VECIMS 2009 - International Conference on Virtual Environments, Human-Computer Interfaces and Measurements Systems
Hong Kong, China
May 11-13, 2009

Authoring Edutainment Content through Video Annotations and 3D Model Augmentation

Abu Saleh Md. Mahfujur Rahman[1], Jongeun Cha[2], Abdulmotaleb El Saddik[1]
Multimedia Communications Research Laboratory
University of Ottawa,
Ottawa, Canada
E-mail: {kafi,abed}[1]@mcrlab.uottawa.ca, jcha@discover.uottawa.ca[2]

Abstract— **In this paper, a real world based interaction metaphor has been adopted to facilitate learning about physical objects in an entertaining fashion. The proposed system incorporates an intuitive video annotation approach in order to catalog and author information about physical learning objects in a scene. The system uses augmented 3D visualization schemes and provides adequate visual cues in order to leverage the hand gesture and voice based interactions with the learning objects. These real world interaction techniques make the system transparent from the young learners and help them to become engaged in their learning activities.**

Keywords-video annotation; multi modal interaction; ambient virtual environment; haptics; edutainment

I. INTRODUCTION

Edutainment systems which combines entertainment with education are becoming popular form of e--learning. Studies [2], [3] have shown that entertainment improves learning skills by increasing learners' interests. However, most of the edutainment systems are based on artificial virtual environment and the interaction with these systems are still not intuitive due to unfamiliar interfaces. To bridge the gap between the virtual environment and the real world, edutainment systems should be presented with intuitive interaction schemes. In order to deliver game-like features many researchers have adopted augmented reality (AR) and 3D virtual environment interfaces [1] [16]. Developing an Augmented-Reality Interface have been successfully used in entertainment, training and other engineering fields [11]. In recent years, great advances have been made in AR rendering and tracking technologies which if used coherently could bring intuitive and entertaining interaction with virtual objects in the AR scene.

Using existing familiar objects and real world interaction techniques may play very important role in getting an interface accepted by the young learners. This paper introduces the AR interface based learning system and proposes an authoring application to create contents. We are proposing an edutainment system that presents real object authoring and navigation through video annotation scheme. The interface makes it possible to interact with the physical objects using natural hand gestures and view learning contents about the objects. The design methodology behind our edutainment system is depicted in the following figure 1:

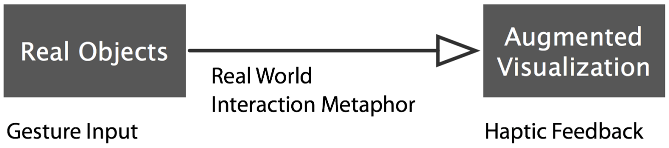

Figure 1. High level structure of the edutainment system.

The system listens to the learner's voice and tracks the hand movements in the scene using a depth-camera [14]. If the learner 1) speaks a word that is found in the annotation database or 2) places the hand over a physical object. The system triggers an event to and the annotated information is then rendered in the visual scene. The augmented information about the physical objects could be rendered in a television screen or projected on the wall for better immersion. While interacting with the objects the learners can see themselves in the scene (with augmented visualization). This helps them to keep track of the physical objects they are interacting with, get timely feedback and become engaged in their learning activities.

The remainder of the paper is organized as follows: firstly in section 2 we discuss some related works. Section 3 gives the illustration of the video annotation scheme. Software architecture and its components are presented in section 4. Finally, in section 5 we discuss the results, point some limitations and provide cues for possible future work.

II. RELATED WORK

Karime et al. [6] developed an edutainment system based on communication and multimedia technologies that process the ambient voice of the young learners in order to obtain meaningful words and display pictures about those words. While the system is an interesting voice based image search scheme, it assumes that the learners are already acquainted with the names of the objects they are going to learn or learn about.

MagicLenses [16] uses a see through AR interface in order to view world map data on specific objects from the real scene. A user has to hold and move a semitransparent tangible device for the interaction.

978-1-4244-3809-9/09/$25.00 ©2009 IEEE

The Magic story cube [11] presents an interactive way to relate children's. The system uses AR technology, in which computer graphics are superimposed on top of a child's traditional "magic cube" to create an animated version of a story. The system requires head mounted display which is uncomfortable for the young. Also, unlike our approach it has limited support for content authoring.

Lastly, Chipman et al [13] illustrates a system that uses tangible flags to advertise the location of interaction. Wireless computer and RFID tags are used to access and collaborate on content (image, data, etc) authoring. While useful, the system is targeted for learners with good technical background. Also, the annotated contents have no relation with the real objects.

III. System Description

In this section we illustrate different elements of our system. These elements and their inter-relations are depicted in figure 4:

A. Gesture Recognition

Our system keeps track of the learner's hand movement and gestures while the learner is moving his/her hands over the polygonal annotations in the graphical display.

1) Basic principle: Listen and track the movements of the hands and trigger events while the hand is over the annotated polygons.

2) Tracking of hand movements: The system uses the z-cam [14] in order to get hand movement and gesture data.

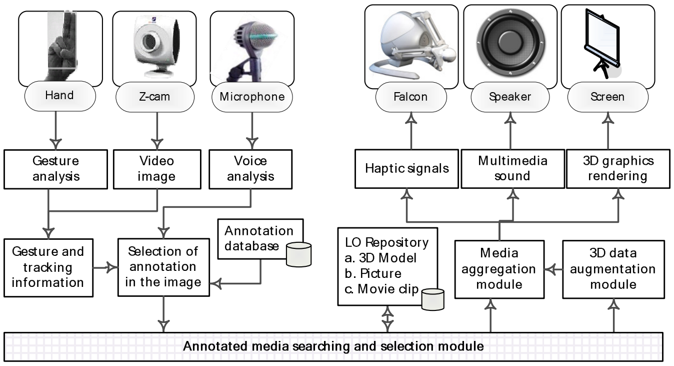

Figure 2. Real world (desktop) interaction metaphor with video annotation tool, haptic device, z-cam and other system elements.

3) Hand gesture commands: The hand gestures similar to the approach in [8] are available for the learners that could be customized to send different commnads to the proposed system. For example commands could be sent to the system in order to instruct it to play the annotation information continuously until the next annotation information is selected. These hand gesture leverages the access to the basic commnads of the system.

B. Voice commands

Our proposed system has a built-in voice recognizer similar to [12]. For each annotation several keywords could be registered for the recognizer. The recognizer then listens for these keywords from the learners and displays information about those keywords in the scene. However, this form of information navigation still needs more research to be more accurate.

C. Haptic feedbacks

When the physical object parts are augmented by 3D models with additional texture or missing information, the system tracks the hand movement and when it selects the models, high-fidelity three-dimensional force feedback could be felt in the attached haptic device [15]. This feature allows the user to feel and sense the augmented information that are presented in the rendered video. For example a skeletal physical object could be annotated and augmented with real skin, volume and texture information. This haptically enhanced feedback will offer the learners to feel the skin type, texture and other information of the virtually augmented objects. Hence, it becomes very useful in order to interact with the virtual representation of the real object. The feedbacks also help to sense the virtual events in the real world [9]. The haptic feature creates huge interest in the younger learners and makes their learning experience more enjoyable.

D. Learning Object play

The multimedia content designated for learning or simply Learning Objects (LOs) [10] selected by the learners while interacting with the system could be rendered and played in many ways:

1) Embedded player: The proposed system has built-in LO player that looks up the format of the content and selects appropriate ActiveX plugin to play the content.

2) Separate pre-associated program: The system could also be customized so that each of these learning LOs could be opened in separate pre-associated programs. This becomes useful when the LO could not be opened with the embedded player.

E. Synthesize speaker

The system has a computer synthesized speaker as described and used in [6, 7] that automatically reads the content of the LOs and speaks them out. Moreover, the speaker can occasionally provide new learners with basic instructions, e.g., on how to use the haptic device or announcing the learner that the system is waiting for a gesture command.

IV. Video Annotation Scheme

In figure 3 a typical setup for the annotation of real world objects has been presented. We annotate real objects with their real space in order to provide easy and natural interaction with those objects. Through interactions the learners can get relevant learning materials about the objects in an entertaining manner. In order to deliver these features in the system, a depth camera, Z-cam [14], constantly monitors the real objects, numbered (1-4) in the figure and supplies the image frame and other sensory information to the authoring application. The depth camera captures color images and synchronized gray-scale depth images containing per-pixel depth information. The depth image is used for hand gesture recognition in this system.

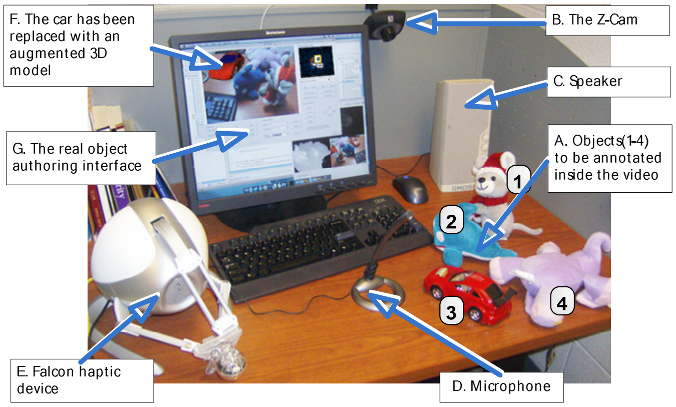
F. The car has been replaced with an augmented 3D model
B. The Z-Cam
C. Speaker
G. The real object authoring interface
A. Objects(1-4) to be annotated inside the video
1
2
3
4
E. Falcon haptic device
D. Microphone

Figure 3. Real world (desktop) interaction metaphor with video annotation tool, haptic device, z-cam and other system elements.

A. *Typical use case for the system*

Before progressing to the details of the video annotation scheme we want to describe a typical use case of the system using the interface depicted in figure 4. In the figure the colored polygons represent the stencils of the physical objects that are drawn as colored polygons using the mouse and annotated with learning contents (pictures, text and video). The interface provides appropriate tools in order to draw, move, rotate, scale and re-adjust the annotated polygons and stores those as an xml file.

While rendering the video the system processes the presence of the hands of the user in the scene. Therefore, at the time the learner is interacting with the physical objects, if the hands (with appropriate gesture) are present on the polygonal dimension of any objects then it renders/plays the learning content specific to that object in the built-in player. When the system senses the presence of the hand in the scene, the colored polygons are also rendered as overlays. By using the colored polygons as references the learners can easily locate the objects that are annotated and can interact with them.

B. *Polygonal annotation scheme*

In the authoring application, the color image frame from the Z-cam is annotated. The application presents tools to draw colored polygons on top of the graphics frame highlighting the parts of the physical objects in the image. These polygons then are associated with a range of video, image or search information. Afterwards, this information about the image frame, annotating the real object's different parts is stored in the form of learning objects in XML format. This gives us the option to load and share the annotated video across multiple platforms.

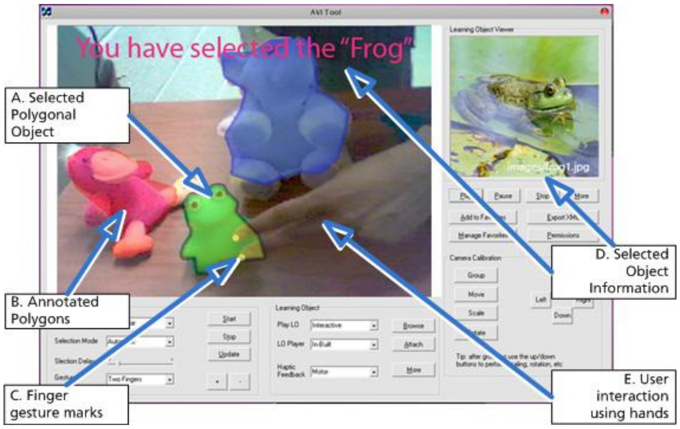

Figure 4. Interaction with real world physical objects, a typical use case

C. *Transformation and calibration*

If the camera's view changes then sometimes the polygonal annotations need to be changed. In order to perform this operation quickly the system offers several options e.g. transformation and re-calibration of the polygonal annotations.

D. *Automated search*

In case the manual selection of learning objects (LOs) is difficult the system provides an alternative by adding the feature of automated searching [5]. The searching paradigm is similar to the one presented in searching of the LOs in virtual gaming environment [1]. However, in the search result filtering process without relying on the system's filtering scheme [6], the LOs obtained should be approved by the designer/parents, i.e, the LOs are selected for the annotations at the design time not at the run time.

E. *Embedding 3D object with ease in the AR scene*

We have included several 3D models in the authoring interface that could be readily included in the AR scene. The 3D models could be customized using colors, positions and other transformation options. Hence, any real object could be enhanced with their complete view in the AR scene very easily. For example, in figure 2 the car has been replaced with a 3D car model in the authoring interface.

V. RESULTS

In this section, we present some performance evaluation to construct the system in terms of processing time and rendering quality.

A. *The system processing times*

It should be noted that considerable time is needed to capture the image and other sensory information from the Z-cam. The image obtained is then processed by the proposed system as a texture and later loaded into the graphics rendering system. Another time consuming system element is the hand tracking and gesture recognition module. Furthermore, when the distance between the learner and the camera changes it increase the processing time. Annotation geometry loading and rendering also requires processing time when the system initializes or the annotation data updates and needs to be reloaded. Different processing times of the system are shown in figure 5.

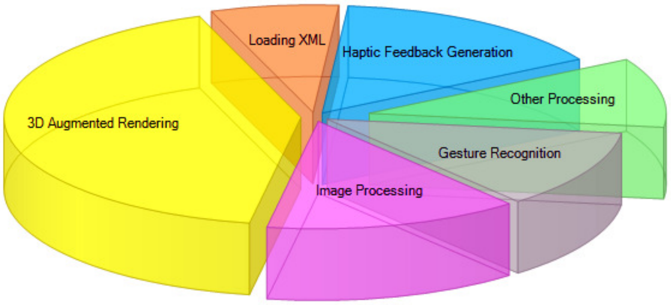

Figure 5. System processing time break down. 3D augmented rendering requires the most processing time.

B. *Rendering quality of the augmented visualization*

The system has been tested on a Pentium 4 workstation, with clock speed 3.4 GHz, 1.5 GB system RAM, and 64 MB AGP memory and running Windows XP service pack 2, version 2002. We have observed that the graphics rendering quality of our system is affected by the number of annotations that are present in the AR scene. More number of annotations results in less responsive system. This is depicted in figure 6. When the number of annotations are less than five the system provides optimal frame rate (48.7/sec) and gives smooth visualization of the AR scene. However, as the number of annotations reaches more than 50 the system results in poor graphics visualization. This happens as the number of annotations increases and the system starts to process their collision detection with the hand tracking information. The voice recognizer also loads more keywords and listens to the user to find a match. This gives less time to the graphics controller because it is rendering the 3D augmented information of the physical objects and at the same time creating haptic collision events to be rendered in the haptic device.

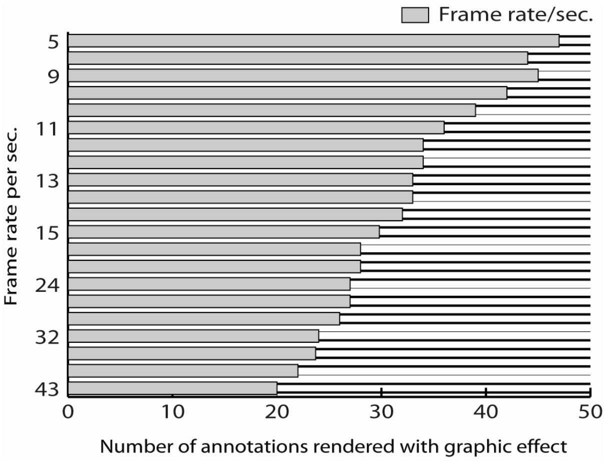

Figure 6. Visual rendering quality of the OpenGL image mapped in frame rate/sec against number of annotations rendered with graphic effects.

C. *Automated LOs fetch based on annotation information*

The system can also automatically recommend LOs for different video annotations. But each of these proposed LOs are filtered based on the learner's customization parameters. Therefore, depending on the number of search results and the selection of filtering options (dictionary based offensive words) the total time required to fetch LOs varies. Also the network sometimes causes delay particularly if a large number of recommendations is requested. In figure 7 the total amount of time required to obtain certain number of LOs are shown:

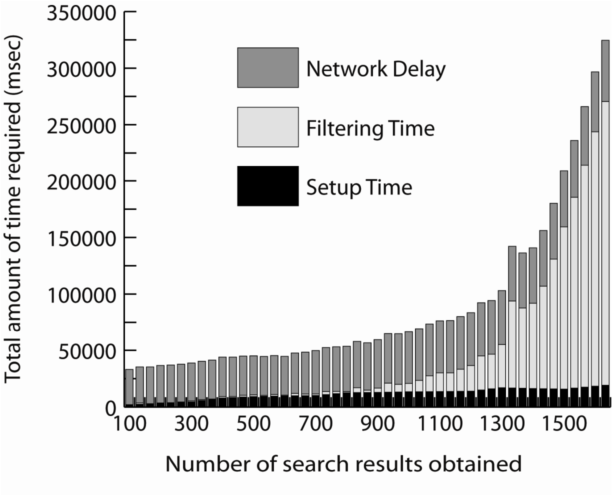

Figure 7. Overall delay in processing the automated annotation request for learning object fetch and filtering in groups (according to type info).

The system has certain other limitations: at the moment only tracking of two hands are supported, hence only two learners can collaborate in the interaction and learning process. Also, the system can not dynamically map the physical objects transformations to that of the virtual annotated polygon transformations. Hence, those need to be readjusted using manual transformation operations. Nevertheless, we are working to solve the later problem by using digital tags..

VI. Conclusion

In this paper, we have proposed and implemented an edutainment system where we can author physical objects using live video imagery. We are able to successfully demonstrate the ability of the system where the learners can interact with the real objects and educate themselves with various aspects of the objects while using and enjoying the intuitive and natural hand gesture based interaction technique.

In future an in-depth usability study will be performed with the system in order to measure different usability parameters from the learner's feedback. Also, in our future endeavor we want to determine the customization parameters which the system should support to serve both young and profession learners while assisting them in the learning process.

REFERENCES

[1] A. El Saddik, ASM Mahfujur Rahman, and M. A. Hossain, "Suitability of Searching and Representing Multimedia Learning Resources in a 3D Virtual Gaming Environment," IEEE Transactions on Instrumentation and Measurement, Braunschweig, Germany, Vol. 57, No. 9, pp. 1830–1839, Sept. 2008

[2] A. Mitchell and C. Savill-Smith, The Use of Computer and Video Games for Learning – A Review of the Literature, M-learning, Information Society Technologies, Learning and Skills Development Agency, UK, 2004, pp. 17-47.

[3] R. Rajaravivarma, "A Games-Based Approach for Teaching the Introductory Programming Course." SIGCSE Bull. Vol. 37, No. 4, Dec. 2005, pp. 98-102.

[4] B. M. Schlining, and N. J. Stout, "MBARI's Video Annotation and Reference System," OCEANS 2006, Boston, MA, pp. 1–5, Sept. 2006.

[5] E. Moxley, M. Tao, H. Xian-Sheng, M. Wei-Ying, and B. S. Manjunath, "Automatic video annotation through search and mining," IEEE International Conference on Multimedia and Expo, Hannover, Germany, pp. 685 – 688, 2008.

[6] A. Karime, M. A. Hossain, A. El Saddik, and W. Gueaieb, "A multimedia-driven ambient edutainment system for the young children," Proceeding of the 2nd ACM international workshop on Story representation, Vancouver, Canada, pp. 57-64, 2008.

[7] ASM Mahfujur Rahman, A. El Saddik, "An Algorithm for Search and Organization of Learning Objects in 3D Virtual Environment", 2006 IEEE International Conference on Virtual Environments, Human-Computer Interfaces, and Measurement Systems, La Coruna, Spain, July 10-12, 2006.

[8] Q. Chen, ASM Mahfujur Rahman, X. Shen, A. El Saddik and N. D. Georganas, "Navigating a 3D Virtual Environment of Learning Objects by Hand Gestures", International Journal of Advanced Media and Communication, Vol. 1, No.4 pp. 351–368, 2007.

[9] ASM Mahfujur Rahman, M. Eid, A El Saddik, " KissMe: Bringing virtual events to the real world", 2008 IEEE International Conference on Virtual Environments, Human-Computer Interfaces, and Measurement Systems, Istanbul, Turkey, pp. 102-105, July 2008.

[10] IEEE Learning Object Metadata, final draft standard, Jul. 15, 2002 http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf, (IEEE 1484.12.1).

[11] Z. Zhou, A. D. Cheok, J. Pan, and Y. Li, "Magic Story Cube: an interactive tangible interface for storytelling," Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology, Singapore, Vol. 74, pp. 364 – 365, 2004.

[12] ASM Mahfujur Rahman, Abdulmotaleb El Saddik, "Traffic Architecture Motivated Learning Object Organization in Virtual Environment," International Journal of Advanced Media and Communication, Vol. 2, No. 1, pp. 96–114, June 2008.

[13] G. Chipman, A. Druin, D. Beer, J. A. Fails, M. L. Guha, and S. Simms, "A case study of tangible flags: a collaborative technology to enhance field trips," Proceedings of the 2006 conference on Interaction design and children, ACM, New York, USA, pp. 1–8, 2006.

[14] 3D Camera & 3D Video Solutions – 3DV Systems, http://www.3dvsystems.com/, accessed Nov, 2008.

[15] Novint Falcon, a haptic based 3D game controller, http://home.novint.com/products/novint_falcon.php, accessed Sep, 2008.

[16] E. Bier, M. Stone, K. Pier, W. Buxton, and T. DeRose, "MagicLenses:The See Through Interface," In Proceedings of SIGGRAPH 93, ACM Press, pp. 73-80, 1993.

[17] A. El Saddik "The Potential of Haptics Technologies" IEEE Instrumentation & Measurement, pp: 10-17, February, 2007