Virtual Reality Markup Framework for Generating Interactive Indoor Environment

W.A.U.Y.S. Wickramasinghe[1], P.S.R.S.De Saram[2], C.P Liyanage[3], L.N.R. Rangika[4],

Dr. L. Ranathunga[5]

[#]*Faculty of Information Technology, University of Moratuwa*
Katubedda, Sri Lanka

{ureshan2011[1],ruwasaram[2],cheraniliyanage[3],liyanage.ruwan27[4]}@gmail.com, lochandaka@uom.lk[5]

***Abstract*—This research proposes an interactive three-dimensional (3-D) application which can be used for modelling and manipulating interior architectural environments. This incorporates the concept of Virtual Reality [1] (VR) giving the user a real world experience. Virtual Reality is a trending technology which is gaining popularity due to the user experience it gives and its effectiveness. The proposed system consists of 4 major components. The user can add objects to the environment either using a description given in natural language or by giving a hand drawn sketch of the objects to be added. Natural Language Processing (NLP) techniques will be used to extract object details from the user description and Image Processing techniques will be used to extract object details from the hand drawn sketch. The user can control the 3D environment either by using hand gestures or voice commands. The hand gestures will be captured using a web camera and the voice commands will be captured using a microphone and Natural Language Processing (NLP) techniques will be used in order to extract the user's command. The 3-D application will be developed using VRML [2] (Virtual Reality Markup Language) which is a language for describing three-dimensional (3-D) image sequences and possible user interactions to go with them. Using the application developed using VRML, the user can interact with the virtual environment by viewing, navigation, moving and rotating objects within the environment thus it gives the user a feeling as actually living in the environment.**

***Keywords*— Virtual Reality, VRML, NLP, Gesture Tracking, Voice Commanding, Image Processing**

I. INTRODUCTION

Virtual Reality (VR) and interactive applications are trending research and development areas in the Information Technology (IT) field. These areas are becoming popular since they give an immersive real-world experience to the users. And also 3-D modeling is extensively used in industries such as architecture, automobile and medicine due to its effectiveness. Therefore this research is done to combine those technologies and come up with an interactive 3D application which can be used for modeling interior architectural environments.

Even though there had been researches and applications developed using Virtual Reality (VR) and 3D modeling, the concept of user interaction with the 3-D virtual environment is not implemented in the field of interior modeling. One of the novel features in this research is to investigate the possibility of generating virtual 3D objects using VRML by processing natural language descriptions which are gathered from the user. Also the users can input a sketch of objects as an image to the system and generate the 3D view of the sketch. VRML models will be articulated based on those analysis and generated in the environment. This research is comprised with Hand Gesture Recognition with Image Computer Vision technologies like "Real Time Web Camera Video Feed Analysis". Users can use their hand gestures to manipulate VRML objects in the space. Touching and grabbing objects, releasing them, pulling and pushing objects based on hand gestures will let users to arrange the indoor environment in an interactive way. This will bridge the gap between the virtual world and real world and deliver a novel user experience. Sinhala voice recognition module further enhances the user experience by allowing the users to navigate in the virtual indoor environment using voice commands. Those voice commands will make an avatar to perform actions such as walk, turn around and fly. Sinhalese voice commands were used since it'll be more familiar for the Sinhalese users and also because Sinhalese voice commands had been rarely used in interactive applications.

II. LITERATURE REVIEW

This research focused on creating indoor environments virtually and letting the users to explore it in an interactive manner.

Generally, virtual reality (VR) systems require 3D computer graphic models that can be interactively experienced by participants [3]. Well-developed immersive VR systems will provide three-dimensional environments in which the user can directly perceive and interact with objects. The underlying factor that motivates most VR researches is that it will lead to a more natural and effective human-computer interface [4]. In a virtual environment there are 3 dimensions and a user is able to interact with the objects in the virtual environment. Hand gestures can be used to interact with the virtual environments.

Various types of sensing methods are used to get gestures for pattern recognition in order to identify hand gestures. The available sensing methods in the previous studies can be classified into three classes. They are movement-based, electromyogram-based (EMG-based), and vision-based approaches. Movement-based technique uses various sensors to measure movement. Glove-based is the mostly used movement-based approach and it gives good performance especially in sign language recognition. Data gloves like the Cyber Glove are used to track different hand gestures [5, 6]. Vision-based methods can detect, track and recognize hand gestures efficiently [7]. These methods are sensitive to scale, rotation, background texture, color, and lighting conditions.

Vision-based hand gesture recognition can be implemented and used for human-computer interaction effectively, particularly for applications in 3D Virtual Environments [8]. Video cameras can be used as the input device. Using the input of the video camera, tracking and detection of hand gestures can be done. Most of the existing vision-based hand gesture recognition systems use a color glove or a marker. It makes

hand gesture tracking easy.[9] Another important part of this research is generating virtual environments. Different methods have been used and mentioned in the literature for creating and interacting with virtual worlds using VRML. Virtual Reality Markup Language (VRML) is a textual language which defines objects and behaviors of objects in three dimensional worlds. WRL is used mostly to enable 3D views for users on Internet using different VRML plug-ins. As with languages such as HTML or XML, VRML is expressed verbatim. Although there are a variety of tools that allow graphic definition, interactive 3D objects that can then be exported as VRML text file, behavior definitions must be specified textually, usually through short programs written in a language of Script such as JavaScript linked to 3D objects using the instructions of the VRML files.

A. Generating VRML using text descriptions from users

NLP approaches have been used in graphic designing field order to generate scenes according to user descriptions. WordsEye automatic text-to-scene system by Coyne and Sproat [10] is one such system. This system allows the user to generate 3D scenes without having to use sophisticated graphic packages. The main drawback of this system is that it stores a large amount of 3D objects (around 2000) in a database and it requires large amount of storage space. Cockburn[11] in his work suggested that adjectives are used describe properties of objects or entities and they can be used as the attributes of the objects identified using nouns.

Some NLP based tools have been developed for automatic extraction of objects. One of such NLP systems is LOLITA proposed by Mich [12]. It's built on a large scale Semantic Network (SN). In this system also every noun is identified as an object and the relationships among the objects are identified using links. The drawback in this system is that, like any other SN system, this doesn't differentiate between classes, attributes and objects.

B. VRML object creation using image processing

Different method have been used in researches to detect shapes and objects. Several edge detection techniques also have been used.

Sobel edge detection is one such approach. In this method sobel operator is applied to the image to detect the edges.

Nowadays with the development of the web, people are able to view 3D virtual worlds and explore the virtual space. Creating and viewing the VRML world is quite easy and possible in most of the browsers. There are only two ways of creating 3D virtual scenes. First way is coding the scene directly using VRML. The other is using existing CAD and modelling software, and saving the created world in VRML format or converting it to VRML from another format. Both of the mentioned methods are time consuming.

Sketch that Scene for Me by Ellen Yi-Luen [13] is a technique which generates VRML world for free hand drawings by improving the idea of Igarashi. It analyzes what user draws and gives the relevant 3D view.

C. Manipulation of Virtual Objects with Real Time Gesture Tracking.

A research had been done which is more focused on loading 3D architectural models, saving attributes of it and manipulation, translation, rotation and scaling objects with gesture commands. This approach is similar to the proposed component up to a certain extent. They also have used hand gestures as input to the system in virtual environment. But they are using hand button mounted input method to get additional support for tracking the gestures and user inputs.

In our approach, the hand gesture tracking component is not based on any electronic hand glove, but instead using a natural hand with one-colorized usual glove for accurately tracking hand gestures. Our image processing and video processing implementations are smart enough to produce better inputs to the VRML framework with considerable percentage of accuracy. The accuracy rate of identifying hand gestures is currently at approximately 78% and this can be further increased with modifications to image processing algorithm. In the other's research approach they are using "Virtual Research V6 HMD as display unit to render the virtual reality model with Performer™, a graphics framework with 3D modeling libraries. Currently our research is capable of rendering models in any computer display with the usage of VRML enabled web browser.

D. Voice Recognition for maneuvering avatar in Virtual Environment with Voice Commands

Automatic Speech Recognition (ASR) is a popular field of research where different language models and applications are involved. There are a lot of ASR related researches out there, but still there is no Voice Recognition application targeting VRML avatar maneuvering. However a research [14] had been done for Sinhala voice recognition but it is not directly linked with VRML markup framework.

And there is another related research conducted which uses English language for voice commanding in an industrial VRML environment. In that approach, a context sensitive speech driven interface is used. But in the proposed approach a specific command set is programed to execute VRML avatar actions like run, stop, turn left, turn right, speedup and slowdown. The approach proposed in [15] invested lot of research effort for noise cancellation and enhancing accuracy level due to the application of the research. Since the research is focused on Industry Level usage, it is essential to have those implementations as well. But in the proposed system, it is not very much essential since the experimenting of the framework is done in a classroom or laboratory environment.

W.A.U.Y.S.Wickramasinghe[1], P.S.R.S.De Saram[2], C.P Liyanage[3], L.N.R. Rangika[4], L. Ranathunga[5]

III. Research Approach

The VRML Markup framework for architectural interior design has system inputs as illustrated in Fig 1.

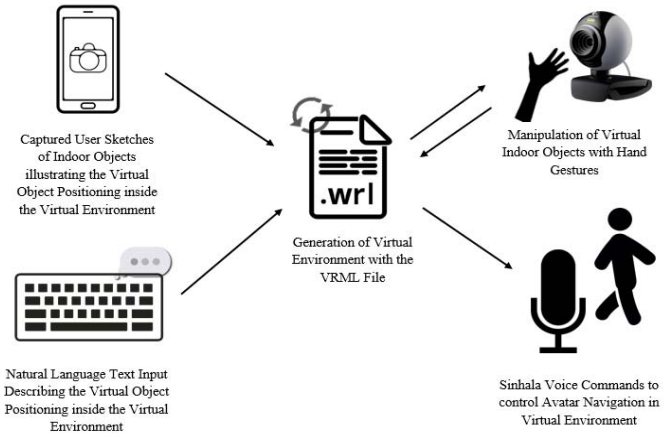

Fig. 1 VRML markup framework system inputs.

Different modules in the system process these inputs accordingly and gives output to the central VRML renderer enabled browser. Following figure describes those input modules and the following subsections will explain the design aspects of each module.

A. Generating the VRML objects using text descriptions

In this module Natural Language Processing (NLP) techniques are used to extract information from a text description written in English language. After identifying the objects and their attributes from the text description, the 3D objects are generated accordingly. The main components of this module and their functions are depicted in the Fig. 2 below.

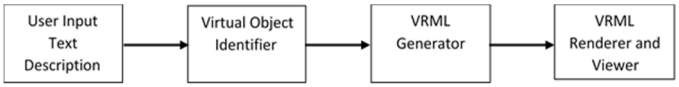

Fig. 2 Architecture of object generator using text descriptions

The process followed by this module is as follows;

1. The Part Of Speech (POS) Tagger is run on the text and the words are tagged according to their POS (ex: Noun, Adjective, Preposition etc.)
2. The sentences are parsed using the grammatical structure of the sentence and the noun phrases are identified. The noun phrases are assumed to be representing a single object.
3. Each noun phrase is further analyzed and the nouns are identified. The nouns are matched with a predefined set of objects and it is determined whether the noun represents a valid object. When determining this, the synonyms or the similar words identified using WordNet [16] are also considered.
4. The adjectives of the noun phrase are identified as the attributes of the objects. The attributes can represent the color and the size of the object.
5. The relative location of the object is identified using the prepositions in the noun phrase.
6. Coreference resolution is used to identify the references to the same object. The set of objects are refined according to this result.
7. The objects along with their respective attributes are stored in a tree structure and passed to the VRML generator sub-module.

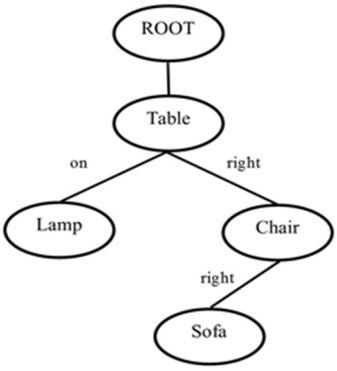

Fig 3 Tree structure generated

This module will take the tree structure that contains the objects and their attributes as a parameter. The tree will be traversed using in-order depth-first traversal and the VRML code to represent the objects and their attributes will be generated. The output of this module will be a text file containing the VRML code. VRML objects that are generated from the previous module can be viewed through the web browser using a plugin such as Cortona3d viewer.

B. Generating the VRML objects using image processing

Generation of 3D modules according to a sketch drawn by the user consists of 2 main parts. They are identification of the object and generation of the VRML world.

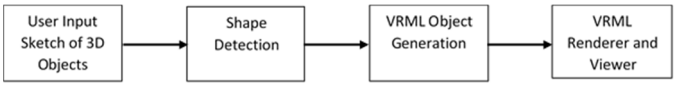

Fig. 4 Architecture of object generator using image processing

Fig. 4 shows the main components of the module with the inputs and the outputs. The user should input a 2D sketch to the system. It can be either the top view or the side view. The user should use shapes such as square, rectangle, circle, oval etc. to denote different objects such as chair, table, round table, coffee table etc. if he is drawing the top view. The user should use defined shapes if he is drawing the side view. There are specific templates for table, chair, bed etc. Following are the main steps followed,

- The user's input, i.e. the sketch is the input for the image processing module. By analyzing the image it identifies the edges of the image.
- In the image processing module, firstly the image is converted to the grey scale image
- Gaussian function is applied to remove the noise in the image.

- Then it is thresholded to find the edges of the image. Adaptive thresholding is suitable where there are different intensities in the input image. Otherwise global thresholding can be used.
- The accuracy of edge detection can be checked by changing the threshold value.
- Contour approximation method is used to identify the shapes. Contour approximation is based on the Ramer-Douglas-Peucker algorithm.
- The contour consist of set of vertices. The number of vertices will determine the shape.
- If vertices=3 shape=triangle
- If vertices=4 and aspect_ratio~1 shape=square
- Circles can be identified considering the ratio between the area and (diameter)^2. The value should be between 0.7 and 0.8.
- Then the relevant VRML object is created by generating the VRML file.

Next task is placing the identified object in the 3D world according to the sketch drawn by the user. In order to place the object in the 3D world x, y & z coordinates should be specified. If the user draws the top view, only the relevant x & z coordinates will be calculated. If the user draws the side view, only the x & y coordinates will be calculated.

C. *Manipulation of Virtual Objects with Real Time Gesture Tracking*

The main purpose of this component is to deliver a user experience of manipulating 3D objects in Virtual Environment (VE) with real time hand gestures. When the architectural design of the house is generated, users will be able to move some objects inside the building.

First users need to tune their camera feed by reducing noise of the capturing by adjusting filters using the GUI controls that are provided by the system. The users will be able to sharpen video frames and make sure to lower the noise levels. If there are too much noise, text label will be displayed to guide the user. If there is high noise, several visual modes are implemented for projecting camera view to screen with contour tracing techniques to detect the boundary of objects perfectly from a colorful background. Figure 6 shows the contour tracing mode. By looking at the visual, users will be able to adjust dilute and blur settings for a sharpened boundary.

Fig. 5 Contour Tracing for identifying hand objects from the camera feed

Next process is the focusing on contour tracing and noise filter application with green filter for realistic view. Users will get more opportunity for tuning the image for higher accuracy.

Fig. 6 Aligning contour tracing and morphing in actual camera feed

The system identifies hand gestures and it is available for maneuvering pc controls like a mouse pointer. The proposed approach is to build an interface between this hand detection modules with the VRML Virtual Environment.

D. *Voice Recognition for maneuvering avatar in Virtual Environment with Voice Commands*

In the proposed approach, one of the objectives is to let users to navigate inside the virtual environment controlled by voice commands. Here an avatar is used as a character controller and Sinhalese language is the language which is used to input voice commands into the system.

For this purpose, the voice recognition library "CMU Sphinix" is used. It is a free and open source library under BSD license. The Sinhalese voice phrases will be captured using a microphone and those clips are segmented into chunks by the middleware, which is implemented to compare those wave files with the trained dataset, library and identify the words. For indoor navigation control, following key terms are used for different actions. Go forward, stop, go backward, turn left, turn right are those voice commands. According to those voice commands, the avatar behaviors vary.

IV. Implementation

A. *Generating the VRML objects using text descriptions*

This module is implemented using Apache OpenNLP natural language processing library. After obtaining the user input it will be passed to the object identifier sub-module. Here each word will be tagged with their respective Part Of Speech (POS) using the POS tagger available in the Apache OpenNLP library. Then the sentences are parsed and the parser will output a set of noun phrases and verb phrases. Here we consider a noun phrase to be representing a single object. For each noun phrase the tag of each word is inspected and the words with the tag "NN" (noun) will be identified as objects (ex: - table, chair). The nouns will be matched with the set of predefined objects before determining it as an object. The words with the tag "JJ" (adjective) is identified as an attribute of the noun following that adjective (ex: - brown, large). The words with the tag "IN" (preposition) are identified as the relative location of the object (ex: - above, below).

Then the coreference resolution function in Apache OpenNLP library is used to identify instances where the same object is referred. For example, in the description "There is a table. It is brown in color" we can identify that "It" in the second sentence refers to the object "table" mentioned in the first sentence. The set of objects and their attributes are adjusted according to the results obtained from coreference resolution.

After identifying the objects and their attributes, the VRML code is generated by the VRML generator sub-module accordingly.

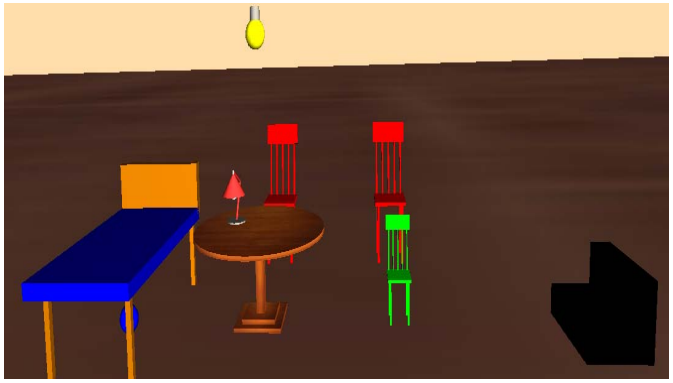

Fig. 7 3D environment generated using VRML

B. Generating the VRML objects using image processing

Image processing techniques are used to identify the shapes in the sketch drawn by the user. OpenCV library is used for this purpose. Relevant VRML file is generated after identifying the object. Contour approximation is used to perform shape detection. Contour approximation is an algorithm for reducing the number of points in a curve with a reduced set of points. That's why it is named approximation. This algorithm is commonly known as the Ramer-Douglas-Peucker algorithm, or simply the split-and-merge algorithm. Contour approximation assumes that a curve can be approximated by a series of short line segments. This results an approximated curve that consists of a subset of points that were defined by the original curve. Contour approximation is implemented in OpenCV(cv2.approxPolyDP method). A contour consists of a list of vertices. The number of vertices can determine the shape of an object.

- If the contour approximated contains 3 vertices, then it should be a triangle.
- If a contour is having 4 vertices, it can be either a square or a rectangle. In order to determine whether it is a rectangle or square, the aspect ratio is calculated. That is simply getting the width and the height of the contour bounding box and dividing the width by the height If the aspect ratio is ~1.0, then the shape a square (since all sides have approximately equal length). Else the shape is a rectangle.
- If a contour is having 5 vertices, it is a pentagon.
- If a contour is having 6 vertices, it is a hexagon.
- If a contour is having more than 6 vertices and the value of area/(diameter)^2 is between 0.7 and 0.8 , it is a circle.

After identifying the relevant shape the object representing that shape will be generated. That is done by generating the relevant VRML code. In VRML objects are defined using geometric objects.

If the user draws the top view, they should use the shapes given. Each shape represents an object. Square represents a chair, rectangle represents a table, hexagon represents a sofa, circle represents a round table etc. From the shape detection module relevant shape and the location of the centroid point can be obtained. This module generates the VRML file relevant to the sketch. Then the 3D model can be viewed.

In the 2D sketch the points are relative to the point the top left corner. But in the 3D world points are relative to the point in the middle. The point relevant to the 3D world should be found. In order to do this height and the width of the image should be found. When the user inputs the top view the y coordinate of the 3D world will be 0. Only x & z coordinates will be considered. Similarly this can be done to the side view as well. Only the x and y coordinates in the 3D world are considered in this scenario.

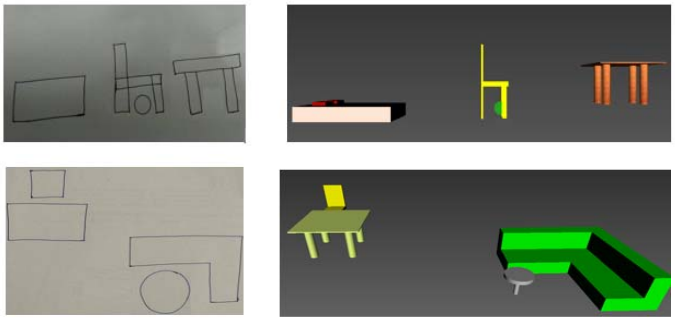

Fig. 8 Detecting multiple shapes from side view and top view sketches

C. Manipulation of Virtual Objects with Real Time Gesture Tracking

This module is mainly coupled with the video camera which is used as the input device to get signals into the component. OpenCV library is used to analyze the input feed. Morphing and contouring image processing techniques are applied to reduce the noise at the second stage. Users also can manually tune the video feed as described in above chapter with the controllers given. Then the component can track the hand gestures under "Trackable Mode" which is described in the 3rd phase.

When describing the process of detecting hand from video feed, it follows an algorithm as follows in fig 9. The video stream is segmented into frames and those frames are sampled. Each frame is converted into its binary representation and contouring is applied.

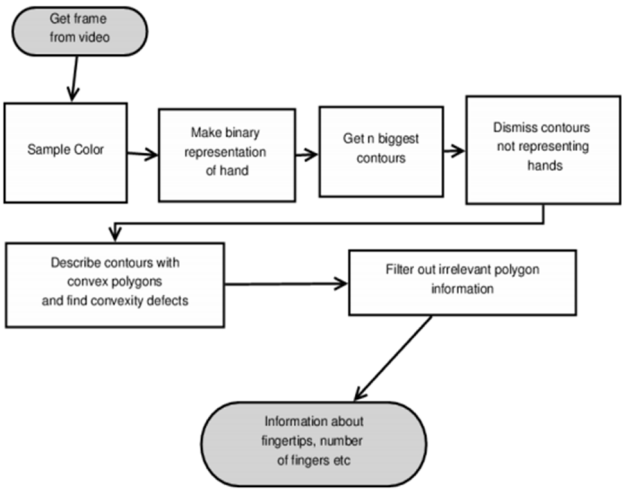

Fig. 9 Flow of processing the hand detection algorithm module

D. Voice Recognition for maneuvering avatar in Virtual Environment with Voice Commands

This component has a language input in Sinhalese. Stereo or Mono based voice input is taken into the module and it decomposes the sound into chunks. The design for this component is mainly under 2 sections which can be named as "Acoustic Model" and "Language Model". The acoustic model is the major component which is detecting phoneme in speech and Language model can detect the connection between words in a sentence. The acoustic model of the component receives speech input as a recording and phonetic strings as scripts to compile and transform into a statistical representation of speech sounds. Language model is a pronounceable word dictionary which is developed using a training dataset. It is referred as bigram language model.

The acoustic data of the training dataset is processed for feature extraction and the output will be the input for acoustic model training. The training dataset script (transcript) is the input to build the language model. The trained model will be tested and results are obtained to do some adjustments for enhancing accuracy.

V. Conclusions

In this ongoing research aspect, all four modules were tested with experimental inputs and measured outputs to compare with expected outcomes.

In the Generating the 3D model using text description model we observed an accuracy rate of 77% when using descriptions with a moderate level of complexity. However the accuracy level is becoming lower as the complexity increases.

Generating the 3D model of a hand drawn sketch. Image processing techniques such as thresholding and contouring is used to identify the shapes in the hand drawn sketch. The threshold value for the image of the sketch should be experimentally determined. Adaptive thresholding can be used when there are different lighting conditions in the image. If the threshold value is not correctly chosen edges of the image cannot be correctly detected. If the threshold value is correctly chosen and if the lighting conditions are good (if the image is not overexposed or underexposed), this has 80% of accuracy for identifying hand drawn sketches.

In hand gesture recognition module, the implemented hand gesture tracking module displayed up to 78% of hand tracking accuracy level with default threshold value assigned by the module automatically.

By combining all four modules, the research application functioning with expected level of user interaction support and it is currently under modifications to further improve its accuracy levels and usability.

References

[1] Woodrow Barfield and Thomas A. Furness III, "Virtual Environments and Advanced Interface Design," Oxford Univ. Press, 1995

[2] Hand Interface for Immersive Virtual Environment Authoring System by Motion Information Team, VR Center, Electronics and Telecommunications Research Institute (ETRI)

[3] Woodrow Barfield and Thomas A. Furness III, "Virtual Environments and Advanced Interface Design," Oxford Univ. Press, 1995

[4] Q. Chen, A. El-Sawah, C. Joslin, and N.D. Georganas. 'A dynamic gesture interface forVE based on hidden Markov models', Proc. IEEE International Workshop on Haptic, Audio and Visual Environments and their Applications (HAVE2005), pp.110–115, 2005.

[5] Y. Wu, and T. Huang. 'Hand modeling analysis and recognition for vision-based human computer interaction', IEEE Signal Processing Magazine, Special Issue on Immersive Interactive Technology, Vol. 18, No. 3, pp.51–60, 2001.

[6] C. Joslin, A. El-Sawah, Q. Chen. and N.D. Georganas. 'Dynamic gesture recognition', Proc. IEEE Instrumentation and Measurement Technology Conference (IMTC2005), Ottawa, Canada, 2005. 15

[7] Nasser H. Dardas and Mohammad Alhaj ,"Hand Gesture Interaction with a 3D Virtual Environment"

[8] Y. Wu, and T. Huang. 'Hand modeling analysis and recognition for vision-based human computer interaction', IEEE Signal Processing Magazine, Special Issue on Immersive Interactive Technology, Vol. 18, No. 3, pp.51–60, 2001.

[9] C. Joslin, A. El-Sawah, Q. Chen. and N.D. Georganas. 'Dynamic gesture recognition', Proc. IEEE Instrumentation and Measurement Technology Conference (IMTC2005), Ottawa, Canada, 2005. 15

[10] Bob Coyne, Richard Sproat .WordsEye: An Automatic Text-to-Scene Conversion System. AT&T Labs— Research, 2001.

[11] A. Cockburn, "Using Natural Language as a metaphorical Basis for Object Oriented Modeling and Programming", IBM Technical Report TR-36.0002, 1992.

[12] L. Mich and R. Garigliano. A linguistic approach to the development of object oriented systems using the nl system lolita. In ISOOMS '94: Proceedings of the International Symposium on Object-Oriented Methodologies and Systems, pages 371–386, London, UK, 1994. SpringerVerlag.

[13] Ellen Yi-Luen Do, "Sketch that Scene for Me": Creating Virtual Worlds by Freehand Drawing

[14] Thilini Nadungodage and Ruvan Weerasinghe, Continuous Sinhala Speech Recognizer, March 2015, University of Colombo School of Computing, Sri Lanka

[15] Stuart Goose, Ingo Gruber, Sandra Sudarsky, Ken Hampel, Brent Baxter, g Nassir Navab , 3D Interaction and Visualization in the Industrial Environment, Multimedia Department, g Imaging and Visualization Department Siemens Corporate Research 755 College Road East, Princeton, NJ 08540, USA

[16] George A. Miller. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41, 1995