Deaf Talk Using 3D Animated Sign Language

A Sign Language Interpreter using Microsoft's Kinect v2

Mateen Ahmed, Mujtaba Idrees, Zain ul Abideen, Rafia Mumtaz, Sana Khalique
NUST School of Electrical Engineering and Computer Science
National University of Science and technology (NUST)
Islamabad, Pakistan
{11besemahmed, 11besemidrees, 11besezabideen, rafia.mumtaz, sana.khalique}@seecs.edu.pk

***Abstract*—this paper describes a neoteric approach to bridge the communication gap between deaf people and normal human beings. In any community there exists such group of disable people who face severe difficulties in communication due to their speech and hearing impediments. Such people use various gestures and symbols to talk and receive their messages and this mode of communication is called sign language. Yet the communication problem doesn't end here, as natural language speakers don't understand sign language resulting in a communication gap. Towards such ends there is a need to develop a system which can act as an interpreter for sign language speakers and a translator for natural language speaker. For this purpose, a software based solution has been developed in this research by exploiting the latest technologies from Microsoft i.e. Kinect for windows V2. The proposed system is dubbed as Deaf Talk, and it acts as a sign language interpreter and translator to provide a dual mode of communication between sign language speakers and natural language speakers. The dual mode of communication has following independent modules (1) Sign/Gesture to speech conversion (2) Speech to sign language conversion. In sign to speech conversion module, the person with speech inhibition has to place himself within Kinect's field of view (FOV) and then performs the sign language gestures. The system receives the performed gestures through Kinect sensor and then comprehends those gestures by comparing them with the trained gestures already stored in the database. Once the gesture is determined, it is mapped to the keyword corresponding to that gesture. The keywords are then sent to text to speech conversion module, which speaks or plays the sentence for natural language speaker. In contrast to sign to speech conversion, the speech to sign language conversion module translates the spoken language to sign language. In this case, the normal person places himself in the Kinect sensor's FOV and speaks in his native language (English for this case). The system then converts it into text using speech to text API. The keywords are then mapped to their corresponding pre-stored animated gestures and then animations are played on the screen for the spoken sentence. In this way the disable person can visualize the spoken sentence, translated into a 3D animated sign language. The accuracy of Deaf Talk is 87 percent for speech to sign language conversion and 84 percent for sign language to speech conversion.**

Keywords—Sign language; Sign to speech; Kinect; Deaf talk; Sign language interpreter

I. INTRODUCTION

People suffering from hearing and speaking disabilities use sign language as simple mode of communication among themselves and with the rest of the world, but unfortunately not every one of the normal people knows sign language, hence, the end result is lack of communication and isolation. According to the surveys conducted by World Health Organization (WHO), over 360 million people in the world suffer hear loss1 and 120,000 are born deaf-mute each year2. According to estimates, it covers around 5.3% of the world population and 91% among them are adults. In order to help people with such disabilities a lot of research has been conducted and some solutions [1] [7] have been made worldwide so far but no major success has been reported hitherto.

Deaf Talk, a sign language interpreter has been proposed which offers a natural way of communication to the hearing and speaking impaired individuals, simplifying their onerous task of communication. This is a dual mode system in a sense that it translates the natural language to its corresponding sign language showing the required gestures on screen and then uses the gesture and pattern recognition to translate sign language to spoken language. The system is based on the latest Microsoft technology of "Kinect v2" which is able to track motion, depth and gestures. The developed system can help the hearing and speaking impaired people to communicate with the world and overcome their disabilities.

The paper explains the two modules of sign language translation i.e. sign to speech translation and speech to sign translation. The tools and techniques along with the components used have been described in detail, giving a complete technical and logical insight of the methodology for each of the modules. Towards the end, results and accuracy of the system is explained which helps evaluating the goodness of the proposed approach and obtained results.

II. RELATED WORK

With Kinect or such 3D cameras, natural human computer interaction has achieved a big improvement in a lot of fields [2] [12] [14] [3]. A number of related researches have been done, such as human detection and 3D modelling [8], human pose and action recognition [10] [11], hand tracking and gesture recognition [5].

Various devices and sensors have been used for detecting sign language. Some that came up with considerable results were Microsoft's Kinect, gloves with sensors, multicolor

[1] WHO, 1995. Executive Summary Child Health
[2] WHO, 2012. Global Estimate Hearing loss

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

cameras, etc [13]. One of the most adaptive technique used for gesture recognition was through depth based information. Main reason for this was user friendly, interactive and cost effective development of depth sensors [6].

Accuracy is more when it comes to depth sensing algorithms in comparison with other techniques and also it covers way more vocabulary than others [6]. With Microsoft's Kinect in the market, many attempts are made to recognize human gestures, especially for hands. Ref. [14] used the depth sensor of Kinect to recognize around thousand phrases from American Sign Language and this recognition is based on hidden Markov model (HMM). Ref. [6] also focused on Kinect base hand gestures recognition. Depth image sensing is also used for recognizing hand gestures but it remained limited to hands only [9]. Whole body movement was not focused in this research.

This domain was not restricted to Kinect based solutions only, many other techniques and methods emerged in this research as well. Google Gesture[3] is a concept which is not under development yet. The concept is to help people with hearing disabilities but its proposed implementation consists of a wrist band which will help in symbol recognition with the help of muscle contraction and relaxation which is not natural and as usable as Deaf Talk, which doesn't require you to wear any gadget for gestures recognition. MotionSavvy[4] is a device that is made using Leap Motion Technology that will help deaf people in their daily life.

It costs around $200. This project has been made available for pre-order in 2014. The limitation of MotionSavvy is, it works only with hands and not capable to handle certain gestures like touching ear, eye or nose etc.

Enable talk[5] is another device that contains two gloves fixed with sensor and a mobile device. This project is in research since 2012. Since then, no prototype has been released and no public announcement made regarding this project.

Microsoft Research has been funding a similar project based on Kinect v1 for Windows [1]. This project is under development in Chinese Academy of Sciences for more than two years. Its first prototype was hailed publically. However, with the release of Kinect v2 for Windows, the work based on Kinect v1offers compromised or limited sign recognition as compared to Kinect v2 due to following reasons. The resolution of the camera has been improved from 640×480 pixels to 1920×1080 pixels, hence improving the detection. Audio input is increased from 16bit channel to 32bit channel increasing the audio recognition quality. Horizontal FOV is increased from 57 degree to 70 degree and vertical FOV is increased from 43 degree to 60 degree thus providing large area coverage. In Kinect v1 only 20 bones of human skeleton can be detected. The quality of skeleton detection has been enhanced in Kinect v2 by sensing 26 bones including 2 more bones in each hand. This feature specifically comprehends fingers movement, which is an essential component of sign language.

III. Methodology (Speech to Sign Module)

This part mainly focuses on real implementation of the proposed system. The system has been divided into two separate modules, speech to sign module and sign to speech module. This part explains "speech to sign" module. This implementation is carried out in controlled environment keeping the track of all the variables and possibilities in real scenarios. The implementation details are as under.

A. Implementation Approach

The strategy used to implement this is illustrated in Fig. 1. System shall get the speech input that will be processed and get converted to the related text form.

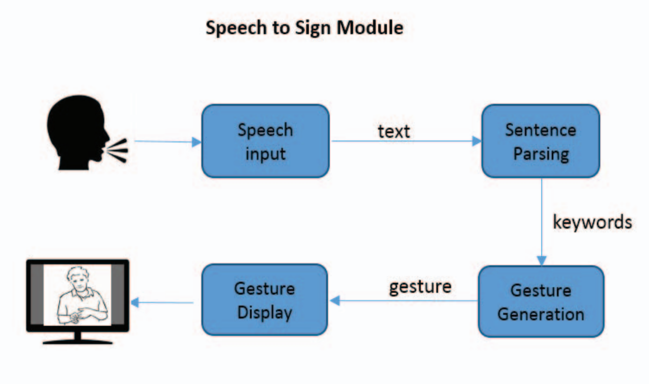

Fig. 1. Speech to sign module workflow

This will be done using external library for speech to text conversion which will take the sentence or word as input and identify the keywords in the input. System will then search for the sign/gesture against the keyword from the database. As there must exist a sign record for each keyword so this module shall get a valid gesture and pass it on to the display module. A data structure, inside Unity3D, is used that will map the gesture animations to a 3D model. Hence, 3D animation corresponding to the mapped gesture will be played on the screen as shown in Figure 1.

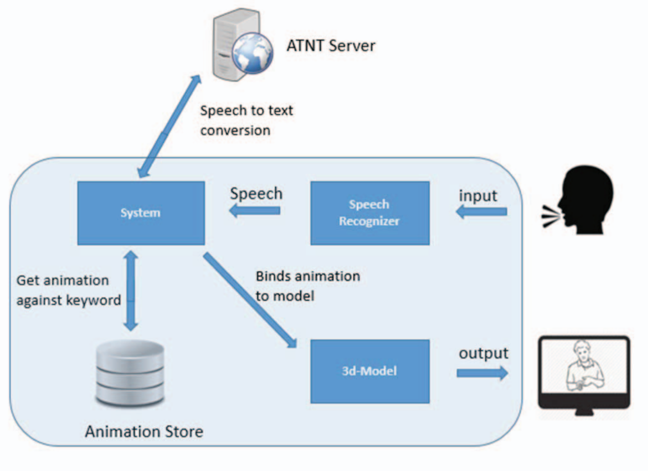

Fig. 2. Speech to Sign Architectural Diagram

[3] David S., August, Ludwig H., 2014. Concept in a student competition Future Lions in Cannes
[4] Ryan H., Alexandr Opalka created Motionsavvy in 2014
[5] Enable Talk presented In Microsoft Imagine Cup (2012)

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

The Architectural Model (as shown in Fig. 2) is both simplistic and interesting at the same time. The system will record speech on user's request and pass it to AT&T's Speech Recognition server. The server will in turn return the text from speech recognition in the form of a string. The system will then parse this string into words and look for their respective animations into the dictionary stored in database. These animations will then be played in sequential manner.

B. *System Components*

Speech to sign part was built using the following fine grained components. Each component along with its purpose is described below:

1) Modeling: A 3D biped model has been used to demonstrate a word or a sentence in the sign language as shown in Fig. 3. The idea is to use modeling as a medium of representing a sentence so that it can be easily comprehended by a person with hearing disabilities. The conversion could be just speech into text but text is not as friendly medium of expression as a 3D model for a deaf person.

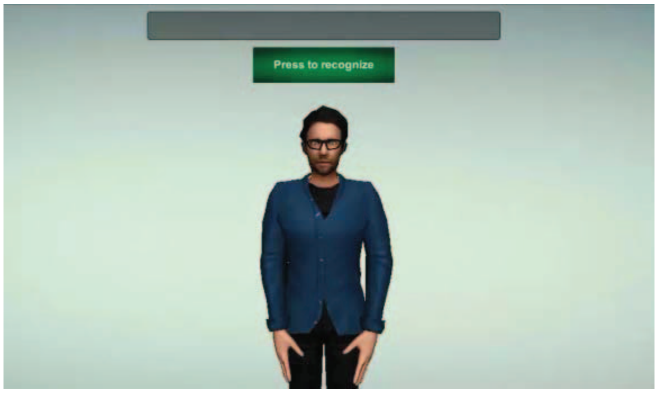

Fig. 3. 3D human model

2) Animation: In computer graphics, the word animation means to simulate a model to show how a physical person or object would behave under similar circumstances in reality. Animations were used to simulate words in real time just like a video or a movie. 3D Studio Max was the tool used to create animations for words that were then mapped against their respective words and these animations were played when such a word was received as input.

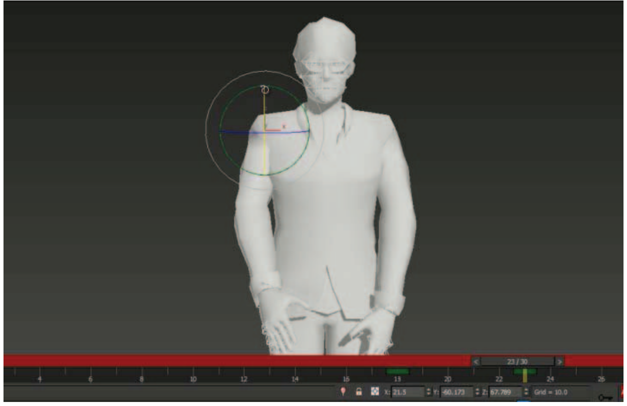

Fig. 4. Speech to Sign Architectural Diagram

3) Speech to Text API: The prerequisite for transforming the speech into animation is to convert the speech into text or a string. For this purpose, AT&T's online API along with its SDK for Unity3D was used.

4) Integration with Unity Framework: All the above mentioned components were then integrated together in Unity3D's development environment. The models were imported into Unity and were placed in the scene at their respective positions. Animations were also imported into Unity3D and they were mapped against the dictionary of words. AT&T's SDK was integrated with the Unity3D's script for speech recognition.

IV. METHODOLOGY (SIGN TO SPEECH MODULE)

This part addresses the implementation techniques used to build the "sign to speech" module. This is done using Microsoft Kinect V2's Continuous Gesture Builder which is responsible for gesture recognition and training. The implementation details of the module is as under.

A. *Implementation Approach*

The Kinect v2 SDK provides a tool, named gesture builder, which has been specifically designed for simple gestures recognition from Kinect manufacturers. It provides the simplicity by abstracting the details of going deep down into coordinates, angles, states and depth for gestures recognition. This tool is trained on the stored gestures and it detects the gesture when the hand motion is performed by the user. The gesture builder tool uses data-driven machine learning algorithms for trainings of the gestures.

For implementation of this module, tools and libraries provided my Microsoft for Kinect development have been used which include Kinect Studio for gestures recording in specific formats, Visual Gestures Builder (VGB) for training of the gestures through which gestures database can be generated. The implementation approach of sign to speech module has been described in Fig. 5.

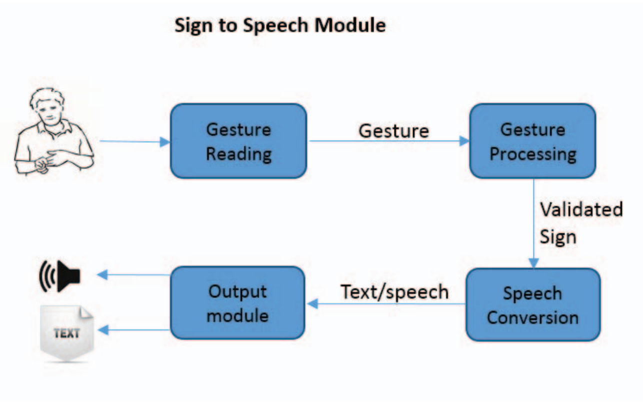

Fig. 5. Sign to speech module workflow

The system gets the gestures input frame by frame using Kinect and it matches each gesture with pre-stored gestures database. For each matched gesture there is a keyword. Then a sentence is constructed using those keywords and ultimately that sentence is converted into speech using text to speech

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

libraries provided by .Net. Now this finalized sentence is spoken by computer machine in the form of natural language which can be easily comprehended by the human.

The Architectural Model of this module is represented in Fig. 6. The system reads the input gestures from sensor, matches them with pre-stored gestures database, identifies keywords in case of hit, sends those keywords to text to speech conversion library and then produces the speech output. System shall get the speech input that will be processed and get converted to the related text form.

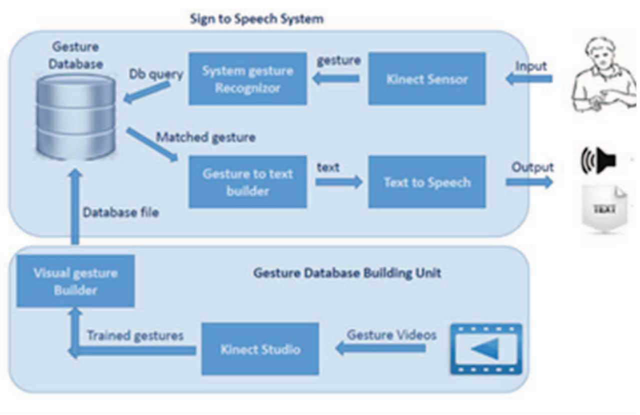

Fig. 6. Architecture model for sign to speech module

B. Algorithms Used (technical overview)

The gestures recognition technologies in Visual gesture builder are grouped in two main categories, AdaBoost and RFRProgress. AdaBoost is a trigger which gives us a true Boolean value while the person is performing a particular gesture, it uses Adaptive Boost machine learning algorithm. RFRProgress on other hand produces continuous results giving us Analog data of progress while the user is performing the gesture enabling the system to detect how much of the gesture is completed and how much is the hit rate at the particular frame of the gesture, it uses Random Forest Regression machine learning algorithm.

For accurate gesture recognition in the system combination of both has been used. The AdaBoost trigger enables the system to detect starting and ending points of classified gestures set having similar nature by returning true/false value while the gesture is being performed. Once AdaBoost trigger returns true value detecting the discrete gesture the system then uses RFRProgress to detect the correct gesture by reading continuous progress of gesture, in this case progress is only enabled if trigger is detected (AdaBoost returns true).

So the AdaBoost determines the context of the gesture being performed and RFRProgress detects gesture by continuously mapping the user’s movement. If the hit rate is enough for a particular gesture by combining both AdaBoost and RFRProgress it is considered a true positive.

C. System Components

Sign to speech part was built using the following fine grained components. Each component along with its purpose is described below.

1) Gesture Recording: In order to make Kinect learn about the gestures, the correct gestures have been recorded in the system first; this is done by recording the ideal gestures in Kinect using Kinect Studio as shown in Fig. 7. It stores3D coordinate points, Depth information, Body heat information and Infrared scan information and then generates its .xef or .xrf file which is then used by gesture builder.

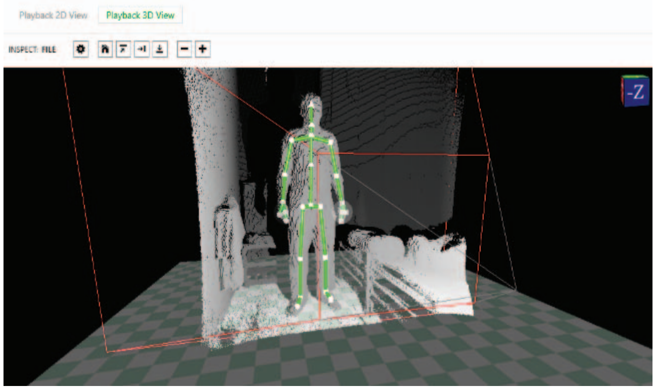

Fig. 7. Recognizing human body in Kinect studio

2) Gesture Tagging: In the gesture tagging process, .xef or .xrf files are imported into visual gesture builder to mark the correct gestures frame by frame. The frames can be marked as positive (always consider right) and as negative (never consider right) for each gesture. Each gesture has been tagged for discrete trigger and continuous progress both to make detection efficient. The decision whether the gesture is detected or not is based on these markings. The same gestures were tested and marked with different persons. Subsequently, these gestures were added in the gestures database which would then be used as a data dictionary to compare the incoming gestures on real time.

3) Gesture Testing and Training: As the system is dealing with real world data which could vary for different users and different situations it is important to understand that it should be able to produce expected results for unprecedented data. Same data cannot be used to train the system and for its testing purpose both. In order to cater this requirement two separate projects were created in visual gesture builder, one for gestures Training/Building and other for Analysis/Testing. The data in both projects is different and the ratio is 2 for 1 i.e. 66% data is used for training the system and 33% for testing and validation.

4) Gesture Recognition: In gesture recognition, the live stream of Kinect sensor or in other words the live gesture or hand motion is read frame by frame. This gesture is compared them with the already stored gestures in the database. If the incoming gesture matches with the stored gesture (there is a threshold value of confidence as well), the keyword against that gesture is added in the sentence. Then ultimately the computer speaks that sentence using text to speech library.

5) Text to Speech Conversion: In order to make the computer speak the translation of gestures of sign language, the text to speech functionality provided by .NET is used.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

V. Results and Discussion

In order to measure the performance and precision of the system, we devised our own method in which we compared the average hit rates of a particular gesture from the test runs by different users. Initially this result is based on statistics taken from 3 people with total number of 100 test-runs. The requirement of different body structures has been catered specifically as it is used to judge the precision in case of sign to speech conversion.

In sign to speech module, accuracy is 84%. This result is based on 100 test runs for different gestures with 3 persons having dissimilar body structures. In order to measure the reliability of the system to produce the expected results for a gesture we counted false positives (when the system detects the wrong gesture against the performed gesture i.e. detects gesture B against performed gesture A) and false negatives (when the system doesn't detect the gesture A performed) against the correct results. Usually false positives and false negatives are two different behaviors of the system but in our case both lead to wrong results so we have given both same value in our calculations. We have also observed that the accuracy of gesture measured heavily depends upon the number of gestures trained in the system and the number of overlapping gestures.

It has been observed that gesture recognition is sensitive to body structures or bone structure. This fact has been deduced while testing the gesture recognition module. For one of the test users whose height was different from the other two users, the gesture recognition results were degraded. This happened specifically due to the significant difference in height as compared to the other two users. This means that the gestures which are used to train one particular sign must cater all possible heights and postures. So when it comes to recognition, it will not treat any incoming gesture out of its scope.

For speech to sign module, the system is giving 87% accurate results with the given data set. This percentage can go up to 95% if the user is not Asian. The reason behind is its sensitivity to US based English accent. We used AT&T speech recognition library which is US based and the system is trained to recognize speech closer to their English accent. It's a little difficult for system to recognize exact words of English language in Asian accent.

The accuracy can be increased by using some other API that can be accent adaptive and offers training for individual users. It is worth mentioning here that if the speech is recognized correctly then the system will give 100% hit rate as it then does one to one mapping of the text to gestures.

VI. Conclusions

The software based solution based on Kinect v2 developed in this research work offers a mechanism through which people with hearing and speaking disabilities can communicate naturally with the rest of the world. These people currently use various gestures and symbols to talk and convey/receive their messages. But this does not solve the communication problem, as natural language speakers don't understand sign language hence there exists a communication gap between these two communities.

The developed system provides dual mode of communication between sign language speakers and natural language speakers. It not only reduces effort and time for a deaf person in communication but would also bridge communication gap between disable community and normal people. Overall this project is a utility for humanity and particularly for deaf community.

Since the system provides dual mode of communication so it has been categorized into two independent modules. The first module, sign to speech conversion, records gestures from the deaf person, comprehends these gestures and convert them into speech that can be understood by natural language speakers easily. In contrast, the second module, speech to sign conversion, takes natural language as an input, understands the language and displays corresponding sign language animation on the screen along with the subtitles. These animations are performed by a 3D humanoid model in real time and a corresponding subtitle is displayed on the screen to help the user in discerning the system.

As of now the project's final prototype has been completed i.e. developed and tested. The system is currently operational on a limited set of words. In order to extend the system, more words will be incorporated in the dictionary in future.

This system is reliable enough considering the accuracy figures 84% for sign language to speech and 87% for speech to sign language conversion. Usability is another plus point of this system as it doesn't require disable people to wear any gadget to perform gestures and they can perform hand motion as easily and flexibly as they do in their real life. As the developed system based on Kinect for Windows v2, so the system is compatible with any device which supports Kinect v2.

References

[1] Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X., Zhou, 2013. Sign Language Recognition and Translation with Kinect. In IEEE Conf. on AFGR, (2013).

[2] Chung,I.C.,Huang,C.Y., Yeh et S.C., et al., 2014. Developing Kinect games integrated with virtual reality on activities of daily living for children with developmental delay. Advanced Technologies, bedded and Multimedia for Human-centric Computing, SpringerNetherlands, Vol. 260, pp1091-1097.

[3] Lai, K., Konrad, J., Ishwar, P., 2012. A gesture-driven computer interface using Kinect, 2012. In IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), IEEE, pp 185-188

[4] Lang S., Block M., Rojas R., 2012. Sign Language Recognition Using Kinect, In 11th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2012), published by Springer in the Lecture Notes in Artificial Intelligence series, Part I, LNCS 7267, pp.394-402

[5] Li, Y., 2012. Hand gesture recognition using Kinect.In IEEE 3rd International Conference on Software Engineering and Service Science (ICSESS), IEEE, pp 196-199

[6] Ren, Z.; Yuan, J.; Meng, J.; Zhang, Z., 2013, Robust part-based hand gesture recognition using kinect sensor. In IEEE Transactions on Multimedia,15(5), 2013, 1110-1120, pp 1110–1120

[7] Sattar K., Irshad S., Talha S., 2014.Kinotherapy, a thesis in NUST School of Electrical Engineering and Computer Science, Pakistan (unpublished)

[8] Smisek,J., Jancosek, M., Pajdla, T., 2013. 3D with Kinect Consumer Depth Cameras for Computer Vision, Springer London, pp 3-25

[9] Suarez, J. Murphy, R.R. 2012. Hand Gesture Recognition with Depth Images: A Review. Proceedings of the IEEE RO-MAN, Paris, France, pp. 411–417.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

[10] Tang, M., 2011. Recognizing hand gestures with microsoft's Kinect, Department of Electrical Engineering of Stanford University

[11] Tian, J., Qu, C., Xu, W., Wang, S., 2013. KinWrite: Handwriting-based authentication using Kinect, Proceedings of the 20th Annual Network & Distributed System Security Symposium (NDSS), San Diego, CA.

[12] Villaroman, N., Rowe,D., Swan,B., 2011, Teaching natural user interaction using OpenNI and the Microsoft Kinect sensor , Proceedings of the 2011 Conference on Information Technology Education, ACM, pp-227-232

[13] Yang,H.Dee, 2015, Sign Language Recognition with the Kinect Sensor Based on Conditional Random Fields, Sensors, 15, pp 135-147;

[14] Zafrulla, Z., Brashear, H., Starner, T.; Hamilton, H.; Presti, P., 2011, American sign language recognition with the kinect. In Proceedings of the Internation

978-1-4673-8460-5/16/$31.00 ©2016 IEEE