INTERACTIVE STOREFRONT DISPLAY WITH FACE AND GESTURE CONTROL

Colin Bellmore, Raymond Ptucha, and Andreas Savakis

Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY

RELEVANCE

We present an interactive storefront display system that incorporates body gestures, head pose and facial expressions. The Kinect depth sensor is used to detect and track an observer's skeletal joints while the RGB camera is used for facial analysis. To demonstrate the applicability of our system to retail applications, a virtual bakery display controllable by simple gestures is introduced. A pointing gesture is linked to the cursor, so the user can point at the location of a display item to bring up additional information. A selection gesture motivated by American Sign Language is performed in combination with pointing to place the selected item into a virtual shopping basket. A wave gesture initiates checkout, which removes all selections from the shopping basket and displays them on the screen.

USEFULNESS

The interactive storefront display is used to simulate retail purchases without the need to enter a store or physically touch any device. The system uses a natural pointing gesture to browse store products, gather more information, and place orders. The system could be placed inside a store to preorder items before arriving at a register, or outside a business to attract new customers. The expression and pose recognition technology may be used to gather additional information about which items are most popular and which experiences users enjoy. We demonstrate the interactive storefront concept through a bakery display where a customer browses and selects items for purchase using intuitive arm and hand gestures. With future improvements, the system could collect demographic information on the customer base and offer personalized service.

TECHNICAL CONTRIBUTION

The availability of joint information through Microsoft's Kinect SDK [1] allows fast and effective interpretation of simple gestures. In this interactive display system, we use three gestures shown in Figure 1: pointing, selection and checkout. The pointing gesture is determined by the line formed by the hand and shoulder joints. The intersection of the hand-shoulder line with the plane of the display corresponds to the target of the point gesture.

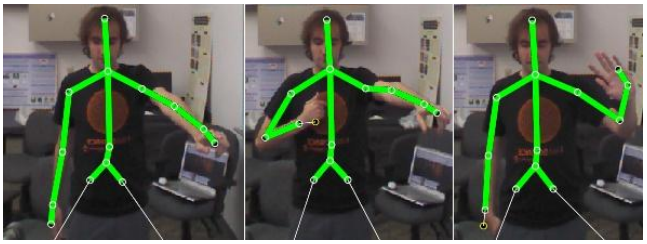

Figure 1: Gestures from left to right: pointing, selection, and checkout.

To transform the intercept location from physical measurements to pixel coordinates on the display, the position of the Kinect camera relative to the display must be known. Conversion to screen pixel coordinates is accomplished in the vertical and horizontal directions separately, given by equations below.

$$T_x = \frac{(P_x - X_1)}{(X_2 - X_1)} \cdot D_x \text{ and } T_y = \frac{(P_y - Y_1)}{(Y_2 - Y_1)} \cdot D_y \qquad (1)$$

The variables P_x and P_y are the gesture locations in the horizontal and vertical directions respectively. The variables X_1 and X_2 represent the distances between the Kinect and the left and right side of the screen, while Y_1 and Y_2 are the distances between the Kinect and the top and bottom of the screen. D_x and D_y define the resolution of the screen in pixels, where D_y is the vertical dimension and D_x is the horizontal dimension. T_x and T_y are the resulting coordinates of the gesture target in pixels.

The selection gesture was inspired by a variation of the American Sign Language gesture for 'self'. The gesture involves bringing the non-pointing hand to the center of the chest and tapping twice against the chest. This gesture provides a straightforward method of recognition when given the position of the hand and center of the body.

The wave gesture recognition uses a similar ray casting technique as the pointing gesture, and works by drawing a line between the elbow and hand joints. The target intersection surface is defined as a plane parallel to the floor, and the recognition algorithm acts as a state machine where the transitions are defined by threshold crossings. This allows many combinations of relative elbow-hand movements to be recognized as a wave. Use of the vertical plane also ensures that this gesture does not interfere with

the pointing or selection gestures.

The face region is identified using the skeletal joints and their distance from the Kinect device. The Microsoft Face Tracker SDK [1] is used to determine the pose and facial feature points of the observer. Head pose is used to control the cursor when the hand selection gesture does not give a valid target on the display, e.g. when the user has his/her hands behind their back. The target location on the display is found using a line from the nose, projecting perpendicular to the face. This rough selection estimate allows the user to interact with the display immediately, without learning the required gestures.

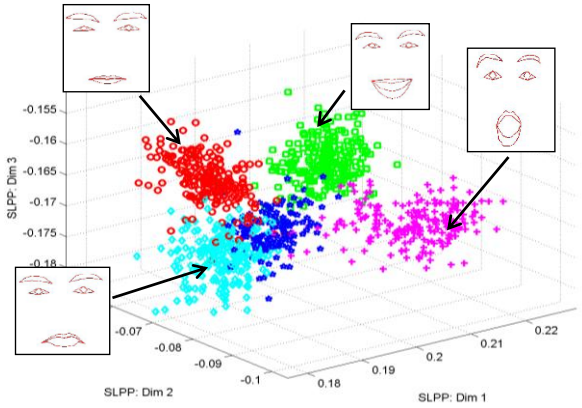

Figure 2: Sample manifold space trained to recognize five emotions: Angry (red), Sad (cyan), Neutral (blue), Happy (green), and Surprised (magenta).

Facial expression is captured to provide information about the user's enjoyment of the selected item and the experience. The RGB facial bounding box area is converted to a luminance image, whereby localized eye and mouth corner points are used to define an affine warp to a canonical face representation. Manifold learning [2] is used to reduce the dimension of the input data by identifying a non-linear low dimensional space where the data resides. In order to support the extension of the manifold model to new examples, linearized techniques such as the Locality Preserving Projections (LPP) [3] solve a linear approximation of the non-linear object and have been successfully applied to expression recognition.

Figure 2 shows the mapping of 1072 expressive faces into low dimensional supervised LPP space [4] where a multi-class linear Support Vector Machine classifier is used to estimate expression. Similar methods have been used by the authors to estimate gender, age, and ethnic background. Such information may be used for demographic data collection and personalized service.

NOVELTY

Combining pose, expression, and gesture recognition into an interactive display offers new opportunities to experiment and improve the user experience. Each piece in the framework can be adjusted independently and the navigation gestures could easily be adapted to other large format interfaces. The use of both head pose and hand-shoulder ray casting makes the display more accessible compared to systems that use only one type of input. The simple and intuitive nature of the wave gesture allows effective recognition of individual variations of waving at low computational cost.

Figure 3: Interactive display interface with active region (Black Forest Cake) selected.

QUALITY OF IMPLEMENTATION

A similar display system [5], shown in Figure 3, was demonstrated at the ImagineRIT innovation festival, which is open to the public, and was well-received by users for its ease of use and fun factor. The small number of issues encountered stemmed from misplaced body joints caused by long hair, hand bags, and baggy clothing. Most users easily learned to navigate and use the display by watching other users or by simple verbal instructions.

ACKNOWLEDGEMENTS

This research is supported in part by Cisco, and the NSF Graduate Research Fellowship Program.

REFERENCES

[1] Microsoft Kinect SDK, URL: http://www.microsoft.com/en-us/kinectforwindows .

[2] A. Ghodsi, "Dimensionality Reduction A Short Tutorial," ed. Ontario, Canada: Department of Statistics and Actuarial Science, Univ. of Waterloo, 2006.

[3] X. He and P. Niyogi, "Locality Preserving Projections," in *Advances in Neural Information Processing Systems 16*, Vancouver, Canada, 2003.

[4] R. Ptucha and A. Savakis, "Facial Expression Recognition Using Facial Features and Manifold Learning," *Int. Symp. Visual Computing, ISVC*, Las Vegas, NV, Nov. 2010.

[5] C. Bellmore, R. Ptucha, and A. Savakis, "Interactive display using depth and RGB sensors for face and gesture control," *2011 IEEE Western New York Image Processing Workshop*, Rochester, New York, 2011.