Final Project Report

Lijun Yin

State University of New York at Binghamton

1. Project Description

This project aims to develop a new instructional and learning technology by utilizing the innovative computer vision and computer graphics techniques to enhance the learning and teaching experience. We have developed an intelligent avatar tutor system (so-called iTutor) and a graphical demonstration system (so-called iDemo) for achieving the goal of visual learning. Visual learning – the use of graphics, images, and animations to enable and enhance learning – is one important strategy that we employed. We further improve the level of simulation and intelligent interaction through a number of innovative components, including face expression recognition, gaze and pose tracking, gesture analysis, voice-to-graphics conversion, and graphical scene generation. Our research addresses fundamental technical issues of those components, and integrates them for proof-of-concept.

2. iTutor System Development

iTutor system is designed to use a synthesized graphical avatar as a virtual instructor to interact with a user. It is capable of understanding the user's expressions, eye gazes, head pose, and reactions, and respond accordingly at various levels. The graphical visualization tools and user-friendly virtual tutor could potentially allow for the learning experience interesting, engaging, individualized, and portable. The key components of the system include algorithm and software development of real-time estimation and recognition of head pose, eye gaze, expressions, voices, and expression synthesis. To do so, we propose to develop a series of novel algorithms, including scene flow and a generic 3D model based 3D pose estimation, 3D iris estimation, and shape index based expression classification.

The system is composed of 3D face model synthesis, facial expression recognition, pose and gaze estimation, and speech recognition. Figure 1 illustrates the general framework of the system. Figure 2 shows a detail composition of a case study.

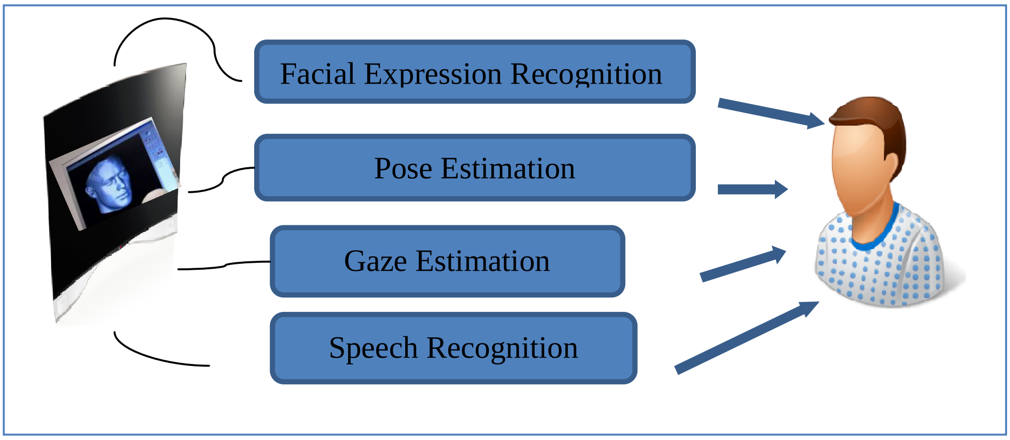

Figure 1: iTutor system composition

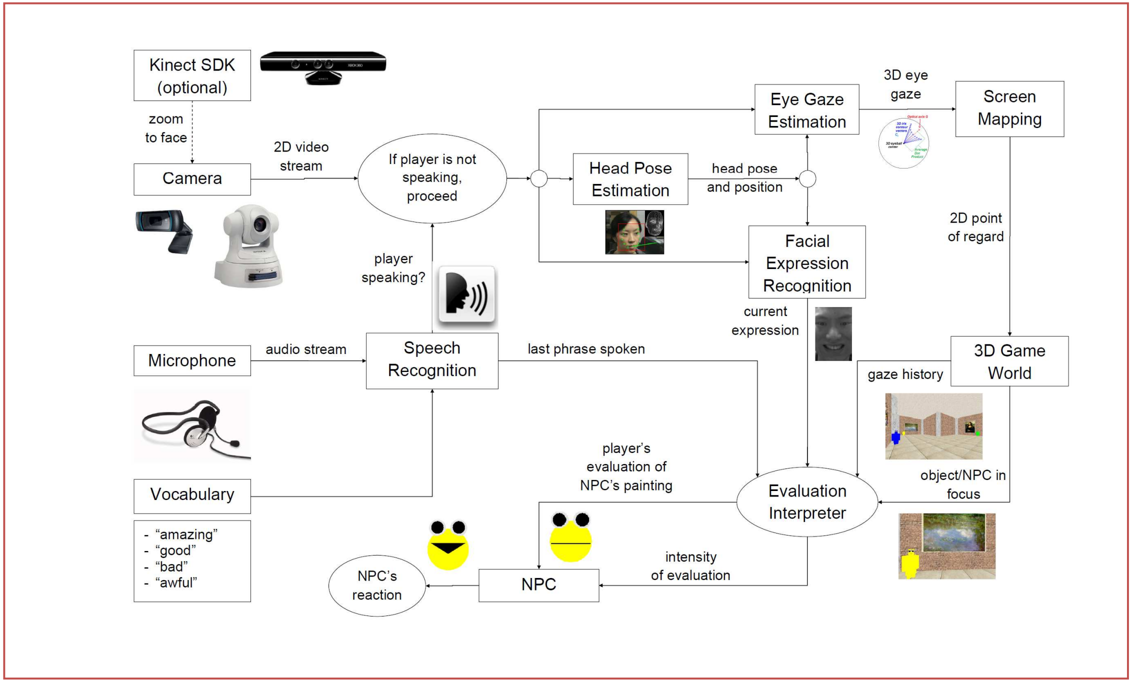

Figure 2: iTutor application case: multi-modal interaction for art critic

We propose to use a simple avatar user interface, and utilize a 3D model of the eye for eye gaze estimation. The head pose is estimated using prior knowledge of the head shape and the geometric relationship between the 2D images and a 3D generic model. The expression recognition module leverages the 2D shape index dynamic texture approach based on the *LBP-TOP* algorithm The Pocketsphinx speech recognition library is employed to recognize words and phrases. We use an art critic as a case study. The system checks whether the user has looked or is looking at a given avatar. When the user looks at the work displayed in the screen, makes a facial expression, and speaks his/her evaluation of the work, the avatar reacts simply based on the user's expression and speech.

2.1 Face avatar model creation

To construct 3D model avatars, we created a large set of 3D dynamic facial expression database, which has 41 subjects exhibiting various expressions. Our 3D dynamic imaging system has two grabbers for two stereo videos' capture and one grabber or a color texture video capture. A master machine is used to control the three-video capture, synchronization, and storage of the raw data. Three slave machines are configured to process the data in parallel for 3D model reconstruction.

To elicit target emotional expressions and conversational behavior, we developed a novel protocol for data collection. The protocol has been approved by the IRB board of Binghamton University. All sessions were conducted by an interviewer, who is a faculty of Binghamton University. The tasks include face-to-face interview, social games, documentary film watching, and other activities. These methods evoke a range of authentic emotions in a laboratory environment.

After participants gave informed consent to the procedures and permissible uses of their data, the experimenter explained the general procedure. After each task, participants completed self-report ratings of their feelings using a user-friendly tablet interface. The rating includes their emotions that have been experienced and the intensity of the emotions during the experience of the task.

3D avatar data with dynamic expressions were created. Participants were recruited from the BU campus. The ethnic background ranges from Asian, African-American, Hispanic, and Euro-American. The database has been released to the public for research purpose. The work has been published in the Journal of Image and Vision Computing [1] (X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A.Horowitz, P. Liu, and G. Girard, BP4D-Spontaneous: A high resolution spontaneous 3D dynamic facial expression database, *Image and Vision Computing,* 32 (2014), pp.692-706). Figure 3 shows one example of 3D dynamic avatar models. The models have been used in our courses (visual information processing and computer graphics) as new course materials for students to design and conduct new term projects for tracking, animation, and recognition.

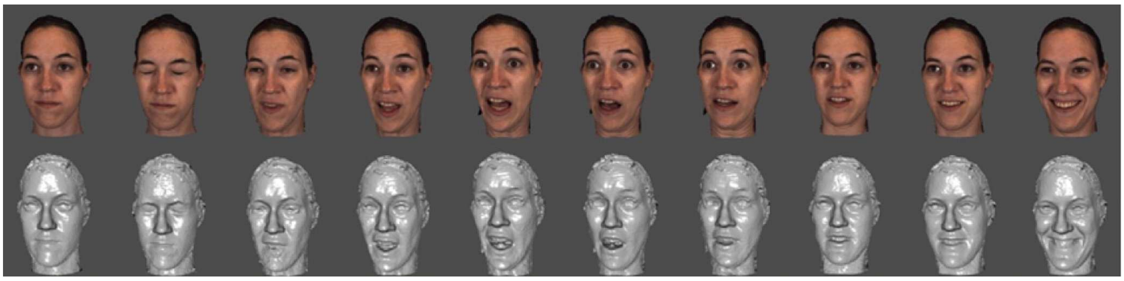

Figure 3: Example of 3D dynamic avatar models

2.2 Face expression recognition

We propose a new approach using 2D Shape Index Based Dynamic Textures for facial expression recognition [1]. Conventional expressions are classified. Dynamic textures (DT) encode texture information across space and time. In this case, these textures are constructed with concatenated Local Binary Pattern histograms from Three Orthogonal Planes (*LBP-TOP*). Basically, for a given image sequence, the LBP histogram for the middle image in the time sequence is computed to give us the XY plane histogram. With the X coordinate set to its center value, an “image” plane is constructed with all variations of Y and T (time) to give us the YT plane, and the LBP histogram is extracted from that as well. A similar process is performed for the XT plane. The histograms for each plane are normalized individually, and the concatenated histograms describe the texture in three dimensions.

Due to the good characterization of facial expression using topographic features, the shape index images are computed as input into the *LBP-TOP* algorithm. Shape index images are relatively robust to different lighting conditions.

We have further investigated the spontaneous expression analysis with head pose information and temperature information [4], and the application of expression analysis with occluded face in virtual reality environment [5].

2.3 Head pose tracking

Head pose is an important indicator of a person’s attention, gestures, and communicative behavior with applications in human-computer interaction, multimedia, and vision systems. The head pose estimation component uses a 2D video stream to determine head direction and position. First, the Viola-Jones approach is applied to detect the frontal face. After the face is detected, the Active Appearance Model technique is employed to detect and track predefined feature points on the face. The feature point coordinates in the 2D images are scaled and mapped to a 3D generic head model. Finally, based on the correspondence between the feature points, the 3D rotation angles can be calculated from the 2D coordinates by the so-called scene flow approach [1]. Figure 4 shows some examples of pose estimation.

Furthermore, we have also developed a new pose tracking approach for 3D data. Robust head pose estimation is challenging, particularly on this spontaneous 3D facial expression video data [7]. Most previous head pose estimation methods do not consider the facial expressions and hence are more likely to be influenced by the facial expression. In this project, we have developed a saliency-guided 3D head pose estimation on 3D expression models. We address the problem of head pose estimation based on a generic model and saliency guided segmentation on a Laplacian fairing model. We propose to perform mesh Laplacian fairing to remove noise and outliers on the 3D facial model. The salient regions are detected and segmented from the model. The salient region Iterative Closest Point (ICP) then register the test face model with the generic head model. The error rates of three rotations (yaw, pitch, and roll) are less than 4 degrees. This work has been published in the 15th ACM International Conference on Multimodal Interaction (ICMI), 2013 [7].

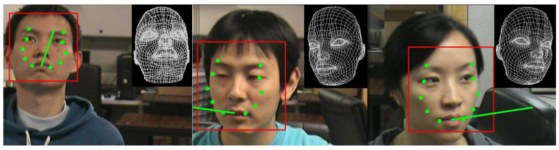

Figure 4: example of the feature points and head poses

2.4 Eye gaze estimation

We propose to determine the eyeball position by an offset from the 3D head pose and position. This offset is calculated from a calibration procedure. The eye detection algorithm maps the current camera image as a 2D texture onto the current position of the 3D eyeballs, rotates the eyeballs in pitch and yaw, renders each rotated eyeball, and picks the rotated eyeball that looks most like the user is looking into the camera. This is evaluated by 1) computing the absolute pixel difference of the center region of the rendered eyeball from a dark, circular template and 2) circle-fitting on the gradient magnitude image. Once the best eyeball rotation and scale are determined, the eyeball is rotated back and projected into image space, giving us our 2D iris center.

Each 2D contour point is converted to a 3D world vector, intersected with the current eyeball sphere, converted to a vector from the eyeball center, and normalized to give us an “iris contour vector”. As a result, the optical axis can be determined. In order to get the visual axis, a fovea offset computed during the calibration procedure is used. The fovea offset is rotated based on the rotation angles of the optical axis. The optical axis is then intersected with the eyeball sphere to get a new estimate for the 3D iris center, and the normalized vector from the fovea to this new iris center is the final gaze direction [1]. Figure 5 shows some sample gazes. Note that our new eye tracking algorithm has been patented with US patent #8,885,882 (granted in November 2014).

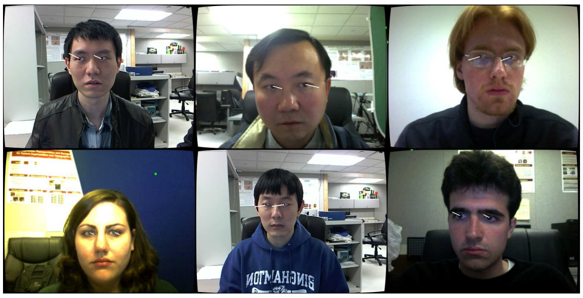

Figure 5: Sample gazes

2.5 Speech recognition and synthesis

To recognize the user’s verbal evaluations and to allow the system to respond with speech, our system has both speech recognition and synthesis components. The speech recognition module makes use of CMU’s Pocketsphinx software, and the synthesis module uses the Festival library. The speech component starts listening as soon as the user begins speaking and stops when the user is silent for more than 1 second. It then extracts Mel-frequency cepstral coefficients (MFCCs) to form the feature vector for the given audio sequence. Given the feature vector, an acoustic model is used to find the “senones” (effectively a complex phone or class of sounds), while a dictionary maps these senones to words. A language model can help ensure that the final word sequence makes sense. The speech recognition module runs concurrently with the main program as a separate thread, allowing simultaneous recognition of all signals.

2.6 Evaluation

To test the effectiveness of our system in the application, we performed a quantitative evaluation with the “Art Critic” interaction system (Figure 6). Each user was asked to evaluate paintings using every combination of facial expression and verbal evaluation in sequence; this was done 3 times with each user. Six subjects were tested, giving us a total of 432 samples. The results show that our facial expression recognition component with other signals in a live context performs well at accuracy of over 95%.

After each user finished playing the game, we asked them to fill out a questionnaire about the experience. The questions focused on whether each component as well as overall system made the interaction of learning comfortable, more immersive, effective, and/or fun. A 5-point scale was used: “Strongly Disagree” (1), “Disagree” (2), “Neutral” (3), “Agree” (4), and “Strongly Agree” (5). An option for “Not Sure” was also included, but it was not used by any of the users. The average and standard deviation of the answers from the questionnaires are 4.5 and 0.7, respectively. The evaluation shows positive feedback on the system developed. All the components have an average of at least 4, and most of them exceed 4.2, which demonstrate the positive experience the system generated. The concept of the system and its application were introduced in the conferences CIT 2014 [11], CIT 2015 [12], and NARST 2015 [10], respectively.

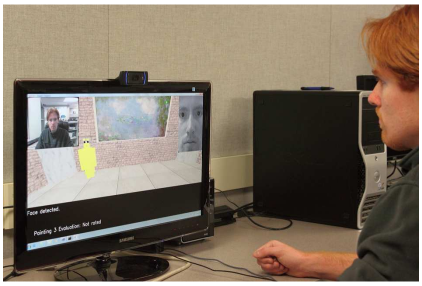

Figure 6: Example of iTutor system of learning based on human-computer interaction

3. iDemo System Development

iDemo system includes two components: speech-to-text-to-graphics software for scene and object composition and visualization, and gesture based interaction for control of visualization. The speech-to-graphics program tool can present and interpret virtual scene and objects intuitively. Such software allows a structure described by a user being visualized in a graphic display, thus making a better way to present ideas and concepts, and further facilitating the learning and training process.

3.1 Graphical scene generation

Automatic scene generation using voice and text offers a unique multimedia solution to classic storytelling and human computer interaction with 3D graphics. Manual scene composition in 3D is a time-consuming and intensive process. Interacting with and using 3D graphics based media and applications require that users adopt and learn specific graphics tools and interfaces. Such a requirement limits the number of potential users and their interactions with a given media. Automatic scene generation reduces such requirements making 3D graphics more accessible to users in non-graphics domains. Using natural language descriptions in conjunction with automatic scene generation enables non-graphics areas to take advantage of the benefits of visualization in 3D while avoiding the need for graphics specific knowledge. Additionally, scene generation is ideal for collaborative and interactive applications.

We developed a system for automatically generating 3D scenes from voice and text descriptions. Our method was designed to support widely varying quality polygon models. We concentrate on supporting a set of common spatial relationships that can serve as a basis for more complex relationship types. To illustrate our method, an initial framework for automatic scene composition using text and voice input phrases was developed. This part includes Bounding box localization, object mesh voxelization, and determination of object placement.

Our current framework composes scenes made up of two to a few objects and one or more spatial relationships. We used a speech recognition engine and a natural language parser to parse the voice and text input. We concentrated on supporting a core set of spatial relationships without any a priori knowledge of the objects. A preliminary framework for scene composition was built. Figure 7 shows an example of scene composition and visualization.

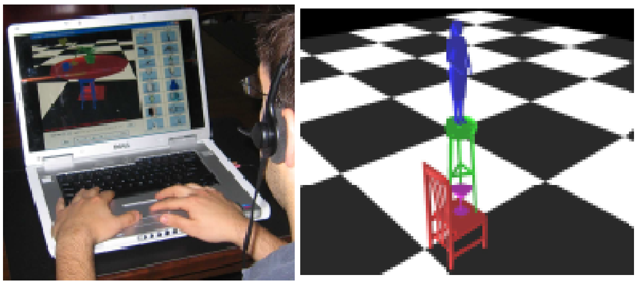

Figure 7: Example of scene generation system for illustration

3.2 Gesture analysis

An ideal demo system with human-computer interaction should function robustly with as few constraints as those found in human-to-human interaction. It is also expected to map human gestures to application control in the most natural and intuitive way possible. We developed a system for 3-D hand pointing gesture estimation and hand gesture recognition to interact with computer and control the visualization of graphics scene. The pointing gesture can resolve ambiguities springing from verbal communication, thus opening up the possibility of humans interacting or communicating intuitively with computers or robots by indicating objects or pointed locations either in 3-D space or on the screen [1].

We developed a novel approach to analyze the topological features of hand postures at multi-scales. Since many postures do not show explicit “holes”, we compute the convex hull of the hand region and consider the complementary space of the hand as holes. We use the multi-scale Betti Numbers matrix inspired by Persistent Homology to describe the multi-scale topological features of the hand posture. The system can detect 12 difference gestures with over 90% accuracy. This work has been published in IEEE International Conference of Computer Vision (ICCV) 2013 [6]. We have further improved the recognition robustness by extending the features to multiple types and multiple layers of geometric shapes. It has also been published in the IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2015.

As an example, Figure 8 Illustrates a Geographic Information Visualization application, where a user can select a region and zoom in or change the scene with the hand cursor and hand gesture.

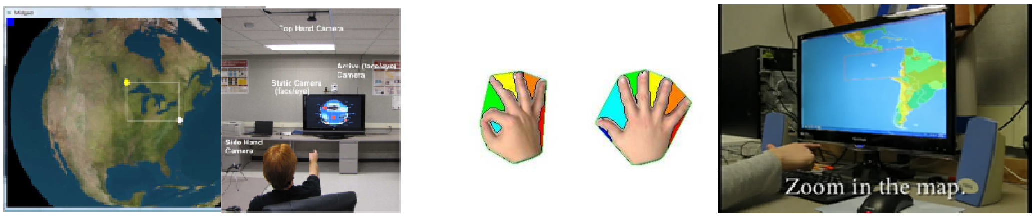

Figure 8: Example of hand gesture based control and interaction for graphical visualization

Note that our new hand pointing technique has been patented with US Patent #8,971,572 (granted in February 2015).

3.3 Evaluation

We have conducted a preliminary evaluation on the system usage. Six subjects were asked to use the system with hand gesture control for graphical scene presentation. We asked them to fill out a questionnaire about the experience. The questions focused on whether the overall system made the interaction comfortable, more immersive, effective, and/or fun. A 5-point scale was used: “Strongly Disagree” (1), “Disagree” (2), “Neutral” (3), “Agree” (4), and “Strongly Agree” (5). The average and standard deviation of the answers from the questionnaires are 4.0 and 0.5, respectively. The evaluation shows the positive experience the system generated. Each component has been discussed in the course of visual information processing and the course of computer graphics.

4. Future Development

(1) The algorithms and systems need be further refined to increase the usability, reliability, generalizability, realism, and fidelity.

(2) We will conduct a large scale evaluation with regards to the usage and efficacy of the systems for both learning and teaching enhancement;

(3) We will apply the new system to the undergraduate and graduate research program, and class instruction.

(4) Potentially, future development could extend the system as a simulation tool for training teachers, participants, students, and other professionals and practitioners.

Publications:

1. M. Reale, P. Liu, L. Yin, and S. Canavan, Art Critic: Multisignal Vision and Speech Interaction System in a Gaming Context, *IEEE Transactions on System, Man, and Cybernetics – Part B*, vol. 43, No. 6, p1546-1559, Dec. 2013
2. X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A.Horowitz, P. Liu, and G. Girard, BP4D-Spontaneous: A high resolution spontaneous 3D dynamic facial expression database, *Image and Vision Computing,* 32 (2014), pp.692-706 (*Note: this paper has been selected as one of 25 representative biometrics papers in the past six years from eleven different Elsevier journals. It has also been published in the virtual special issue of Biometrics 2014, Pattern Recognition Journal, Oct., 2014, Elsevier.)*
3. S. Canavan, P. Liu, X. Zhang, and L. Yin, Landmark Localization on 3D/4D Range Data Using a Shape Index-Based Statistical Shape Model with Global and Local Constraints, accepted by *Computer Vision and Image Understanding* (Special issue on Shape Representations Meet Visual Recognition), Elsevier, 2015.
4. P. Liu and L. Yin, "Spontaneous Facial Expression Analysis Based on Temperature Changes and Head Motions", *The 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG15),* 2015
5. X. Zhang, U. Ciftci, and L. Yin, Mouth Gesture based Emotion Awareness and Interaction in Virtual Reality, *ACM SIGGRAPH* (poster program), Aug., 2015
6. K. Hu and L. Yin, "Multi-scale topological features for hand posture representation and analysis", *14th IEEE International Conference on Computer Vision (ICCV)*, December 2013.
7. P. Liu, M. Reale, and L. Yin, "Saliency-guided 3D head pose estimation on 3D expression models", *15th ACM International Conference on Multimodal Interaction (ICMI)*, December 2013.
8. K. Hu and L. Yin, "Multiple Feature Representations from Multi-Layer Geometric Shape for Hand Gesture Analysis", *The 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG15),* 2015
9. X. Zhang, Z. Zhang, D. Hipp, L. Yin, and P. Gerhardstein, "Perception Driven 3D Facial Expression Analysis Based on Reverse Correlation and Normal Component", *AAAC 6th International Conference on Affective Computing and Intelligent Interaction (ACII 2015)* (Association for the Advancement of Affective Computing), 2015
10. A. Cavagnetto, R. Lamb, B. French, O. Adesope, L. Yin, and M. Taylor, ”A Potential Future in Education: The Application of Intelligent Systems in Teacher Education", *Proceedings of the International Conference of National Association for Research in Science Teaching (NARST),* Chicago, IL. 2015
11. Presentation and Talk "3D computer vision and graphics based intelligent interaction system for enhancing teaching and learning" at the 23rd Annual Conference on Instruction & Technology (CIT 2014) at Cornell University. May 2014. SUNY Faculty Advisory Council on Teaching & Technology (FACT2)
12. Presentation and Poster "Applying an intelligent simulation system for improving teacher education" at the 24th Annual Conference on Instruction & Technology (CIT 2015) at SUNY Geneseo. May 2015. SUNY Faculty Advisory Council on Teaching & Technology (FACT2)