What You Write Is What You Get: A Novel Mixed Reality Interface

ZhiYing Zhou, Syed Omer Gilani, and Stefan Winkler

Interactive Multimedia Lab
Department of Electrical and Computer Engineering
National University of Singapore (NUS)
{elezzy,elegso,winkler}@nus.edu.sg
http://www.ece.nus.edu.sg/iml

Abstract. Kids develop their understanding of the world by interacting with different objects. This cognitive development process can be further enriched through the use of interactive multimedia technology. This paper presents a novel application that combines the flexibility of the virtual world with a handwriting recognition engine to create a tangible mixed reality learning environment.

Key words: mixed reality, tangible user interface, real-time, handwriting recognition

1 Introduction

Teachers play an important role in the social and cognitive development of children. However, traditional teaching approaches put the focus on teachers and consider students as secondary [1]. This results in lower learning efficiency, since students may become passive and not fully attentive to the material. To improve learning among the diverse group of children, the initiative must come from the learners rather than the teachers. This can be achieved by the use of novel teaching approaches, integrating intuitive design and multimedia technologies. Advances in sensors and media technologies have made it possible to design real-time HCI applications suitable for the task. Such applications are receptive to the user's needs, easy to use, and offer multi-sensory experiences. With this in mind, we propose a novel, media-rich (text, graphics, video) application which combines the flexibility of the virtual world with a handwriting recognition engine to create a tangible mixed reality learning environment.

Although the importance of human tutors cannot be denied in the cognitive and social development of children, it can be enhanced with interactive media technology to enrich the learning process. Emerging interactive media technologies have the potential to not only help young people learn, but also engender a true love of learning [2]. Smart toys have been designed to stimulate creative and imaginative abilities in children [3]. Researchers have been working on making the learning process for children more intuitive by exploring new paradigms of

interaction [4]. There are also some haptics-enabled learning environments [5], but they require more expensive equipment.

Here we go one step further by using mixed reality in conjunction with tangible user interfaces (TUI) to make the learning process a fun and exciting experience [6]. In mixed reality the user has complete presence in the real world, while he or she interacts with objects in the virtual world that overlap with the real world. In our system, children use paper and a marker pen to write the name of an object and have it appear on screen in real-time. The system can be deployed on a desktop PC or a notebook without the need for special equipment. We term this interface "What you write is what you get"(WYWIWYG), as a counterpart to our previous work, "What you say is what you get" (WYSIWYG) [7].

2 System Design

The system hardware setup consists of a white board, a fiducial marker, a black marker pen, a webcam and a PC. The white board is used here for the reusability of the writing area. A fiducial marker is used to accurately position the 3D virtual object in mixed reality space, and to locate the text relative to its position.

Fig. 1 shows the system diagram. A digital image of the writing area is captured by a webcam. This image is first analyzed for the presence of text. Once the text is detected, the image is forwarded to a pre-processing stage. A character matrix is generated by the pre-processing stage for each character detected in the image. This character matrix is then analyzed by the recognition engine, which generates a prediction matrix. The final recognition in terms of noun, color and digit is obtained by cross-referencing the output with the knowledge base.

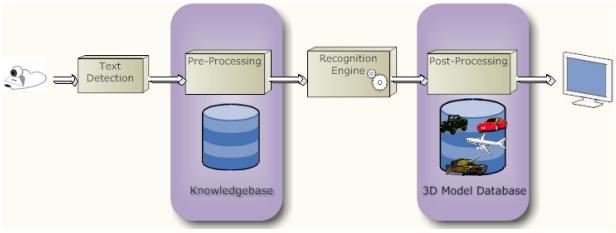

Fig. 1. WYWIWYG system diagram.

3 Methodology

Using simple tangible objects like a marker pen and white paper, users can write a noun, color, digit or any combination thereof to have corresponding 3D action displayed in real-time. There are two modes in the system: offline mode

and online mode. The offline mode is used to train the system, and there is no display of 3D objects. In online mode the system recognizes the input text and displays the 3D object in real-time.

3.1 Text Detection

The first step is capturing image frames of 640x480 resolution with a webcam. Since the white board is fixed with respect to camera, the system can initiate the text detection sequence by monitoring the hand occlusion. The captured image is in 24-bit, 3-channel RGB format. It is converted to an 8-bit, single-channel intensity image for subsequent text processing.

The text is searched in the area around the fiducial marker. This has two benefits: first, only a portion of the image has to be processed, and second the fiducial marker characteristics will not affect the calculation of the thresholds. Once the scanning area is determined, the image is segmented for further processing. Since we are writing with a black marker pen on a white board, it is easy to locate the text presence by the contrast variance of the text relative to the uniform background.

3.2 Pre-Processing

Once the text is located, the segmented image undergoes a series of pre-processing steps before input to the recognition engine. The image is binarized using Otsu's Adaptive Threshold method [8]. The bi-level image obtained is forwarded to character-level segmentation. This segmentation is done by Connected Component Analysis (CCA), followed by the grouping of characters. Characters are grouped under the alphabet and digit categories, by measuring the distance between the centers of gravity of adjacent character areas. This grouping of characters along with the word compendium builds up the knowledge base, which is used in the post-processing stage to get the final recognition results in terms of digit, color and noun. The knowledge base is updated each time a new word or digit is written by the user. The word compendium remains static, unless updated in offline mode. Once the characters are extracted, they are normalized to 32x32 pixels using bilinear interpolation. Normalized characters are then formatted for processing by the recognition engine.

3.3 Recognition

We use a C-SVC type vector machine[9] with RBF kernel support as a recognition engine. The performance of the SVM model depends on the fine tuning of its parameters C (error cost) and γ (kernel parameter). These parameters are selected with the help of ν-fold cross-validation technique, where ν denotes the number of subsets the training data is divided into. First we perform a coarse grid search with parameters C and γ having wide range of values, with exponential steps$(2^{-5}, 2^{-4}...2^{4}, 2^{5})$. We select the parameter values with the highest

cross-validation rate and perform a finer grid search in the neighborhood of these values. This ensures the best possible recognition rates in online (recognition) mode. Once the best parameters are found we train the recognition engine with the complete training data set. Current recognition rates for both alphabets and digits are shown in Table 1.

Table 1. Characteristics and performance of text recognition.

Data sets	Training samples	Testing samples	Attributes	Classes	Accuracy
Digits	390	60	1024	10	95%
Alphabets	1560	104	1024	26	81%

3.4 Post-Processing

The output of the recognition engine is a 1xN prediction matrix, corresponding to each character detected in the input text. To achieve good recognition results at the word level, we cross-reference the prediction matrix with the knowledge base. First the prediction matrix is divided into digit and alphabet groups, using the pre-grouping from before. Then the alphabets are further classified into colors and nouns by comparing them with the available word compendium. This results in the final recognition in terms of number, color and noun.

Once the complete text is recognized, corresponding action is taken in real-time. If only the color is present, the color of the default model (a tea-pot) is changed (Fig. 2(b)). If only the noun is present, the corresponding 3D model is loaded in default color (white) (Fig. 2(c)). Digit recognition clones the virtual model according to the number recognized (Fig. 2(d)). The 3D models are in wavefront geometry definition format [10] and are rendered using MXR Toolkit, our open source mixed-reality API.

4 Results

Since the objective of this mixed reality environment is to serve as an interactive learning tool, we encourage input in terms of handwritten characters, thus facilitating the goal of learning while playing. Currently the system can display 9 multiple instances of 20 models, corresponding to 20 nouns in our word compendium, in 8 different colors. Recognition and display are done in real-time. This is necessary for the user to have a consistent mixed reality perception. The processing speed of the system is 10 frames per second. The average recognition time is 400 msec on a current Pentium D. The system is developed in C++, using MXR Toolkit and OpenCV.

(a) (b) (c) (d)

Fig. 2. Real-time online recognition results.

5 Future directions

The aim of our system is to display multiple instances of the model in different colors. In the future we plan to add more media types, like sound and animation, to further enrich the user experience. We will also implement a text-to-speech engine to teach pronunciation. Finally, we will add performance indicators that reflect how well characters are written in order to improve the writing style.

References

1. Liu, K.R., Chen, Y.S., Chen, K.T., Chen, H.S.: New media learning for children – interact with color and tempo. In: Proc. IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing. Volume 2., Taichung, Taiwan (2006) 124–129
2. Wartella, E., O'Keefe, B., Scantlin, R.: Children and interactive media – a compendium of current research and directions for the future. Markle Foundation (2000)
3. Itoh, Y., et al.: TSU.MI.KI: Stimulating children's creativity and imagination with interactive blocks. In: Proc. 2nd International Conference on Creating, Connecting and Collaborating through Computing, Kyoto, Japan (2004) 62–70
4. Resnick, M.: Technologies for lifelong kindergarten. Educational Technology Research and Development **46**(4) (1998)
5. Portillo, O., Avizzano, C., Raspolli, M., Bergamasco, M.: Haptic desktop for assisted handwriting and drawing. In: Proc. IEEE International Workshop on Robot and Human Interactive Communication, Nashville, TN, USA (2005) 512–17
6. Zhou, Z., Cheok, A.D., Li, Y., Kato, H.: Magic cubes for social and physical family entertainment. In: CHI Extended Abstracts on Human Factors in Computing Systems, Portland, OR, USA (2005) 1156–1157
7. Zhou, Z., Cheok, A., Liu, W., Chen, X., Farbiz, F., Yang, X., Haller, M.: Multisensory musical entertainment systems. IEEE Transactions on Multimedia **11**(3) (2004) 88 – 101
8. Otsu, N.: A threshold selection method from grey-level histograms. IEEE Transactions on Systems, Man and Cybernetics **9**(1) (1979) 62–66
9. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001)
10. Chen, D.Y., Ouhyoung, M., Tian, X.P., Shen, Y.T., Ouhyoung, M.: On visual similarity based 3D model retrieval. In: Proc. Eurographics, Granada, Spain (2003) 223–232