Sign Language Interpretation System

The University of Sydney
Faculty of Engineering & Information Technologies
Ryan Michael Thomas, 307204154

1. Abstract

An estimated 360 million people worldwide suffer from hearing loss. The vast majority of able hearing individuals do not understand sign language, and consequently communication between those who can hear and those who are deaf present difficulties and challenges. While communication methods exist, a cost-effective, efficient interface does not exist to translate sign language into readable language, and readable language into sign language, in real time. This report outlines a web camera based application which converts sign language into readable language.

2. Background

2.1. Sign Language

Sign language is a language which uses manual communication and hand signals to convey meaning. Hand shapes, orientation and movement of the hands, express a speaker's thoughts. Signing is not only used by the deaf, it is also used by people who can hear but cannot physically speak. Hundreds of sign languages are in use around the world and are at the core of deaf cultures. A common misconception is that all sign languages are the same worldwide, or that sign language is international. Each country generally has its own native sign language. It is believed there are up to 137 sign languages. Methods exist for translating sign language and aiding communication between a signer and non-signer. The most common approaches include a human sign language interpreter, voice-to-text transcription, and pen and paper. It is only in recent years that sign language interpretation has been explored using modern software.

2.2. Related Work

Researchers at the Chinese Academy of Sciences and Beijing Union University teamed up with Microsoft Research Asia to create a feasible sign language interpretation system. Microsoft Kinect technology was used to create a cost-effective and capable technology prototype that enables sign language communication between signer and non-signer, and offers translation between different sign languages.

The gestures of sign language are converted to spoken and written language, and vice versa. The system thus captures a conversation from both sides. For example, the signer is shown signing, with a written and spoken translation being rendered in real-time for the hearing person. The nonsigner is represented by an avatar that takes his or her spoken words and turns them into accurate, understandable signs to accompany the written translation for the deaf person to "read," all in real-time. However, the technology is dependent upon Microsoft's Kinect technology which is both large and costly, so the system still faces limitations in terms of cost, ease of use, and accessibility.

2.3. Significance

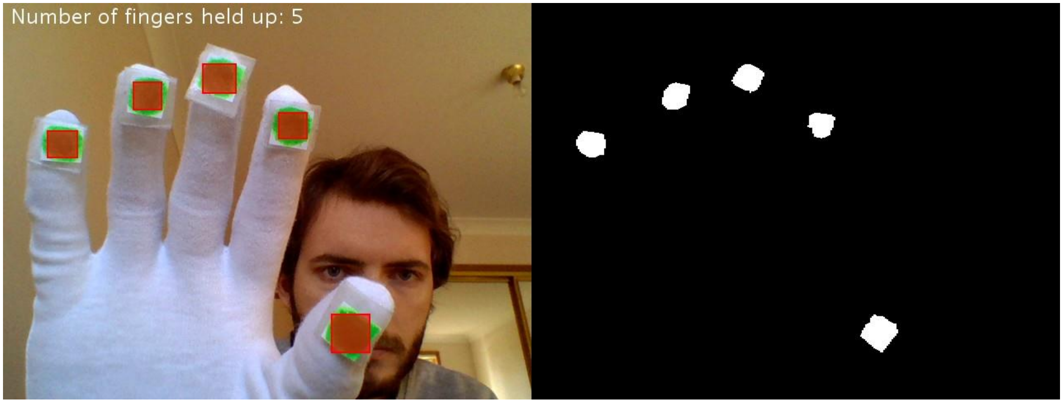
Consider a scenario in which a person with deafness visits a physician who doesn't understand sign language. A patient could pre-schedule an interpreter or resort to communicating with paper and pen, but this is either too expensive or too awkward. But with the sign language interpretation technology, the deaf patient and the physician could simply use sign language to communicate. Such technology would empower deaf people around the globe, allowing for effective, independent communication between those who are deaf and those who are not deaf

3. System Design

3.1. Glove

The aim of the sign language interpretation system is to identify meaning through hand gestures. In order to achieve an interpretation of the hand gestures, green coloured dots were added to each fingertip of a white glove. Two additional green coloured dots were added to the knuckles of the index and pinkie fingers. An example is shown in figure 1.

Figure 1: Glove setup.



Number of fingers held up: 5

3.2. Real Time Output

Two screens are displayed. The left screen will display the video in real time, and should be used for guidance and symbol detection. A red bounding box appears overlayed on the green dot once successfully identified by the algorithm. Text appears in white in the top left hand corner, identifying the number of fingers held up. Additional text is presented in red below when the appropriate hand gesture is made. The screen on the right shows the binary threshold version of the detected colours, and should be used to determine whether or not the colours are being adequately detected, and whether the removal of background noise is sufficient. If the white circular dots are not present, the lighting conditions of the room may be inadequate to accurately identify the coloured dots.

3.3. Algorithm

The interpretation of the hand gesture is dependent upon both the position of the fingers and the movement of the fingers. Hence, the algorithm for sign language interpretation must be able to determine how many fingers are held up, and determine the position of those fingers relative to one another. The algorithm determines how many fingers are held up via a colour tracking algorithm, and then uses the position of these colours relative to one another to determine the hand gesture. Clearly a computational system which could interpret every hand signal within a sign language would take many researchers working over several months, or even years, to accurately implement. To ensure a working product is created within the short amount of time allocated, the approach has been simplified to create a system which identifies simple hand gestures. Numbers from zero to five have been interpreted. The gestures; “peace”, “thumbs up”, and “surf’s up” have also been interpreted.

4. Technical Approach

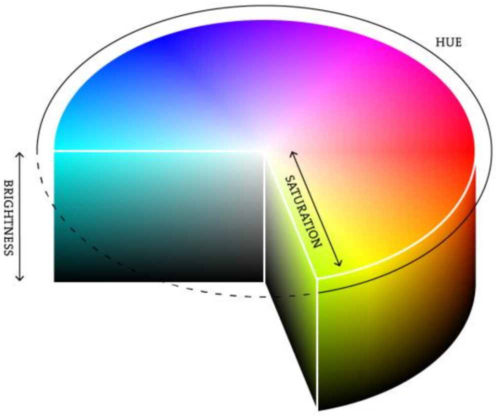
4.1. Colour tracking

Colour tracking is the ability to take an image, isolate a particular colour and extract information about the location of a region of that image that contains just that colour. In order to specify colour, minimum and maximum allowable values must be defined for three colour channels. Every unique colour is represented by a red, green, and blue value [1]. Since light is not perfectly uniform and the colour of an object is not perfectly uniform, a range of allowable values for all three colour channels is required. The bounds should not be too wide as unwanted colours will be accepted, yet not too thin so that coloured dots are not identified.

4.2. RGB to HSV

Detecting coloured dots using the HSV (hue, saturation, value) format, rather than the RGB (red, green, blue) format, improves the reliability of the colour tracking algorithm. HSV has the advantage of only having to use a single number to detect the coloured dot, even when several shades of colour are present. The green RGB value used is RGB(40,180,80), which is then converted into the corresponding hue value in Processing. Note, HSV is sometimes referred to as HSB. An example of the HSB system can be found in figure 2 [2].

Figure 2: HSB chart.



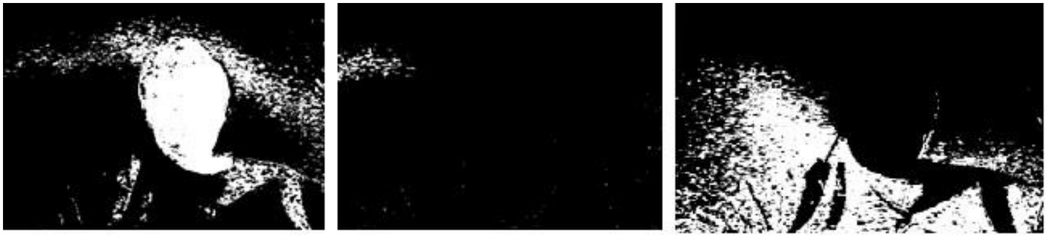
4.3. Thresholding

The webcam image was binary thresholded in order to improve the colour detection. The hue (with a hue range of ±12 either side of the hue value) was cast as white and the rest of the image was cast as black. Thresholding makes the detection of blobs and contours much easier and hence improves the accuracy of the colour tracking algorithm [3].

4.4. Background noise

In order to optimise the accuracy of the sign language identification algorithm, background noise must be minimised. Hence, the choice of colour used for the coloured dots is important. Human skin contains a considerable amount of red hues, which creates substantial background noise when trying to identify red dots. As a result, red is a poor colour choice for colour tracking in the sign language interpretation system. Similarly, there is a considerable amount of background noise in the blue hue, due to the common usage of blue hues in clothing. Green, however, is a more appropriate choice as there is little green hue in human skin tone, and little green hue in clothing, which reduces the amount of background noise. An example of background noise due to red, green, and blue can be found in figure 3. Clearly, the green hue contains the least amount of background noise, and therefore was the chosen colour for the sign language interpretation system.

Figure 3: Background noise of red, green, and blue.

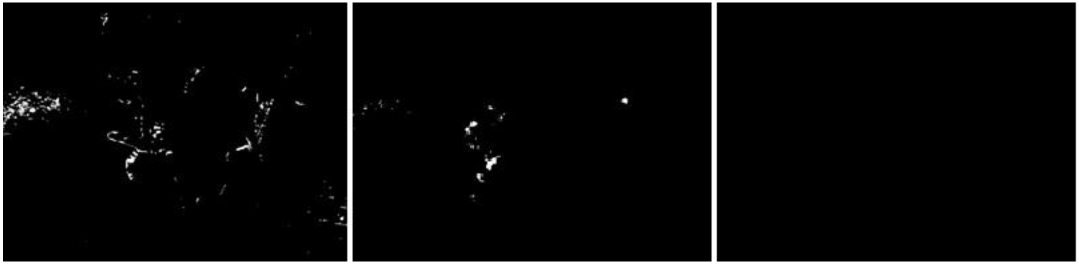


Red RBG(200,0,0) Green RGB(0,200,0) Blue RGB(0,0,200)

4.5. Erosion

Even though the choice of colour substantially reduces the amount of background noise, the problem of incorrectly identifying green is still possible. To further reduce the amount of background noise, the frame was eroded several times, the results of which can be found in figure 4. Without erosion, there is a significant amount of background noise. Eroding the image once reduces the background noise, but leads to clumping of blobs which may be mistakenly identified as the coloured dots of the glove. Eroding the image several times yields the most accurate result. Erosion also decreases the circle size of the green dots of the glove, which is undesirable. Hence, there must be a balance between removing background noise and eroding the green dots on the glove.

Figure 4: Erosion of background noise.



No erosion Eroded once Eroded several times

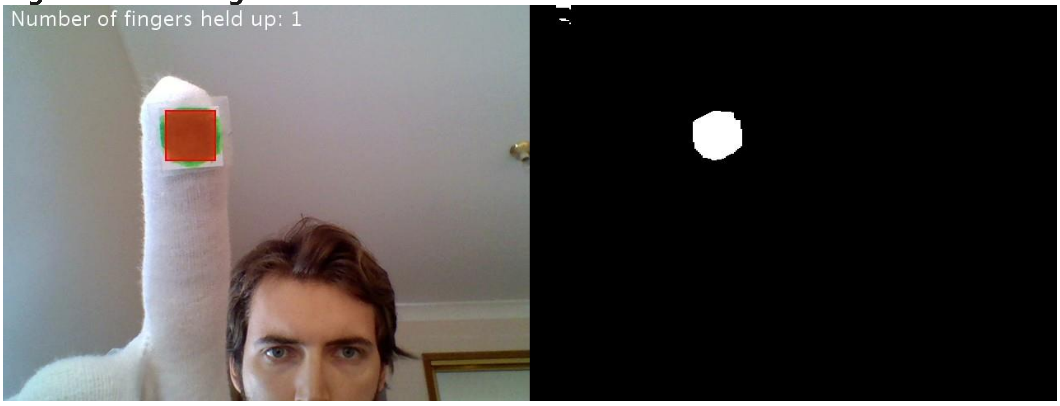
4.6. Contours

A contour line is a curve along which the function has a constant value. In this case the contour line is the white circle on a background of black, which is identified as the green dot on the glove. The amount of contours in the image corresponds to the amount of fingers held up, assuming the contour passes the bounding box requirement. This allows the representation of a number to be conveyed through sign language. An example of a white contour on a black background can be found in figure 5.

4.7. Bounding boxes

Bounding boxes are important and are the final step in determining the number of fingers held up. The bounding box is the minimum enclosing box for a given contour, subject to imposed constraints. If the bounding box is greater than the height and width constraints imposed, then the bounding box is valid and a finger is identified as held up. This final step eliminates any background noise which may have been missed during the erosion process. It also determines the distance at which the hand should be held form the camera in order to be correctly identified. If the glove is too far away from the screen, the bounding box will be too small, and the green dots will not be identified. An example of a bounding box for a single finger can be found in figure 5. The bounding box is seen as a red box overlayed on the green dot on the glove. The amount of bounding boxes in the image corresponds to the amount of fingers held up

Figure 5: Bounding box.



Number of fingers held up: 1

4.8. Geometric distance

Identifying the number of fingers held up requires counting the number of contours and establishing an appropriate bounding box, but this is only half the story. To identify additional hand gestures the relative positioning of fingers must also be known. Knowledge regarding the position of the bounding boxes is required. The relative position between bounding boxes is stored in an array, which corresponds to finger position. Then, using the distance between the x and y positions of the bounding boxes, the system is able to identify the hand gestures; "peace", "thumbs up" and "surf's up", according to the position of bounding boxes relative to one another.

4.9. Libraries

Two libraries were used to create the sign language interpretation system. Processing's video library (processing.video.*) and the open source Processing openCV library (gab.opencv.*). By leveraging from these libraries, a robust system was created which utilises existing functions from the respective libraries.

5. Versatility of the model

5.1. Background noise removal

Choosing an appropriate dot colour reduced the background noise. The webcam frame had very little green in the background, and hence green was chosen to improve the accuracy of the algorithm. Conceivably though, the user of the sign language interpretation system may be wearing green clothing, or may be outside amongst green objects, in which case the background noise may interfere with the accuracy of the algorithm. Hence the success of background noise removal depends upon the environment in which the web camera is used. A more versatile model which changes the colour identification based on the background noise may improve the accuracy and reliability of the system. The glove must also be versatile, where the coloured dots can be changed.

5.2. Geometric distance between fingers

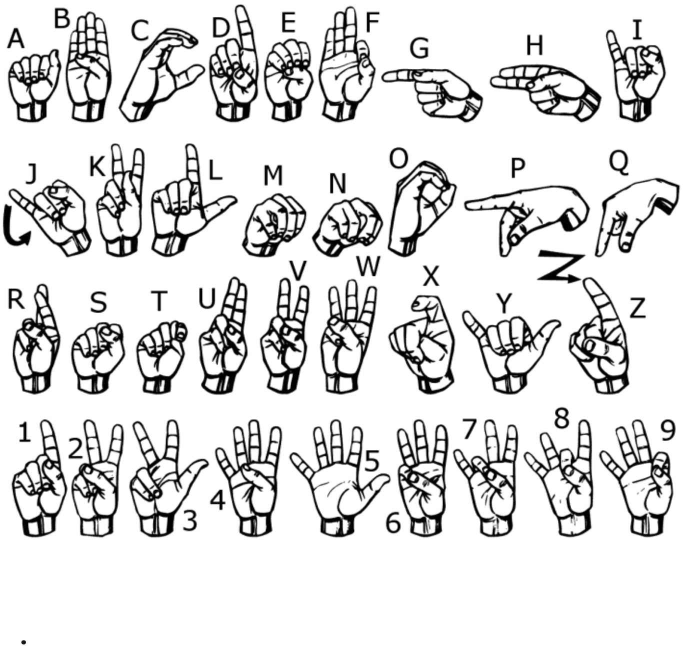
The geometric distance between fingers is used to determine the hand gesture being made by the person in the frame. However, there is natural variance in human hand size, some hands are larger than others. Hence, the algorithm is sensitive to the size and relative position of the hand.

6. Future Work

6.1. Sign language alphabet

Identifying each letter of the sign language alphabet would be a first step in comprehensively cataloguing all sign language gestures. With identification of each letter of the alphabet, any word may be spelt, and any sentence may be constructed. With just 24 characters, the sign language interpretation system could interpret any English word. An example of the sign language alphabet can be found in figure 6 [4]. Because most of the alphabet hand gestures are static, the complexity of the algorithm for detecting the letter is reduced. Using single colour dots for the glove fingertip, however, is problematic. Now the actual fingers help up (middle, index, etc.), rather than just the number of fingers held up, must be identified. This can be done by using different colours on each fingertip. The identification of the colour is associated with a fingertip, and accurately interpreted by the algorithm. Unfortunately, as mentioned above, when introducing additional colours the complexity of background noise also increases. The additional background noise must be overcome in order to create a robust, and correctly functioning system.

Figure 6: Sign language alphabet.



7. Conclusion

While prototype systems have been created, and provide effective communication between those who are deaf and those who are not deaf, no commercially available system exists. Microsoft's sign language interpretation system is promising, but has three major shortcomings. It is expensive, it is large, and it requires external technology, namely Microsoft Kinect. The sign language interpretation system will utilise only a web camera which is readily available on all smart phones. This approach is cheaper, smaller and does not require external technology. Those suffering from deafness, who wish to communicate with those who are not deaf, need not go any further than their own smart phones. The successful implementation of the sign language interpretation technology offers a new, effective and cheap interface between those who are deaf and those who are not deaf. The sign language interpretation system prototype suggests a promising future for those who are burdened by deafness.

[1] Burg, J., *The Science of Digital Media*, Pearson Education, 2008

[2] Processing. 2015. *Tutorials: Color*. [ONLINE] Available at: https://www.processing.org/tutorials/color/imgs/hsb.png. [Accessed 27 May 15].

[3] Gonzalez, R.C., Woods, R.E., *Digital Image Processing*, Prentice Hall, 2007

[4] Lifeprint. 2015. *Finger spelling*. [ONLINE] Available at: https://www.processing.org/tutorials/color/imgs/hsb.png. [Accessed 27 May 15].