VAST) provides suggestions for improvement in the quality of three technical factors within an ASL vlog. These technical factors include lighting, digital sign space and amount of clutter contained in a vlog. The aims of this study are to gather participant's opinions and impressions of their experience with the use of and recommendations provided by the VAST system, whether recommendations provided by VAST would influence a user's decision to modify/improve their video and which technical factor has the most impact on user perception of quality of self-created vlogs.

Description of the Study: First, you will be asked to complete a pre-study questionnaire to collect some background information and experience with video blogging (vlogging). Once the pre-questionnaire is completed, you will be asked to sign the response to five different questions while being video taped. Between each of the five questions some of the technical factors will be changed. For example, the brightness of the lights may be increased or decreased slightly in order to view the effects lighting has on perceived video quality. The video you create will then be analyzed by the VAST software and it will provide recommendations on whether the technical factors are at optimum levels or should be adjusted. You will then be asked to complete a short questionnaire on your opinion of the software's advice and whether you would take it into consideration should you record another vlog. Once all responses are completed, we will ask you to fill out one final questionnaire to gather your overall impressions of the VAST system. The entire session will be videotaped so that we can record any commentary or discussion that happens during the study.

Principal Investigator: Joseph Moscatiello, Ryerson University

(416) 979-5000 ext. 2523 or jmoscati@ryerson.ca

Faculty Supervisor: Deborah Fels, Ph.D., P.Eng. Ryerson University

(416) 979-5000 ext. 7619 or dfels@ryerson.ca

Project Title: **VAST: Vlog Analysis and Suggestion Tool**

Risks or Discomforts: The risks associated with the study are minimal. You might feel uncomfortable or fatigued while responding to the individual questions or questionnaires. If you feel tired or uncomfortable, you may take a break to rest or discontinue participation in the study either temporarily or permanently. You may feel uncomfortable being video-taped. We will turn on the camera during the pre-questionnaire so that you can become use to it being on. If that does not help, then we will stop the study.

Benefits of the Study: It is not foreseen that you will personally benefit from participation in this study. However, the results from this research may contribute to the development of VAST which may in turn help vloggers improve the technical quality of their vlogs.

Confidentiality: All data will remain confidential; will be secured at the Inclusive Media and Design Centre at Ryerson University and destroyed after five years. Furthermore, only the principal investigator and faculty supervisor of this study will have access to the data for analysis purposes. Data will only be presented in summary form and no one individual will be identified. Number codes will be used to link data with personal information. We will also be recording the study on video. We will not use this footage in any public setting, and the footage will be stored on our password protected lab servers located at Ryerson University.

Costs and/or Compensation for Participation: There are no costs associated with your participation. You will be compensated with $30.00 in cash for completing the entire study. If you choose not to finish the study, you will still be given $30 for your participation.

Voluntary Nature of Participation: Participation in this study is voluntary. Your choice of whether or not to participate will not influence your future relations with Ryerson University. If you decide to participate, you are free to withdraw your consent and to stop your participation at any time without any penalty. At any particular point in the study, you may refuse to answer any particular question or stop participation altogether.

Questions about the Study:

We sincerely appreciate your co-operation. If you have any questions or concerns, please do not hesitate to call Joseph Moscatiello at 416-979-5000 ext. 7110 or Deborah Fels at 416-979-5000 ext. 7619.

Questions/ Concerns about Participant's Rights:

The Research Ethics Board Chair, Dr. Lynn Lavallée, may be contacted at (416) 979-5000 ext. 4791 or at rebchair@ryerson.ca should there be any complaints or concerns about the participant's rights as a participant, c/o Office of Research Services, Ryerson University EPH-241, 350 Victoria St., Toronto, ON M5B 2K3, Tel: (416) 979-5042.

Principal Investigator: Joseph Moscatiello, Ryerson University

(416) 979-5000 ext. 2523 or jmoscati@ryerson.ca

Faculty Supervisor: Deborah Fels, Ph.D., P.Eng. Ryerson University

(416) 979-5000 ext. 7619 or dfels@ryerson.ca

Project Title: VAST: VLOG ANALYSIS AND SUGGESTION TOOL

Consent Form to Participate in Study

Agreement:

Your signature below indicates that you have read the information in this agreement, have had a chance to ask any questions you have about the study, and know that your participation is entirely voluntary. Your signature also indicates that you agree to be in the study and have been told that you can change your mind and withdraw your consent to participate at any time. You have been given a copy of this agreement.

You have been told that by signing this consent agreement you are not giving up any of your legal rights.

Name of Participant (please print)

__________________________________ ________________

Signature of Participant Date

__________________________________ ________________

Signature of Investigator Date

Your signature below indicates that you **agree** to be video-taped during the study.

______________________________ _________________

Signature of Participant Date

C. Study Questionnaire

5.2.3 Pre-Study Questionnaire

VAST Pre-study Questionnaire

Purpose of pre-study questionnaire: The purpose of these questions is to collect general information about you and your video blogging (Vlogging) habits when using a Signed Language for communication. It should take less than 10 minutes to complete this questionnaire.

1. **Please indicate your age:**
   - □ 18 – 24
   - □ 25 – 34
   - □ 35 – 44
   - □ 45 – 54
   - □ 55 – 64
   - □ 65 +

2. **Please indicate your gender:**
   - □ Male
   - □ Female

3. **What is your highest level of education completed? (Please check one)**
   - □ No formal education
   - □ Elementary school
   - □ High School
   - □ College
   - □ University
   - □ Graduate School

4. **How often do you use a computer? (Please check one)**
   - □ Everyday
   - □ Every 2 – 3 days
   - □ Once a week
   - □ Once a month
   - □ Never

5. **How often do you create video blogs (vlogs)? (Please check one)**
   - □ Everyday
   - □ Every 2 – 3 days
   - □ Once a week
   - □ Once a month
   - □ Never

6. **What types of vlogs do you typically create? (Please check all that apply)**
   - □ Personal
   - □ For my coursework
   - □ Political

- ❑ Informational
- ❑ Other:________________
- ❑ I don’t create vlogs

7. **Which recording tool/s do you use to create vlogs with? (Please check all that apply)**
   - ❑ Adobe Premiere
   - ❑ iMovie
   - ❑ YouTube
   - ❑ Movie Maker
   - ❑ Other _________________
   - ❑ I don’t create vlogs

8. **How often do you view vlogs created by other people?**
   - ❑ Everyday
   - ❑ Every 2 – 3 days
   - ❑ Once a week
   - ❑ Once a month
   - ❑ Never

9. **Please indicate in which year you are currently studying**
   - ❑ 2nd Year
   - ❑ 3rd Year
   - ❑ 4th Year

10. Rate your level of agreement with the following statements:

	Strongly Disagree	Disagree	Don’t Care	Agree	Strongly Agree
Direct sunlight is the best lighting for vlog creation					
The lighting in this room is not ideal for vlog creation					
Mood lighting should be used when creating vlog content					
A solid colour painted wall with no items on it is an ideal background for vlog creation					
An office environment would not be an ideal background to use when creating vlog content					

Producing vlog content in front a window with incoming light is best for vlog creation					
Adjusting the camera so that your face and a small space around your head is in the frame is best for vlog creation					
Adjusting the camera so that it is able to view your body but not your face is best for vlog creation					
Adjusting the camera so that your head and body are in the video frame is not ideal for vlog creation					

5.2.4 **Between Study Questionnaire**

VAST: After each vlog recording

Purpose of the intermediate questionnaire: The purpose of this questionnaire is to collect your feedback about your suggestions of the VAST system on the vlog you created. This feedback consists of the intelligibility of the vlog, as well as your understanding of the various visual elements. It should take about 5 minutes to complete these questions.

1. **Please summarize, in one or two sentences, what the system told you.**

2. **How helpful was the recommendation on lighting provided by VAST in determining the technical quality of your vlog (Please select one)?**

Not at all Helpful	Not very Helpful	Neutral	Helpful	Very Helpful
❑	❑	❑	❑	❑

Why?

3. **How helpful was the recommendation on signing space/video frame from VAST in determining the technical quality of your vlog (Please select one)?**

Not at all Helpful	Not very Helpful	Neutral	Helpful	Very Helpful
❑	❑	❑	❑	❑

Why?

4. How helpful was the recommendation on the vlog background provided by VAST in determining the technical quality of your vlog (Please select one)?

Not at all Helpful	Not very Helpful	Neutral	Helpful	Very Helpful
□	□	□	□	□

Why?

5. If you knew another vlog had similar visual quality to the one you just made, how likely would you be to watch that vlog (Please select one)?

Not at all Likely	Somewhat Unlikely	Likely	Somewhat Likely	Very Likely
□	□	□	□	□

Why?

6. You have been instructed to submit the vlog you just created to a professor. Given the time it takes to recreate the vlog, how likely are you to remake it to change the technical quality factors based on the VAST system recommendations (Please select one)?

Not at all Likely	Somewhat Unlikely	Likely	Somewhat Likely	Very Likely
□	□	□	□	□

What would you do?

Lighting

- □ Increase Lighting Levels
- □ Decrease Lighting Levels
- □ Would not change the lighting levels

Signing Space

- □ Zoom out with camera
- □ Zoom in with camera
- □ Would not change camera settings

Background

- □ Add items to the background
- □ Decrease the items in the background
- □ Would not modify the background

7. **You want to submit this vlog to a popular video sharing website, such as YouTube. Given the time it takes to recreate the vlog, how likely are you to remake it to change the technical quality factors based on the VAST system recommendations (Please select one)?**

Not at all Likely	Somewhat Unlikely	Likely	Somewhat Likely	Very Likely
❑	❑	❑	❑	❑

What would you do?

Lighting

- ❑ Increase Lighting Levels
- ❑ Decrease Lighting Levels
- ❑ Would not change the lighting levels

Signing Space

- ❑ Zoom out with camera
- ❑ Zoom in with camera
- ❑ Would not change camera settings

Background

- ❑ Add items to the background
- ❑ Decrease the items in the background
- ❑ Would not modify the background

5.2.5 **Post Study Questionnaire**

VAST Post-Study Questionnaire

Purpose of the post-study questionnaire: The purpose of this questionnaire is to understand the effect of each visual quality factor, the difficulties of comprehending individual vlogs as well as your likes and dislikes of the experience.

1. **Rate the level of difficulty of importing a vlog into the VAST system?**

Very Difficult	Difficult	Neutral	Easy	Very Easy
❑	❑	❑	❑	❑

2. **Rate the level of difficulty of generating recommendations for the vlogs you created using the VAST system?**

Very Difficult	Difficult	Neutral	Easy	Very Easy

□	□	□	□	□

3. Rate the level of difficulty of reading the information provided by the VAST system?

Very Difficult	Difficult	Neutral	Easy	Very Easy
□	□	□	□	□

Why?

4. Would you watch a vlog if:

Vlog Information	Would not watch vlog	Might not watch vlog	Don't Care	Might still watch vlog	Would definitely watch vlog
It was filmed with a blue coloured wall in the background.					
The vlog is recorded where you can see the face and shoulders of the signer.					
It is recorded with optimal lighting levels					
There is a lot of					

clutter in background of the signer					
The light in the vlog is too bright					
The vlog is recorded where all you can see is the signers face					
It is recorded with dim lighting					
The vlog is recorded where you can see the face and body of the signer					

It is recorded with bright lighting					
No clutter is found in background of signer					
It is recorded with dark lighting					

5. When deciding whether a vlog is good quality, which technical factors do you consider important. Please select your top <u>three</u> choices.

- ❑ Lighting
- ❑ Video sharpness
- ❑ Number of items on screen at one time
- ❑ Signing speed of signer
- ❑ Frame rate
- ❑ Clothing on signer
- ❑ Viewable signing space
- ❑ Other:________________

6. Using the VAST system has had the following effects on me:

- ❑ I learned about making vlogs with good quality
- ❑ It made me pay attention to vlog quality
- ❑ It helped me to understand the impact of certain technical factors on viewers
- ❑ It would help me to improve the quality of my future vlogs
- ❑ I was confused by the recommendations
- ❑ It took too much time
- ❑ I was bored
- ❑ There was nothing new
- ❑ Don’t care, not important

7. Why is lighting important or not important to you when creating vlogs?

8. Why is digital signing space important or not important to you when creating vlogs?

9. Why is the chosen background of a vlog important or not important to you when creating vlogs?

10. What did you like most about the VAST system?

11. **What did you dislike most about the VAST system?**

D. Thematic Analysis Descriptive statistics

Below are the descriptive statistics for the ICC values from the open ended questions.

Question	ICC value
Q1. Rate the level of difficulty of reading the information provided by the VAST system? Why? (Likert scale question and explanation)	0.83
Q2. What did you like most about the VAST system?	0.93
Q3. What did you dislike most about the VAST system?	0.92

E. References

[1] YouTube. “YouTube statistics.” Internet - https://www.youtube.com/yt/press/statistics.html, 2004.

[2] G. Valentine and T. Skelton, "Changing spaces: the role of the internet in shaping Deaf geographies," *Social & Cultural Geography,* vol. 9, pp. 469-485, Aug. 2008.

[3] H. Zettl, *Television Production Handbook, Tenth Edition.* Michael Rosenberg, Boston: Wadsworth Publishing Company, 2009.

[4] S. C. Herring *et al.*, "Bridging the gap: A genre analysis of weblogs," *in Proceedings of the 37th Annual Hawaii International Conference on System Sciences,* 2004, pp. 1-11.

[5] E. Keating and G. Mirus, "American Sign Language in virtual space: Interactions between deaf users of computer-mediated video communication and the impact of technology on language practices," *Language in Society,* vol. 32, pp. 693-714, 2003.

[6] S. Hooper *et al.*, "The effects of digital video quality on learner comprehension in an American Sign Language assessment environment," *Sign Language Studies,* vol. 8, pp. 42-58, 2007.

[7] M. Sun *et al.*, "Active lighting for video conferencing," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 19, pp. 1819-1829, 2009.

[8] E. S. Hibbard and D. Fels, "The Vlogging Phenomena: A Deaf Perspective," in *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility,* 2011, pp. 59-66.

[9] R. van den Berg *et al.*, "A crowding model of visual clutter," *Journal of Vision,* vol. 9, pp. 1-11, 2009.

[10] J. R. Gannon, *Deaf Heritage.* Washington: Gallaudet University Press, 1981, pp. 1-22.

[11] A. Goodrum *et al*., "The Creation of Keysigns: American Sign Language Metadata," in Proceedings of the Tenth International ISKO Conference, 2008, pp. 282-286.

[12] K. Emmorey, "Processing a dynamic visual—Spatial language: Psycholinguistic studies of American Sign Language," *Journal of Psycholinguistic Research,* vol. 22, pp. 153-187, 1993.

[13] R. Conrad, "Lip-reading by deaf and hearing children," British Journal of Educational Psychology, vol. 47, pp. 60-65, Feb. 1977.

[14] K. Emmorey *et al*., "Visual imagery and visual-spatial language: Enhanced imagery abilities in deaf and hearing ASL signers," *Cognition,* vol. 46, pp. 139-181, 1993.

[15] Bettger *et al*., "Enhanced facial discrimination: Effects of experience with American Sign Language," *Journal of Deaf Studies and Deaf Education,* vol. 2, pp. 223-233, 1997.

[16] I. Matthews *et al*., "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 198-213, 2002.

[17] N. Michael *et al.*, "Computer-based recognition of facial expressions in ASL: From face tracking to linguistic interpretation," in *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC, Malta,* 2010.

[18] M. Kemp, "Why is learning American Sign Language a challenge?" *American Annals of the Deaf,* vol. 143, pp. 255-259, Jul. 1998.

[19] C. T. Akamatsu *et al*., "An investigation of two-way text messaging use with deaf students at the secondary level," *Journal of Deaf Studies and Deaf Education,* vol. 11, pp. 120-131, 2006.

[20] S. M. Wilson and L. C. Peterson, "The anthropology of online communities," *Annual review of anthropology,* pp. 449-467, 2002.

[21] W. C. Stokoe, "Sign language structure: An outline of the visual communication systems of the American deaf," Journal of Deaf Studies and Deaf Education, vol. 10, pp. 3-37, 2005.

[22] J. Hoem, "Videoblogs as Collective Documentary," presented at BlogTalk 2.0, Vienna, 2004. [23] W. Gao *et al*., "Vlogging: A survey of videoblogging technology on the web," *ACM Computing Surveys,* vol. 42, pp. 151-157, Jun. 2010. [24] Heather Molyneaux *et al*., "Exploring the Gender Divide on YouTube: An Analysis of the Creation and Reception of Vlogs," *American Communication Journal*, vol. 10, No. 2, 2008.

[25] D. Fels et al., "Providing inclusive video-mediated communication," *Annual Review of Communications,* vol. 57, pp. 593-601, 2004.

[26] E. S. Hibbard and D. Fels, "The Vlogging Phenomena: A Deaf Perspective," in *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility,* 2011, pp. 59-66.

[27] B. A. Nardi *et al*., "Blogging as social activity, or, would you let 900 million people read your diary?" in *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work,* 2004, pp. 222-231.

[28] L. J. Muir and I. E. Richardson, "Video telephony for the Deaf: Analysis and Development of an Optimised Video Compression Product," pp. 650-652, 2002.

[29] Gallaudet University, "American Sign Language Video Assignment Rubric," Internet: http://www.gallaudet.edu/Documents/Grad/VideoASSGN.docx, [June 12, 2013].

[30] R. Pfau *et al.*, Sign Language: An International Handbook. Walter de Gruyter, 2012. [31] T. Byrd, "Deaf Space", Publication:2007.

[32] L. Muir *et al*., "Gaze tracking and its application to video coding for sign language," in *Picture Coding Symposium,* 2003, pp. 23-25.

[33] A. Cavender, R. E. Ladner and E. A. Riskin, "MobileASL:: Intelligibility of sign language video as constrained by mobile phone technology," in Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, 2006, pp. 71-78.

[34] L. Muir, I. Richardson and S. Leaper, "Gaze tracking and its application to video coding for sign language," in Picture Coding Symposium, 2003, pp. 23-25.

[35] D. Agrafiotis *et al.*, "Perceptually optimised sign language video coding based on eye tracking analysis," *Electronics Letters,* vol. 39, pp. 1703-1705, Nov. 2003.

[36] P. Sleegers *et al.*, "Lighting affects students' concentration positively: findings from three Dutch studies," *Lighting Research and Technology,* vol. 45, pp. 159-175, 2012.

[37] C. McCloughan et al., "The impact of lighting on mood," *Lighting Research and Technology,* vol. 31, pp. 81-88, 1999.

[38] N. P. Erber, "Effects of angle, distance, and illumination on visual reception of speech by profoundly deaf children," *Journal of Speech, Language and Hearing Research,* vol. 17, pp. 99-112, Mar. 1974.

[39] T. Goodman, "Measurement and specification of lighting: A look at the future," Lighting

Research and Technology, vol. 41, pp. 229-243, 2009.

[40] B. K. Horn, "Understanding image intensities," Artif. Intell., vol. 8, pp. 201-231, 1977.

[41] P. Boyce, "Age, illuminance, visual performance and preference," Lighting Research and

Technology, vol. 5, pp. 125-144, 1973.

[42] W. L. Braje, D. Kersten, M. J. Tarr and N. F. Troje, "Illumination effects in face recognition," 1998.

[43] D. L. Butler and P. M. Biner, "Preferred Lighting Levels Variability Among Settings, Behaviors,

and Individuals," Environ. Behav., vol. 19, pp. 695-721, 1987.

[44] J. E. Flynn, T. J. Spencer, O. Martyniuk and C. Hendrick, "Interim study of procedures for investigating the effect of light on impression and behavior," Journal of the Illuminating Engineering Society, vol. 3, pp. 87-94, 1973.

[45] J. Veitch and G. Newsham, "Preferred luminous conditions in open-plan offices: Research and practice recommendations," Lighting Research and Technology, vol. 32, pp. 199-212, 2000.

[46] I. Knez, "Effects of indoor lighting on mood and cognition," J. Environ. Psychol., vol. 15, pp. 39-51, 1995.

[47] C. Shi, K. Yu, J. Li and S. Li, "Automatic image quality improvement for videoconferencing," in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04).

[48] I. Lewin and W. Bell, "Luminance measurement by photographic photometry," Illuminating Engineering, vol. 63, pp. 582-589, 1968.

[49] H. J. Andersen and E. Granum, "Classifying the illumination condition from two light sources by color histogram assessment," Josa a, vol. 17, pp. 667-676, 2000.

[50] L. Tao and V. Asari, "Modified luminance based MSR for fast and efficient image enhancement," in Applied Imagery Pattern Recognition Workshop, 2003. Proceedings. 32nd, 2003, pp. 174-179.

[51] L. O. Beltran and B. M. Mogo, "Assessment of luminance distribution using HDR photography," in ISES Solar World Congress, ISES Solar World Congress, 2005, .

[52] V. Vezhnevets, V. Sazonov and A. Andreeva, "A survey on pixel-based skin color detection techniques," in Proc. Graphicon, 2003, pp. 85-92.

[53] W. Niblack, An Introduction to Digital Image Processing. Strandberg Publishing Company, 1985.

[54] S. Bezryadin, P. Bourov and D. Ilinih, "Brightness calculation in digital image processing," in International Symposium on Technologies for Digital Photo Fulfillment, 2007, pp. 10-15.

[55] Techniques for Accessibility Evaluation and Repair Tools.

[56] R. Rosenholtz, Y. Li and L. Nakano, "Measuring visual clutter," Journal of Vision, vol. 7, 2007.

[57] G. A. Alvarez and P. Cavanagh, "The capacity of visual short-term memory is set both by visual information load and by number of objects," Psychological Science, vol. 15, pp. 106-111, 2004.

[58] M. J. Bravo and H. Farid, "A scale invariant measure of clutter," Journal of Vision, vol. 8, 2008.

[59] A. Oliva, M. L. Mack, M. Shrestha and A. Peeper, "Identifying the perceptual dimensions of visual complexity of scenes," in Proc. of the 26th Annual Meeting of the Cogn. Sci. Soc, 2004.

[60] R. Rosenholtz, Y. Li, J. Mansfield and Z. Jin, "Feature congestion: A measure of display clutter," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2005, pp. 761-770.

[61] VirtualDeafChurch –"A New Look!" http://www.youtube.com/watch?v=P2_XtejsGao, 2013

[62] D. Whitney and D. M. Levi, "Visual crowding: A fundamental limit on conscious perception and object recognition," Trends Cogn. Sci. (Regul. Ed.), vol. 15, pp. 160-168, 2011.

[63] F. Heylighen, "The growth of structural and functional complexity during evolution," The Evolution of Complexity, pp. 17-44, 1999.

[64] D. Bavelier et al., "Do deaf individuals see better?" Trends in Cognitive Sciences, vol. 10, pp. 512-518, Oct. 2006.

[65] N. Senthilkumaran and R. Rajesh, "Edge detection techniques for image segmentation- a survey of soft computing approaches," International Journal of Recent Trends in Engineering, vol. 1, 2009.

[66] R. Maini and H. Aggarwal, "Study and comparison of various image edge detection techniques," International Journal of Image Processing (IJIP), vol. 3, pp. 1-11, 2009.

[67] J. Canny, "A computational approach to edge detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, pp. 679-698, 1986.

[68] M. A. Oskoei and H. Hu, "A survey on edge detection methods," University of Essex, UK, 2010.

[69] K. Emmorey, "Space on hand: The exploitation of signing space to illustrate abstract thought." 2001.

[70] J. G. Kyle and B. Woll, Sign Language: The Study of Deaf People and their Language. Cambridge University Press, 1988.

[71] A. Braffort and P. Dalle, "Sign language applications: preliminary modeling," Universal Access in the Information Society, vol. 6, pp. 393-404, 2008.

[72] H. Bauman and J. Murray, "Reframing: From Hearing Loss to Deaf Gain," Deaf Studies Digital

Journal, vol. 1, 2009.

[73] Z. Rasheed, Y. Sheikh and M. Shah, "On the use of computable features for film classification," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 15, pp. 52-64, 2005.

[74] P. Robertson, "Robojourno: reframing the talking head," 2002.

[75] P. Xu, L. Xie, S. Chang, A. Divakaran, A. Vetro and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video." in Icme, 2001, pp. 928-931.

[76] M. H. Kolekar, K. Palaniappan, S. Sengupta and G. Seetharaman, "Semantic concept mining based on hierarchical event detection for soccer video indexing," Journal of Multimedia, vol. 4, pp. 298-312, 2009.

[77] Y. Wang, Z. Liu, G. Hua, Z. Wen, Z. Zhang and D. Samaras, "Face re-lighting from a single image under harsh lighting conditions," in Computer Vision and Pattern Recognition, 2007. CVPR'07.IEEE Conference on, 2007, pp. 1-8.

[78] S. Gundimada, L. Tao and V. Asari, "Face detection technique based on intensity and skin color distribution," in Image Processing, 2004. ICIP'04. 2004 International Conference on, 2004, pp. 1413-1416.

[79] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Tech. rep., Microsoft Research, 2010.

[80] Y. Lu, J. Zhou and S. Yu, "A survey of face detection, extraction and recognition," Computing and Informatics, vol. 22, pp. 163-195, 2012.

[81] P. Viola and M. J. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57, pp. 137-154, 2004.

[82] K. Sung and T. Poggio, "Example-based learning for view-based human face detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, pp. 39-51, 1998.

[83] M. Stoerring, H. J. Andersen and E. Granum, "Skin colour detection under changing lighting conditions," in 7th Symposium on Intelligent Robotics Systems, 1999, .

[84] N. Habili, Automatic Segmentation of the Face and Hands in Sign Language Video Sequences, 2001.

[85] P. Viola, M. J. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, 2003, pp. 734-741.

[86] C. Kotropoulos and I. Pitas, "Rule-based face detection in frontal views," in Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, 1997, pp. 2537-2540.

[87] X. Lu and A. K. Jain, "Ethnicity identification from face images," in Defense and Security, 2004, pp. 114-123.

[88] D. Chai and K. N. Ngan, "Locating facial region of a head-and-shoulders color image," in Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, 1998, pp. 124-129.

[89] D. Keren, M. Osadchy and C. Gotsman, "Antifaces: A novel, fast method for image detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, pp. 747-761, 2001.

[90] C. P. Papageorgiou, M. Oren and T. Poggio, "A general framework for object detection," in Computer Vision, 1998. Sixth International Conference on, 1998, pp. 555-562.

[91] K. Grauman, Ed., Visual Object Recognition. Oregon: Morgan & Claypool, 2011.

[92] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands, 1993, pp. 206-213.

[93] Government of Canada. "Canada Occupational Health and Safety Regulations (SOR/86-304)." http://laws.justice.gc.ca/eng/regulations/sor-86-304/page-19.html, 2004

[94] The Engineering Toolbox. "Illuminanace - Recommended Light Levels." http://www.engineeringtoolbox.com/light-level-rooms-d_708.html, 2003

[95] J. Jacoby and M. S. Mattel, "Three-Point Likert Scales Are Good Enough." Journal of Marketing Research (JMR), vol. 8, 1971.

[96] A. Field, Discovering Statistics using IBM SPSS Statistics. Sage, 2013.

[97] IBM statistics, SPSS. http://www-01.ibm.com/software/analytics/spss/solutions.html. Accessed 2014

[98] Laerd Statistics: One-way repeated measures ANOVA in SPSS Statistics. Laerd.statistics.com

[99] J. Aronson, "A pragmatic view of thematic analysis," The Qualitative Report, vol. 2, pp. 1-3, 1994.

[100] S. Hooper, C. Miller, S. Rose and G. Veletsianos, "The effects of digital video quality on learner comprehension in an American Sign Language assessment environment," Sign Language Studies, vol. 8, pp. 42-58, 2007.

[101] W. Lidwell, K. Holden and J. Butler, Universal Principles of Design, Revised and Updated: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, make Better Design Decisions, and Teach through Design. Rockport Pub, 2010.