Sign Language Recognition using Kinect

Edon Mustafa[1], Konstantinos Dimopoulos[2]

[1]South-East European Research Centre, University of Sheffield, Thessaloniki, Greece
[2]CITY College- International Faculty of the University of Sheffield, Thessaloniki, Greece
edmustafa@seerc.org, k.dimopoulos@city.academic.gr

Abstract. Sign language is used for communication among hearing impaired people. Their communication is hardly understood by normal hearing people. To facilitate communication between these two groups this paper presents a system that uses Microsoft Kinect device to build a translation system that translates signs from sign language to spoken language. The developed system uses SigmaNIL framework to provide hand shape recognition which is not supported by official and other SDKs for Kinect. The resulting system is capable of translating a limited vocabulary of Kosova Sign Language. The system has been tested with native and non-native speakers of sign language and in general achieves high accuracy varying by components of sign language.

Keywords: Sign Language Recognition Kinect Hand Shape

1 Introduction

Sign language is used by the community of Hearing Impaired (HI) people as the main means of communication. Compared to Normal Hearing (NH) people that use oral communication, HI people use visual signs that involve the use of both hands, the head, facial expressions and body postures. Exchange of information between a HI and a NH person is difficult since the NH person must be able to understand the sign language of the HI person. Because of the impairments that HI people have on hearing, the responsibility of learning the sign language falls on the NH person or alternatively a third person that understands both oral and sign languages and would do the necessary translation from one language to the other. A third alternative, is to use a computer to act as a translator. Ideally the computer system should watch HI person performing sign language and then translate it to speech for NH person; to accommodate of a dialogue, the system should also listen the NH person speaking and should translate it into a form that the HI person could understand; that is either sign language or text. Such a system has been developed here, focusing on the first translation direction: from signs to speech, since it is challenging and not developed at levels of being used in real world scenarios. The developed system employs the Microsoft Kinect device to infer necessary features for sign description. The Kinect

consist of various sensors to supports voice, movement and gesture recognition [1]. The skeleton tracking feature that is made possible through its depth sensor is used in this system [2]. In order to program the Kinect a number of SDKs exist, which offer skeletal tracking but not hand shape tracking (thus not allowing for finger level precision, which can be important for this application), with one exception: SigmaNIL. SigmaNIL wraps around other SDKs and offers the hand shape recognition feature [3]. Programming with SigmaNIL is challenging since it is on beta stage. It lacks the documentation, support and some functionalities like hand shape tracking of both hands and skeleton positions of other body parts. Solutions for these functionalities have been developed and they are presented at later section. The resulting system is capable of recognizing alphabet letters, digit numbers, words and sentences from the Kosova Sign Language (KSL). The system was tested by native speakers of KSL with recognition accuracy that depends on the component being recognized and depends on whether one hand or two were involved in sign making, whether they occluded each other and whether they included movement or not.

2 Background Research

2.1 Sign Language

Hearing impairments can vary from limited hearing to complete deafness. Since the process of learning how to speak involves the use of auditory feedback, people who are born with hearing impairments also face difficulties with speech even though they may have nothing physically wrong with their vocal system. Communication between two people with hearing impairments involves the use of hands to describe the shape of something, or to describe actions [4]. For example it is common (also to people with no hearing impairments) to express the action “go from this place to that place”, by pointing first to the source and “drawing” a line toward the destination. Communication is augmented with the simultaneous use of facial expressions and body postures together with hands gestures to fully express in this language any meaning [4]. There are many situations where hand gestures are used in place of oral communication. Babies may use simple gestures to express their needs before they learn to speak or adults in situations where speech is impossible or not appropriate. However, in contrast to these situations sign language is structured and has rules for composition and interpretation [4]. Sign language is taught to children with hearing impairments as a mother language [5], but it is a not universal language since as each vocally spoken language has its own sign language dialect [6]. This means for example that Greek sign language is different from English sign language. Specifically, a sign language mimics the fundamental properties that are present in the spoken language, like the grammar and vocabulary and it is used to express complex as well as abstract meanings [7]. Sign language is classified as a natural language and in many countries it is legally recognized.

Sign making involves the upper part of human's body, where parts of upper body are categorized as being manual or non-manual (automatic) features. Manual features involve hands [4] while non-manual features involves facial expressions (eye blinking, mouth shapes etc.), body postures and head movements [7]. Because manual features express most of the meaning in sign language and because of the limited time of the project the developed system addresses only manual features (those related to hands) while non-manual features although important for expressing grammatical features [8] they are not addressed in this system. Comparing sign languages with vocally expressed languages, there is a significant difference: signs (being visual in their nature) allow for simultaneous (parallel) processing of information, contrary to vocally expressed languages where only one sound can be perceived at a time, and thus are linearly processed [8]. However the amount of information conveyed in a given interval is the same with both languages [9]. Sign languages do not use articles (e.g. "*the*"), conjunctions (e.g. "*and*") and copulas (e.g. "It *is* raining") [8]. However there are also many vocally expressed languages that miss these features. Non-manual features are used for showing grammatical features like making questions, negatives [8] and showing boundaries between sentences [7]. Yes and no questions are signaled by raising the eyebrows, negatives by shaking the head and mouth shapes to signal degree (intensity) of something being communicated [7]. For example in American Sign Language to express the sentences "it is raining" and "it is *not* raining", the same gestures are used with the difference that in the second case, the head is horizontally rotated to a negation. The sign location is used to indicate tense (present, future, past). Signs performed near body refer to present, signs in front of the body to the future and signs behind the shoulders indicate something in the past [10]. Sign language is a natural language, very rich in features as vocally expressed languages.

2.2 Kinect

Kinect is voice and body recognition sensor that serves as controller for the XBOX gaming console, giving the users the ability to play games through voice and body gestures, without wearing or caring additionally accessories to track their body movements [11]. It features an RGB video camera, a depth sensor for 3D representation of the environment, a multi-array microphone for voice recognition and a property Microsoft software that enables human body recognition [1]. The software for skeletal tracking (tracking of the various body parts) is based on the data that comes from depth camera. It allows tracking of up to six people in its field of view, while offering full skeleton tracking for two of them. In full skeleton tracking mode, twenty (20) joints are tracked, while in seated mode, half of them [12]. The accuracy depends on the distance of the person from the camera where the distance from 1.2 meters to 3.5 meters gives the most accurate position of skeleton joints [12]. Joints are recognized as 3D points (with x, y, z coordinates), with the Kinect placed at the origin of the coordinate system and from Kinect viewpoint: positive z-axis increases towards the user, positive y-axis increases upward and positive x-axis increases to the left [13]. This is shown in figure 1.

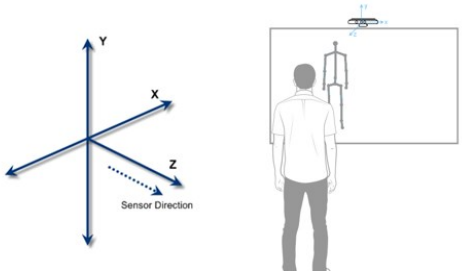

Figure 1 - Kinect skeleton coordinate space

The process of enabling skeleton tracking feature at Microsoft started by taking depth images for which body parts were known and used them to generate more depth images in order to cover all body types and positions [14]. So, 100,000 real depth images taken using motion capture system were used to generate one million depth images using computer graphic techniques [14]. Randomized decision forests were used to match these newly generated images with corresponding body parts [14]. Finally at the end these body parts were transformed into joint positions by using mean shift algorithm that finds densest region in each body part and considers it as joint position [14]. Hand shape is crucial in sign making and hand shape recognition is not offered through the standard SDKs. SigmaNIL is a framework that is capable of recognizing hands and has been used to enable the hand shape recognition. Its use will be detailed further in sections below that describe how the system was implemented.

2.3 Sign Language Recognition Process

In the literature there are many different approaches for the process of automatically recognizing sign language, but in general the process has three phases: the data acquisition phase, followed by the feature extraction phase and finally the analysis phase which concludes whether or not the combination of features that describe a particular sign happened. The process is depicted in figure 2.

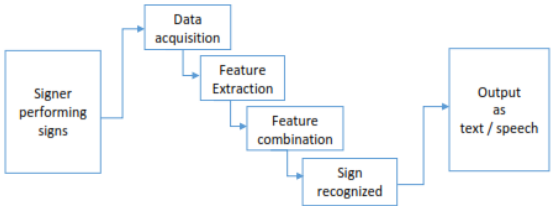

Figure 2 - Sign language recognition process

During the data acquisition phase, information about the movements of the user are collected. This may be achieved in one of two ways: the first, requires from the sign language performer to wear sensors (usually accelerometers) in body parts that are involved in sign making. These are usually gloves with embedded sensors that transmit the hand configuration to the connected computer wirelessly. These kind of

systems have been developed in [15], [16], [17] where gloves were connected to a computer for sign information transmission. Another system that required gloves was built in [18] but this time the glove transmitted information wirelessly to the computer. This approach is cumbersome and not convenient although it transmits accurate information. The second approach is more advanced as it is based on computer vision and offers a more natural interaction. Sign features in this category are derived by image and sensor data processing (from cameras). Such system has been developed in [19] where two ways were used for feature extraction: the first extracted features based on hand's skin color while the second required colored gloves for better segmentation of the hand. Computer vision techniques allows for more flexibility. For example the developed system [20] used two experiments: in the first the camera was placed on the desk and in the second the camera was placed on a cap worm that user had to wear; in another system [21] three camera were placed orthogonally to extract features in 3D. The second approach (vision based) is more favorable also in the cases were non-manual features of sign language are considered [22], [23] since the first approach would require wearing of cap worm for sensing head movements and facial expressions. The Kinect device fits in the second category. Kinect as presented earlier has two cameras: an ordinary one and a depth camera. Usually in SLR systems Kinect's depth camera is used to infer human skeleton positions. Kinect has been used in CopyCat game developed by CATS which tries to improve memory skills of HI children by allowing them to practice sentences in ASL [24]. In another system Kinect is part of humanoid robot NAO and is used to teach sign language to HI children: the robot performs signs, encourages the children to imitate and then watches whether the sign was performed correctly [25]. Some other SLR systems with Kinect are [26], [27], [28], [29], [30], [31] and [10]. The second phase (feature extraction) involves the use of the information stream that was acquired at the first phase, in order to identify important features (like the hand positions). The final phase uses the features that where extracted at the second phase in order to conclude which sign was performed. As it was discussed previously, sign language is multimodal, where information is conveyed simultaneously from different modes at the same time (hands move at the same time, may have different hand shapes, while at the same time head is moving and different facial expressions are present). To achieve correct analysis in the third phase usually Hidden Markov Models (HMM) are used [32]. HMM have been successfully used in speech recognition [33], but for sign recognition they have to be adopted for multimodal use, and this may be challenging [32]. Such systems have been developed in [16], [19], [20], [21], [22] and [10]. An alternative could be Artificial Neural Networks (ANNs) [32]. Sign combination can be seen as patterns. ANNs try to mimic the brain functionality for pattern recognition and simultaneous processing of information from different channel [34]. Some of the systems that use ANNs are [17], [35], and [28]. A more complete cover of SLR systems is available at [32], [35] and [36]. However this method has not had a great success. Our system does not use any of the above mentioned methods but rather employs an ad-hoc approach since the interest is inclusion and testing on hand shape feature in Sign Language Recognition (SLR) systems that are build using Kinect. This method has the advantage that it is simple in

getting results, but within a limited range of movements, as each hand gesture has to explicitly programmed at the system.

3 The Developed System

3.1 Description of the Arrangement

The idea is to facilitate the communication between HI and NH people by building a translation system that uses Kinect. The system will play the role of human sign language interpreter that understands the sign language and translates it into speech for NH people. Idea visual representation of the proposed arrangement is illustrated in figure 3. Three parties will be involved: the HI person, system that consist of Kinect and a computer and NH person. The HI person perform signs in front of the Kinect. Kinect tracks the HI person and transmits sign information to the computer. Computer analyzes the sign and provides its meaning as speech to NH person.

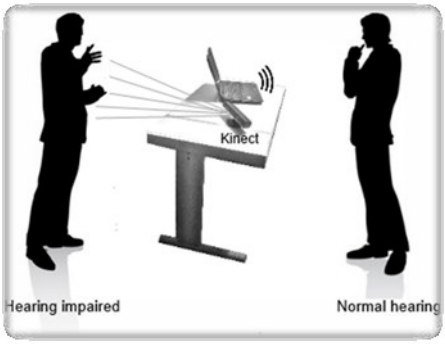

Figure 3 – Proposed system. The Hearing Impaired person performs signs in front of a Kinect, which translates the signs to text in a computer screen for a Normal Hearing person to read.

3.2 Constraints

As discussed in section 2.1, signs are composed from manual and non-manual features (such as eye blinking). In general, manual features involve using hands while non-manual features involve facial expressions, body postures and head movements. Due to time limitations the developed system deals only with recognition of manual features. Further more, the system is thought to be functional only in in-door environments, since the infrared lights that Kinect employs as part of depth measuring are destructed by sun-lights in outdoor environments.

3.3 System Design

From physical viewpoint the system consists of two components: the Kinect device and a computer. Kinect senses the scene and transmit the information to the computer where they analyzed and concluded which sign was performed. The developed software components are located on the computer only. The Kinect sensor supplies raw RGB and depth sensor's data which then are processed by software components residing on the computer. The software components are organized in a layered architecture. At the lowest layer are base SDKs that take raw data and construct RGB and depth images and then from depth images they infer skeleton positions. Going one layer up, the SigmaNIL framework utilizes these data to generate hand related features (e.g. the shape of the hand). Finally at the top of the hierarchy resides the developed components of the system and these components utilize skeleton positions and hand shape features to derive additional features like hand location and hand movement direction in order to describe a sign. In this layered architecture the communication is event driven. SigmaNIL waits for events from underlying SDKs and the developed application waits for events from SigmaNIL. The developed system consists of these components: the hand shape recognition component, hand position relative to other body parts, hand movement direction and text-to speech.

3.3.1 Hand Shape Recognition

Many alphabet letters and digit numbers of a sign language can be recognized if only the hand shape is recognized. The process of hand shape recognition passes by in two phases: first the selected hand shapes to be recognized are recorded, labeled and put in the SigmaNIL Training tool. This is a build-in tool that uses an algorithm for hand shape recognition and as a final output gives a database file with rules for distinguishing different hand shapes. Then this file is utilized in the software, where for each recognized shape a unique label is generated. SigmaNIL framework is in beta stage, with lack of documentation, support and furthermore it does not offer hand shape recognition in both hands (limited to one). Nevertheless, a way has been found to get recognition of hand shape in both hands at the same time. This was done by using the same database of hand shapes but introducing two additional SigmaNIL engines (segmentation and shape) for other hand, a solution similar (not identical) as for recognition of shapes with one hand and of course with a side effect on decreasing the performance, however not very noticeable. Hand shape recognition is provided by SigmaNIL framework. However the shapes that it will recognize must be pre-selected and they have to go through a training phase before they are able to be recognized (the training phase is elaborated more in system implementation section). Since recognition of all hand shapes is a long process a set of twenty (20) hand shapes were selected. Among selected shapes are those that enable recognition of basic sign language components and the shapes that may be used in different signs. However in a fully developed system, more hand shapes can be used. Using hand shape recognition the system can recognize all letters and numbers.

3.3.2 Hand Position in Relation to Each Other and to Other Body Parts

Hand shape identification alone, is not enough for recognition of larger sets of signs. Another important aspect is hand position relative to other body parts. This involves checking the position of hands relative to each other (are they near or are they touching and in which direction), to the head or to the chest and many more. SigmaNIL does not offer this feature, and therefore it was necessary to augment with appropriate code. Detection of hand position relative to each other was done by using a 2D representation of hands as they are presented in computer graphic coordinative system. A rectangle border around each hand was defined. The intersection of the two borders around the hands was used to define whether the hands touched each other. Distance between hands is measured as Euclidean distance between points of interest and if the distance decreased bellow a threshold it was concluded that the hands intersected. Figure 4 shows how the proximity of two hands is calculated: first two rectangular borders are drawn around each hand (figure 4.a and 4.b). The top two corners that are closer to each other are used then as points of interest, and the x and y difference is then calculated (figure 4.a and 4.b). To calculate the intersection between the two areas, the same borders are used. This is shown in figure 4.c and 4.d.

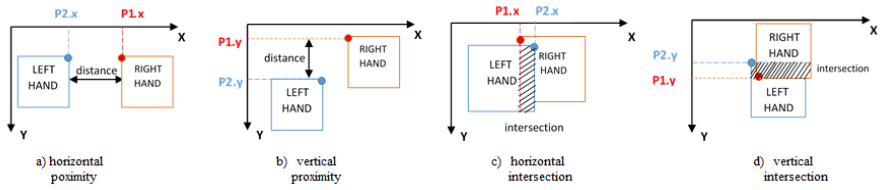

Figure 4 – Calculation of hands proximity: a) horizontal, b) vertical and hands intersection c) horizontal and d) vertical

In order to identify the relative position to other body parts (like the head, the chest and not between hands) the SigmaNIL framework had to be modified to draw borders around three body parts of special significance: the head, the torso and the inner torso (see figure 5.a). However these are larger areas compared to the hands and determining hand relative positions is done by introducing an additional method, by identifying that a hand is completely with in the border of a larger area. This is shown in figure 5.b.

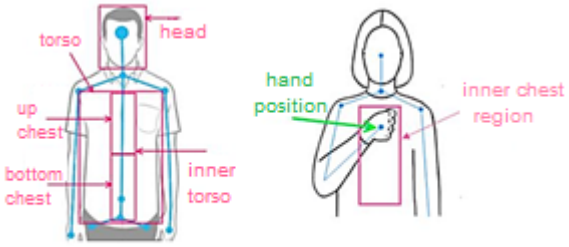

Figure 5 – Hand position relative to body parts: a) important regions and b) determining hand position

3.3.3 Hand Movement Direction

Hands can move in one of three directions in space: right-left, up-down, and away-towards the user. These are seen as basic movements of hands although more advanced movements exist. Movement can be understood as change of hand position from one location to the other. Since hand position is provided as skeleton position from Kinect SDK, implementation of this feature requires first understanding of Kinect skeletal coordinate system. The origin of the coordinative system is at the Kinect center as it is shown in figure 1. Each skeleton joint is represented by three dimensional (x, y, z) point. Furthermore as illustrated in figure 1, the x dimension represents the horizontal direction, the y dimension represents the vertical direction and the z dimension represents the direction from the Kinect towards the users' body. So movement of the hand along the x dimensions can be represented as *left* or *right* movement, movement along the y dimension as *up* or *down* movement and movement along the z dimension as *away from user* or t*owards the user* movement.

3.3.4 Alphabet, Number, Word and Sentence Recognition

The alphabet and number recognition can be demonstrated by showing the recognition of signs that involve one and two hands. For recognition of one hand signs, hand shape recognition is enough. This is showed in figure 6 that shows recognition of letter C (figure 6.a) and number one and number five (figure 6.b). For recognition of two hand signs within alphabet and number category, relative position of hands is also important. The figure 6 shows also the recognition of number nine (figure 6.c), letter A (figure 6.d) and letter G (figure 6.e).

Figure 6 – Alphabet and number recognition: a) letter C, b) number 1 and number 5, c) number nine, d) letter A and e) letter G

The word recognition is demonstrated by showing the recognition of the word "HELLO". This word is a composition of a particular hand shape and a move in a particular direction starting from a particular body position. The hand shape, movement direction and hand location are shown in figure 7. To recognize this word, first is checked if the hand has this particular shape (similar to number four). Next is checked the location of the hand in relation to the head. If the sign is performed with the right hand as in figure 7, the hand must touch the right side of the head. Then after the hand touches the head, it must move in two directions at the same time: away from the user (toward Kinect) and a little up. The implementation is done by starting a timer when the hand touches the head. Then if within four seconds the hand moves

away from the user and little up it is concluded that sign for “HELLO” word is implemented. In a similar way other words were programmed. Composing sentences is then an issue of recognizing consecutive words. Sentences are treated as sequential occurrence of words that happen within a period of time. Sentence recognition is demonstrated through following example. The figure 8 shows sequential occurrences over the time of the signs that represent two words. The first is the sign for “HELLO” word and the second is the sign for “DAUGHTER” word, both in KLS. When they happen sequentially within a period of time they form the sentence “HELLO DAUGHTER”The recognition of sentence “HELLO DAUGHTER” is done by starting the timer when the HELLO signs is performed and if within five seconds if sign DAUGHTER happens, it is concluded that sentence “HELLO DAUGHTER” was performed. The last component provided is sign to speech translation. Each recognized sign is translated into spoken language and outputted as a text and as speech. This section concludes the functionalities provided in this system.

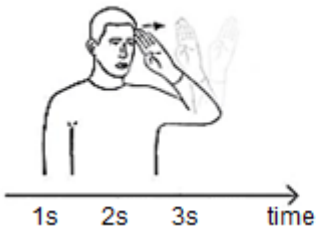

Figure 7 - Recognition of *Hello* word

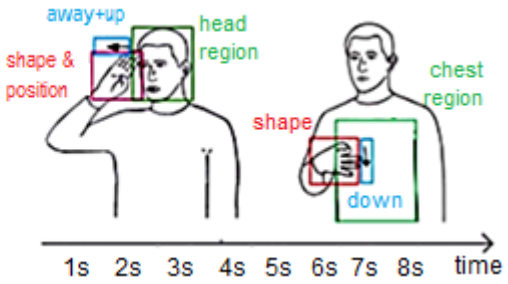

Figure 8 - Recognition of sentence: *Hello Daughter*

3.4 Testing and Evaluation of the System

The functionalities for which the system is tested are: number, alphabet letters, words and sentence recognition and text to speech as well. The system was tested under different light conditions, in real time. The system was tested from two (2) native speakers of sing language and one (1) non-native speakers of sign language. The testers were of different ages, body types and sizes. Different testers assure testing for different sizes and shapes of hands thus providing more insights into using the hand shape feature in systems for sign language recognition. The system was tested for nine numbers, fifteen alphabet letters, four words and one sentence. Testers performed each signs ten (10) times. The non-native speakers had to be trained to perform the signs whereas non-native speakers were required just to perform the signs. The results from testing are shown in table 1 and graphically presented in figure 9. In table 1 the second column lists the tested signs and three other columns lists the accuracy results from three different system testers. Each time the tester performed the sign, it was observed whether the system correctly recognized it and if it did it was counted. As it is shown in table 1, the sign for number one was performed ten times by each tester and in the case of the first tester the system recognized correctly it each time, in case

of second tester the system recognized correctly it nine (9) times and in the case of the last tester the sign was recognized correctly eight (8) times. In the case of the sign for letter A and B the tests were canceled for native speakers because they were not recognized at all after initial testing with non-native tester. This happened because of hand occlusion where one hand was occluded by the other and not observable by Kinect viewpoint.

Table 1 - Results from testing

Sign	Non-native tester	Native tester 1	Native tester 2
Number 1	10/10	9/10	8
Number 2	10/10	7/10	9/10
Number 3	10/10	10/10	10/10
Number 4	9/10	8/10	8/10
Number 5	10/10	10/10	10/10
Number 6	10/10	1/10	3/10
Number 7	10/10	7/10	9/10
Number 8	10/10	9/10	10/10
Number 9	8/10	10/10	10/10

Sign	Non-native tester	Native tester 1	Native tester 2
Letter A	5/10		
Letter B	5/10		
Letter C	10/10	9/10	10/10
Letter E	8/10	8/10	10/10
Letter L	10/10	10/10	10/10
Letter LL	8/10	10/10	10/10
Letter O	10/10	10/10	10/10
Letter U	9/10	7/10	4/10

Sign	Non-native tester	Native tester 1	Native tester 2
Word: OK	8/10	6/10	6/10
Word: Hello	9/10	10/10	10/10
Word: Daughter	8/10	8/10	8/10
Word: Are you OK ?	7/10	10/10	8/10
Sentence: Hello daughter	4/10	8/10	10/10

The graph in figure 9 presents the recognition accuracy for each component of sign language for each tester. It can be seen that accuracy is higher in case of non-native speaker for number and alphabet letters while it decreases in case of words and sentences.

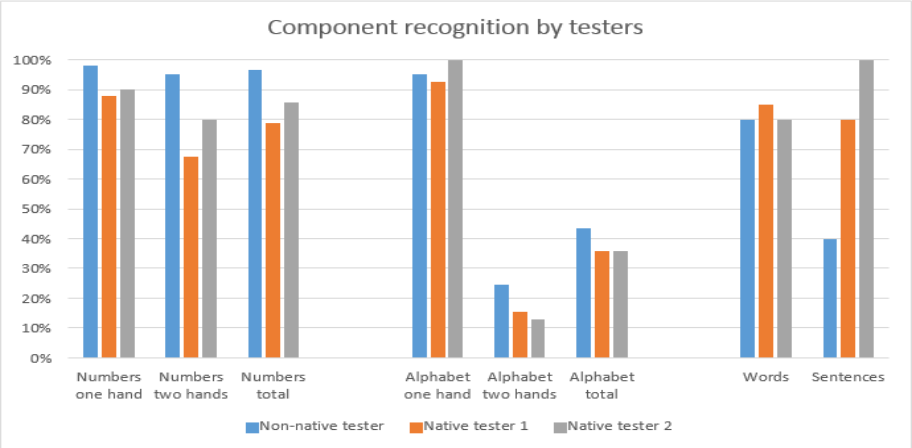

Figure 9 – Graphical representation of test results for each component of sign language for each tester

Higher accuracy for non-native speaker can be explained from the fact that non-native speaker was trained to perform signs from system developers and thus un-intentionally influenced the way signs were performed. In the other side, native speakers have interpersonal variation in performing same signs and thus the accuracy varied. However in the case of more advanced signs like words or sentences the system recognition accuracy with native speakers was higher. One possible explanation is that more complicated signs are harder to learn for non-native speakers.

As mentioned earlier in this section the system had difficulties in correctly recognizing signs that involved both hands and especially cases where hands occluded each other. This is shown also in the graph where the recognition accuracy for signs with one hand is much higher.

3.4.1 Evaluation of the System

The system is able to recognize numbers from 1 to 9 in sign language with high accuracy. However not all numbers of sign language were recognized such as numbers starting from ten and up. In addition, the KSL alphabet consist of thirty-six alphabet letters. From them fifteen were implemented and tested. The system showed high recognition accuracy for alphabet letters that involved one hand. In cases where two hands are involved the recognition accuracy drops. This holds for signs when hands touch each other, while in the signs where hands do not touch (near only) the accuracy is similar to alphabet letters made with one hand. Furthermore the alphabet letters that were combination of other alphabet letters were not implemented at all. The system implements a limited vocabulary of words and sentences. More specifically four words and one sentence were implemented and tested. Although recognition accuracy is higher and at acceptable levels, the system does not employ common methods (HMMs or ANNs) for word and sentence recognition that could have increased recognition accuracy. As an additional consequence the system is not easily scalable for addition of new words and sentences. Sign languages incorporate non-manual features (facial expression, body posture, hand movement) that were not taken into consideration in this project. The capability for continuous recognition of sign language communication is not provided. Continued communication in sign language does not require recognition of alphabet letters, since sequential combination of words is what forms sentences that are used for continuous communication. Sentence recognition was tested for one sentence only and followed a simple algorithm that is not scalable for inclusion of new sentences. Although SigmaNIL proved to have good potential for hand shape recognition it had to be modified in order to provide hand shape recognition in both hands and also to proivde tracking of other joints than hands. These two modification resulted in decrease on system performance (fortunatley not noticable by users) and deviation from coding standards (implicatoin on future system releases). Beside these facts it is very promising framework to be used in building SLR systems and hopefully it will be enhanced in next releases.

4 Future Work

The system can be improved further either by enhancing existing futures or by providing new ones. The recognition accuracy of hand shape component can be increased further by playing with parameters of the algorithm that is available through SigmaNIL training tool. Furthermore recognition accuracy can be increased by employing proven methods like HMMs, ANNs and other that were used in similar systems (refer to section 2.3). Vocabulary of recognized hands shapes can be

increased to include large data set. The hand position relative to body feature can be enhanced further by defining all necessary hand position to be detected and then by enhancing the algorithms for detection of these positions. The 2D and 3D feature were used to detect different hand positions. Implementing algorithms that are based only in 3D futures would increase the performance by eliminating the need of conversation of 3D into 2D and then using 2D for comparison purposes. Hand movement feature also can be enhanced further. It provides recognition of six basic hand movements that can be combined to detect more movement directions. Identification of other movement directions involved in sign making and their detection would provide the possibility for recognition of larger vocabulary of signs. While numbers, alphabet letters and words recognition can be enhanced by enhancing these features, sentence recognition requires implementation of more sophisticated mechanism like HMMs, ANNs and other methods, in order to provide continuous SLR. Finally the system can be enhanced further by providing new features, like bi-directional translation and tracking of non-manual features. While incorporation of non-manual features is important, providing the speech to signs translation will definitively provide better communication between hearing impaired and normal hearing people.

5 Overview and Conclusions

The Kinect sensor was investigated for building a system that translates sign language to spoken language. While sign language uses manual and non-manual signs to conveying meaning in this work only manual signs were investigated. Manual signs are composition of hand shape, hand location, hand orientation, and hand movement. The developed system incorporates these components and uses them to recognize and translate to speech: numbers, alphabet letters, words and sentences from Kosovo Sign Language. Four components of this language were recognized and tested: numbers from 1 to 9, 15 alphabet letters, four words and one sentence. The system is tested by two (2) native speakers of sign language and one (1) person without prior knowledge on SL to observe the effect of sign language proficiency in recognition accuracy of the system. Results of testing system with native and non-native speaker showed that proficiency with sign language and interpersonal variation are important factors that have impact on recognition accuracy. Thus recognition accuracy of non-native speaker was higher for alphabet letters and numbers explained by the fact that they were easy to learn and since the tester was trained by system developers, they possibly influenced the tester in performing signs in the way system expects. In the other side recognition accuracy of native speakers for words and sentences was higher and can be attributed to their proficiency in sign language while variation in the accuracy to interpersonal variations. In general the system achieved higher recognition accuracy for one hand signs and lower recognition accuracy for signs where both hands were involved. While the recognition accuracy for static signs was high (around 90%), further work is needed in order to improve recognition accuracy of hand shape component when it is combined with hand movement and hand location features. In the other side, although Kinect is well suited for feature

extraction, proper software recognition method has to be followed in order to understand continuous communication in sign language. Finally as most of conducted researches do not incorporate the hand shape component, the incorporation and investigation of it in this work is seen as main contribution in the area of Sign Language Recognition systems with Kinect.

References

[1] Microsoft, "Kinect for Windows, Sensor Components and Specifications," 2013. [Online]. Available: http://msdn.microsoft.com/en-us/library/jj131033.aspx. [Accessed: 06-Dec-2013].

[2] M. Andersen, T. Jensen, and P. Lisouski, "Kinect depth sensor evaluation for computer vision applications," 2012.

[3] SigmaNIL, "SigmaNIL: Technical documentation of the framework." [Online]. Available: http://www.sigmanil.com/docs/SIGMANIL API DOCvs1.htm. [Accessed: 06-Dec-2013].

[4] W. C. Stokoe, "Sign language structure: an outline of the visual communication systems of the American deaf. 1960.," *J. Deaf Stud. Deaf Educ.*, vol. 10, no. 1, pp. 3–37, Jan. 2005.

[5] K. Emmorey, *Language, cognition, and the brain: Insights from sign language research*. Psychology Press, 2001.

[6] M. P. Lewis, G. F. Simons, and C. D. Fening, "Ethnologue: Languages of the world (17th edition)," SIL international Dallas, TX, 2013.

[7] S. K. Liddell, *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, 2003.

[8] R. Battison, "Lexical Borrowing in American Sign Language.," 1978.

[9] U. Bellugi and S. Fischer, "A comparison of sign language and spoken language," *Cognition*, 1972.

[10] S. Lang, M. Block, and R. Rojas, "Sign language recognition using kinect," *Artif. Intell. Soft Comput.*, pp. 394–402, 2012.

[11] Microsoft, "Kinect Fact Sheet," 2010. [Online]. Available: www.microsoft.com/enus/news/presskits/xbox/docs/kinectfs.docx.

[12] Microsoft, "Kinect for Windows, Human Interface Guidlines." [Online]. Available: http://msdn.microsoft.com/en-us/library/jj663791.aspx. [Accessed: 06-Dec-2013].

[13] Microsoft, "Kinect Coordinate Space." [Online]. Available: http://msdn.microsoft.com/enus/library/hh973078.aspx#Depth_Ranges. [Accessed: 06-Dec-2013].

[14] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[15] T. Takahashi and F. Kishino, "Hand gesture coding based on experiments using a hand gesture interface device," *ACM SIGCHI Bull.*, vol. 23, no. 2, pp. 67–74, 1991.

[16] C. Lee and Y. Xu, "Online, interactive learning of gestures for human/robot interfaces," in *Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on*, 1996, vol. 4, pp. 2982–2987.

[17] J. Weissmann and R. Salomon, “Gesture recognition for virtual reality applications using data gloves and neural networks,” in *Neural Networks, 1999. IJCNN’99. International Joint Conference on*, 1999, vol. 3, pp. 2043–2046.
[18] K. Thomas, “Glove lends the deaf a hand,” *USA Today. Retrieved Oct.*, vol. 10, p. 2002
[19] T. Starner and A. Pentland, “Real-time american sign language recognition from video using hidden markov models,” in *Motion-Based Recognition*, Springer, 1997
[20] T. Starner, J. Weaver, and A. Pentland, “Real-time american sign language recognition using desk and wearable computer based video,” *Pattern Anal. …*, vol. 20, no. 12, pp. 1371–1375, 1998.
[21] C. Vogler and D. Metaxas, “ASL recognition based on a coupling between HMMs and 3D motion analysis,” in *Computer Vision, 1998. Sixth International Conference on*, 1998, pp. 363–369.
[22] D. Kelly, J. Reilly Delannoy, J. Mc Donald, and C. Markham, “A framework for continuous multimodal sign language recognition,” in *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009, pp. 351–358.
[23] P. Dreuw, H. Ney, G. Martinez, O. Crasborn, J. Piater, J. M. Moya, and M. Wheatley, “The SignSpeak Project - Bridging the Gap Between Signers and Speakers,” 2010.
[24] Z. Zafrulla, H. Brashear, and H. Hamilton, “American sign language recognition with the kinect,” *Work*, pp. 279–286, 2011.
[25] H. Kose, R. Yorganci, E. H. Algan, and D. S. Syrdal, “Evaluation of the Robot Assisted Sign Language Tutoring Using Video-Based Studies,” *Int. J. Soc. Robot.*, vol. 4, no. 3, pp. 273–283, Mar. 2012.
[26] F. Huang and S. Huang, “Interpreting American Sign Language with Kinect,” pp. 1–5, 2011.
[27] E. Rakun, M. F. Rachmadi, K. Danniswara, and others, “Spectral domain cross correlation function and generalized Learning Vector Quantization for recognizing and classifying Indonesian Sign Language,” in *Advanced Computer Science and Information Systems (ICACSIS), 2012 International Conference on*, 2012, pp. 213–218.
[28] F. Trujillo-Romero and S.-O. Caballero-Morales, “3D data sensing for hand pose recognition,” in *Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on*, 2013, pp. 109–113.
[29] M. Oszust - “Polish sign language words recognition with Kinect,” in *Human System Interaction (HSI), 2013 The 6th International Conference on*, 2013, pp. 219–226.
[30] H. Takimoto e “A Robust Gesture Recognition Using Depth Data.”
[31] D. Capilla, “Sign Language Translator using Microsoft Kinect XBOX 360 TM,” *Dep. Electr. Eng. Comput. …*, 2012.
[32] H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” in *Visual Analysis of Humans*, Springer, 2011, pp. 539–562.
[33] P. Dreuw and D. Rybach, “Speech recognition techniques for a sign language recognition system,” …, Figures 7 and 8 have some problems.2007.
[34] Y. F. Admasu and K. Raimond, “Ethiopian sign language recognition using Artificial Neural Network,” in *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on*, 2010, pp. 995–1000.
[35] P. Vijay and N. Suhas, “Recent Developments in Sign Language Recognition: A Review,” *irdindia.in*, no. 2, pp. 21–26, 2012.
[36] M. Majid and J. Zain, “A REVIEW ON THE DEVELOPMENT OF INDONESIAN SIGN LANGUAGE RECOGNITION SYSTEM,” *J. Comput. Sci.*, vol. 9, no. 11, pp. 1496–1505, 2013.