M.I.T. Media Laboratory Vision and Modeling Group Technical Report No. 312
(Appeared, Intl. Workshop on Automatic Face and Gesture Recognition, Zurich, July 1995)

Attention-driven Expression and Gesture Analysis in an Interactive Environment

Trevor Darrell and Alex P. Pentland
trevor,sandy@media.mit.edu

Abstract

To provide natural user interfaces to interactive environments, accurate and fast recognition of gestures and expressions is needed. We adopt a view-based gesture recognition strategy that runs in an unconstrained interactive environment, which uses active vision methods to determine context cuess for the view-based method. Using vision routines already implemented for an interactive environment, we determine the spatial location of salient body parts and guide an active camera to obtain foveated images of gestures or expressions. Face recognition routines used to obtain an estimate of the identity of the user, and provide an index into the best set of view templates to use. The resulting system combines low-resolution, user-independent processing with high-resolution, user-specific models, all of which are computed in real time as part of an interactive environment.

1 Introduction

Gesture and expression are important interface modalities for interactive environments. Previously, we developed a method for view-based recognition of spatio-temporal hand gestures [1] and a similar mechanism for the analysis/real-time tracking of facial expressions [3]. These methods offered real-time performance and a relatively high level of accuracy, but required user-dependent training of model views, and a high-resolution image of the hand or face. There are many domains/tasks for which these are not unreasonable assumptions, such as interaction with a single user workstation or an automobile with a single driver. However the method had limited usefulness in unconstrained domains, such as "intelligent rooms" or interactive virtual environments, when the identity and location of the user are unknown.

Interactive virtual environments provide perhaps the greatest generalization of the notion of a user interface, as they seek to "immerse" a user into a machine-made world. A user should be able to interact with objects or agents in these environments in as natural a manner as possible. Previously, we have explored the use of computer vision routines as non-invasive interfaces to virtual worlds. In this approach visual routines analyze images of users and allow them to interact with a virtual world using both direct manipulation and gestural interfaces.

Our ALIVE system, which provides a video-based environment for interacting with artificial life creatures, implemented vision routines to find a user in a room and locate certain salient body parts [2]. The ALIVE system assumed little prior knowledge about the user, and operated on coarse-scale images of the user. A simple mechanism for recognition of hand gestures was implemented in the ALIVE system, which made no use of high-resolution view models, and could only recognize pointing and waving motions definied by the motion of the centroid of the hand. To extend this work to accommodate a broader range of gestures through the use of a view-based approach, we obtain high-resolution images of the users hand, and estimates of users identity. In this paper we propose attention mechanisms based on visual routines in the ALIVE system which can provide the spatial and model-indexical context that view-based gesture recognition methods require to be successfully applied.

2 Context-dependent Gesture Analysis

Several approaches to the problem of recognizing hand and/or face gestures have been proposed recently. In general, one can distinguish between methods which assume a physically valid model of the hand or face, and those which do not extract or impose these types of 3-D constraints. While 3-D models can provide powerful prior models for interpretation, we have investigated the latter approach, using an ensemble of 2-D models to represent a complex articulated object as it undergoes a particular gesture. This approach offers feasible real-time performance using currently available computational resources.

In our approach, a set of view-based correlation models is used to represent spatio-temporal gesture patterns. We take a sequence of images representing the gesture to be trained, and build a set of view models that are sufficient to track the hand as it performs the gesture to be recognized. Our view models are normalized correlation templates; in the examples presented here we use intensity-based models, but it is also possible to use band-pass or wavelet-based models. The latter would have the advantage of being less dependent on illumination direction in addition to being invariant to the overall level of illumination.

Figure 1(a) shows two training sequences of a hand waving; from these sequences a set of normalized correlation view models are found using an unsupervised clustering procedure. Full details of the clustering procedure are in [1]; in essence we use a "leader" algorithm, in which the first image is used to create a view model, and subsequent images also used to create new view models when the largest correlation score in the current set of view models drops below some threshold. Using a normalized correlation threshold of $\theta = 0.7$ and the two waving sequences as

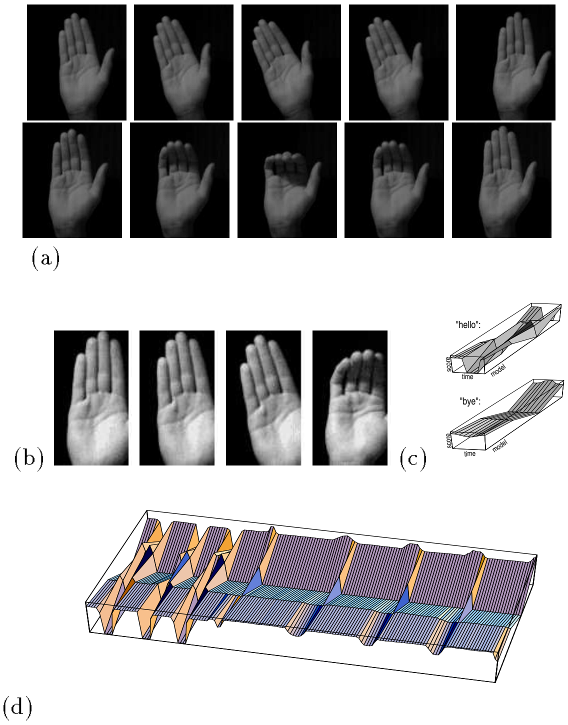

Figure 1: Gesture recognition using correlation-based view models. (a) Frames from two image sequences of a hand performing waving gestures. (b) "View-based" correlation templates. (c) Correlation scores of templates for each frame of sequences. (d) Correlation scores for new data containing three of each gesture. Adapted from [1].

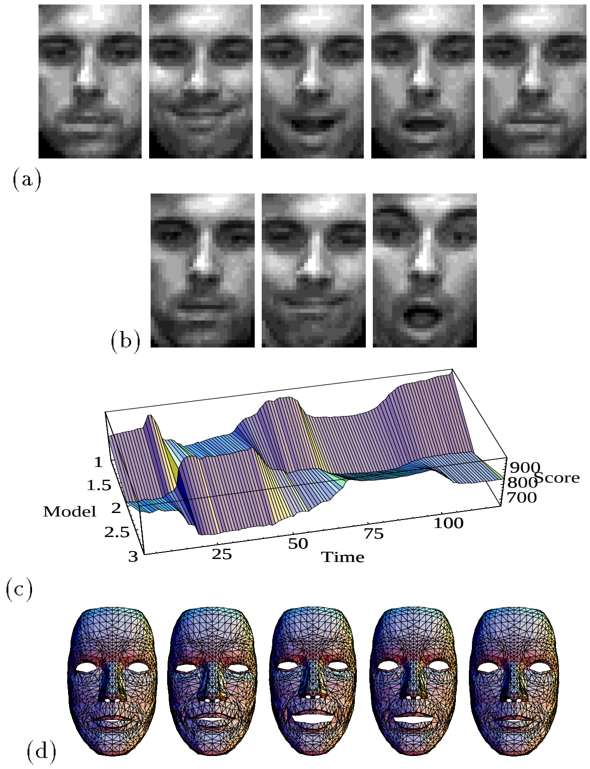

Figure 2: View-based analysis/synthesis of facial expression. (a) Sequence of face images performing expressions. (b) Expression view models. (c) View model scores. (d) Interpolated 3-D model faces. Adapted from [3].

input, four view models were found to be sufficient to track the hand through the two waving gestures. Figure 1(b) shows these four templates.

The view models can be thought of as a dimensionality reduction mechanism, since for a sequence of new images it becomes sufficient to examine the vector of correlation scores of the view models to detect gestures. Figure 1(c) shows the scores for the two waving sequences, displayed as a surface plot.[1] Figure 1(d) shows the view model scores for an input sequence which contained three instances of each gesture. One can see that in this representation, recognition is straightforward. In [1] we reported recognition accuracy of over 90% in tests conducted on multiple users, given assumed conditions of high-resolution (foveated) hand images, little global rotation, and user-dependent templates.

We have also applied the view based approach to the task of tracking facial expressions [3]. Figure 2(a) shows a sequence of faces performing a surprise expression. Figure 2(b) shows the view models learned for this sequence, and Figure 2(c) shows the normalized correlation score for these models. For the task of facial tracking, we implemented an interpolation paradigm to map vision input to motor control output dimensions, using the facial model presented in [4].

Interpolation requires a set of control points or exemplars from which to derive the desired mapping. In this example pairs of real faces and model faces for different expressions are presented to the interpolation method during a training phase, by generating a 3-D model face and asking the user to match it. We use the Radial Basis Function (RBF) method presented in [7], and define the interpolated motor controls to be a weighted sum of radial functions centered at each example:

$$\mathbf{Y} = \sum_{i=1}^{n} c_i \mathcal{G}(\mathbf{X} - \mathbf{X_i}) \tag{1}$$

where $\mathbf{Y}$ are the model face parameters, $\mathbf{X}$ are the observed view-model scores, $\mathbf{X_i}$ are the example scores, $\mathcal{G}$ is an RBF (and in our case was simply a linear ramp $\mathcal{G}(\S) = ||\S||$), and the weights c_i are computed from the example motor values $\mathbf{Y_i}$ using the pseudo-inverse method.

With this simple formalism, we were able to track expressions of a real user and interpolate equivalent 3-D model faces (Figure 2(d)) in real time. However, as with the hand tracking example, this face tracking method requires high-resolution images of faces and user-dependent face templates. To apply the method

[1] The axes of this surface plot are as in Figure 2(c); the ordering of the view model dimension in these plot is arbitrary, however structures are easier to see in this display than in four line graphs.

Figure 3: The ALIVE system for vision-based interaction with a virtual environment. (a,b) A user sees him/herself in a "magic mirror", composited in a virtual environment. Computer vision routines analyze the image of the person to allow him/her to effect the virtual world through direct manipulation and/or gestural commands. (c) Results of hand and head tracking routine. The resolution of hand images from the single camera in previous versions of ALIVE was insufficient to apply the view-based recognition techniques; hand gestures were modeled only using the motion of the hand location. Adapted from [2].

in unrestricted interactive domains, we need to find mechanisms which can both find the face and provide cues for detecting who the user is. Fortunately such mechanisms have already been implemented in our interactive video environment, described in the next section.

3 Interactive Environments

We have implemented an environment that uses computer vision-based routines to allow a user to interact with virtual creatures. ALIVE, an Artificial Life Interactive Video Environment, can provide the contextual environment within which view-based gesture analysis methods can be successfully applied.

ALIVE presents a user with a "magic-mirror" metaphor, in which the user sees him/herself presented in a mirror along with virtual creatures. This allows a completely non-invasive (no wires, gloves, goggles) mode of virtual interaction. A wide field-of-view video camera acquires an image of the user which is combined with computer graphics images and projected on a large screen in front of the user (Figure 3(a,b)).

As described in [2], computer vision routines in ALIVE analyze the image of the user, compute figure/ground segmentation, and analyze the contour/silhouette of the user to determine the location of head, hands, and other salient body features. We use only a single, calibrated, wide field-of-view camera to determine the 3-D position of these features. By assuming the the user is sitting or standing on the ground plane, we use the imaging and ground plane geometry to compute the location of the bounding box of the user in 3-D. To do this, we project a ray from the camera through the 2-D projection of the bottom of the bounding box of the user. If the user is indeed on the ground plane, the intersection of the projected ray and the ground plane will establish the user's 3-D location.

The vision routines in ALIVE can locate the head of a user and crop it from the scene, but the images of the head obtained in this manner are too low-resolution to apply the view-based expression analysis methods. What we need is a foveated imaging sensor, which can obtain a high-resolution image of the head or hand.

4 Attention Framework

To provide high resolution images for gesture recognition, we augment the existing wide field-of-view camera in our interactive environment with an active, narrow-field-of-view camera. Information about head/hand location from the existing ALIVE routines is used to drive the motor control parameters of the narrow field camera.

In our augmented ALIVE system, figure/ground segmentation and contour analysis routines are run on the image of the user, and determine head or hand location (Figure 3(c)). The head location is translated into gaze angles for the active camera's motor system, and a foveated image of the body part is acquired. Face recognition routines [5, 6] are then run on the foveated image to estimate the identity of the user and select the appropriate set of view-based templates to use when multiple users are present. Figure 4(a) shows the overall architecture of the active imaging system we have implemented in ALIVE. Figure 4(b,c) shows example output from the wide angle camera and the narrow angle camera, as the narrow camera tracked the users head given the head position information computed by the ALIVE routines on the wide angle image.

We have integrated the active vision sensor with the view-based gesture analysis method in ALIVE and show preliminary results which demonstrate the feasibility of the combined methods. Using a highly simplified, two expression model of facial expression (neutral and surprised), we tracked facial expressions as the user moved about the scene and the narrow angle camera followed the face. Figures 5,6 show the results of tracking expressions using the narrow angle camera input when the user is in two different locations in the scene, using a single set of view models. The view models were acquired at a location in the scene different from the two locations where we ran these tracking experiments.

For each of these experiments a surprise measure was interpolated from the view scores using the ra-

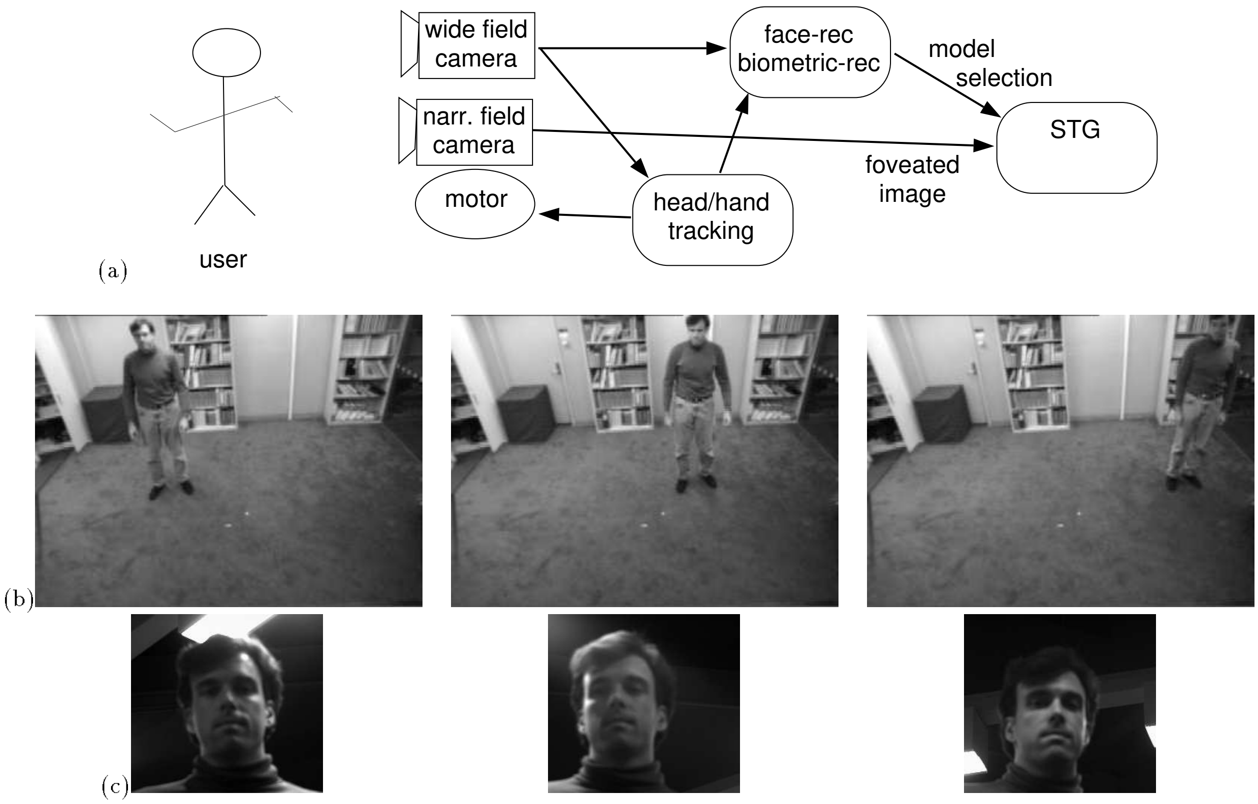

Figure 4: Combination of gesture analysis and interactive vision systems. (a) Active vision architecture for augmented ALIVE system. (b,c) Images acquired from wide and narrow field of view cameras as user moved across scene and narrow camera tracks head.

dial basis function method described in Section 2. To produce the interpolated surprise measures shown in these figures, we mapped the vision scores to a one-dimensional motor value, labeled “surprise”. Equivalently we could have mapped the vision scores directly to the motor controls corresponding to a “surprise” face, as was done in [3]. With this formulation, the distinction between expression recognition and expression tracking becomes blurred; the surprise measure can be used directly for animation, or peaks can be found and used for recognition/detection.

Finally, Figure 7 shows the plot of the two view model scores and the interpolated surprise measure for the entire run. During this run, the user began standing in the middle of the scene, made three surprise expressions, then moved to the left, back to the middle, and finally to the right of the scene, and repeated the three expressions at each location. In the graphs we can see that the view scores fall to zero as the user moves to a new location and the camera saccades to find the face again. When fixated on the face, the two fixed view-models extract useful information about the surprise expression, as is evidenced by the four sets of three peaks in the interpolated surprise measure. Each peak corresponds to the user performing the smile, which he did three times at each of the four locations in the scene.

Using previously implemented visual routines, we can find the size of the users bounding box and the location of their centroid within this bounding box in real terms, and compute a statistical model of these parameters for each user. Preliminary experiments have indicated that these cues can be used to separate between a population of known users, when the number of users is not too large. In addition, more sophisticated face recognition routines may be run on the foveated image [5]. Face recognition and biometric cues can then be combined into a single estimate of identity and used to perform model selection on the view models. The implementation of integrated face and biometric recognition is currently in progress in the ALIVE gesture recognition system.

5 Conclusion

Attention methods can offer considerable performance enhancement in recognition systems as processing is restricted to a particular subregion of physical space (spatial attention) and/or model parameter space (model indexing), without sacrificing ultimate accuracy of the estimated result. In this paper we have demonstrated preliminary results on the use of attentional mechanisms to provide the contextual environment for view-based spatio-temporal gesture analysis methods. By combining the early vision routines available in the ALIVE system with an active, high-resolution camera, we were able to implement view-based gesture analysis in an unconstrained in-

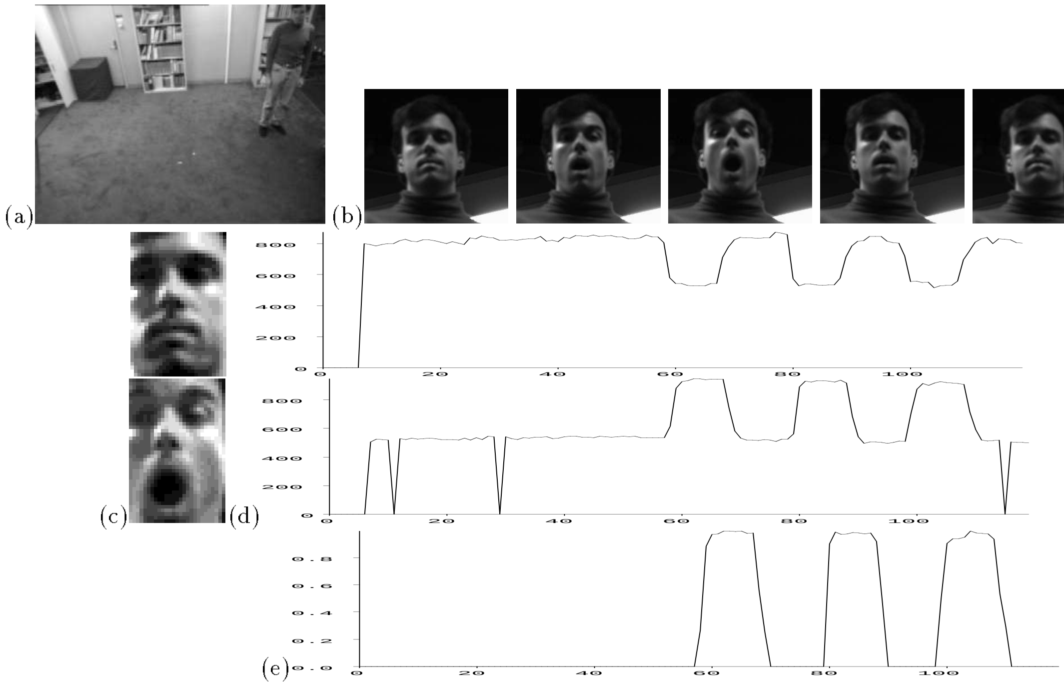

Figure 5: View-based expression tracking using foveated face images. (a) Wide-angle view of scene. (b) Foveated images of face while user performs "surprise" expression. (c) Normalized correlation "view" templates of neutral and surprise expression. Views were trained while user was at a different location in the scene. (d) Normalized correlation score of view templates evaluated on sequence in (b). User performed three surprise expressions in the sequence. (e) Plot of surprise measure interpolated from view template scores. Three peaks are present corresponding to the three surprise expressions.

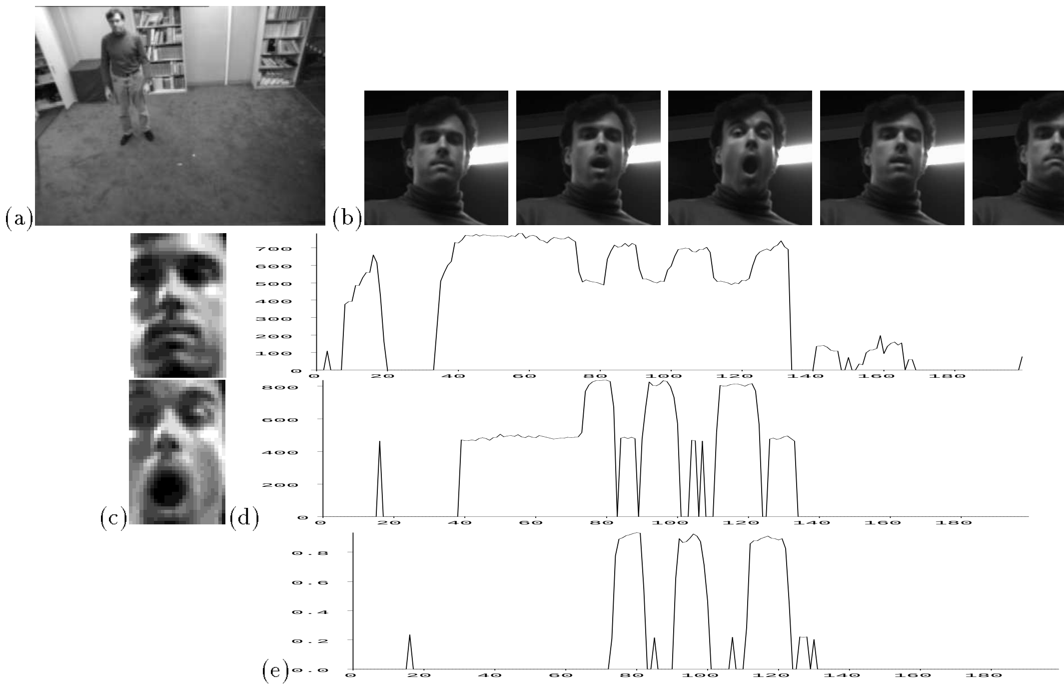

Figure 6: Same surprise expression performed at different location in scene, analyzed using same foveated view templates as as in previous figure.

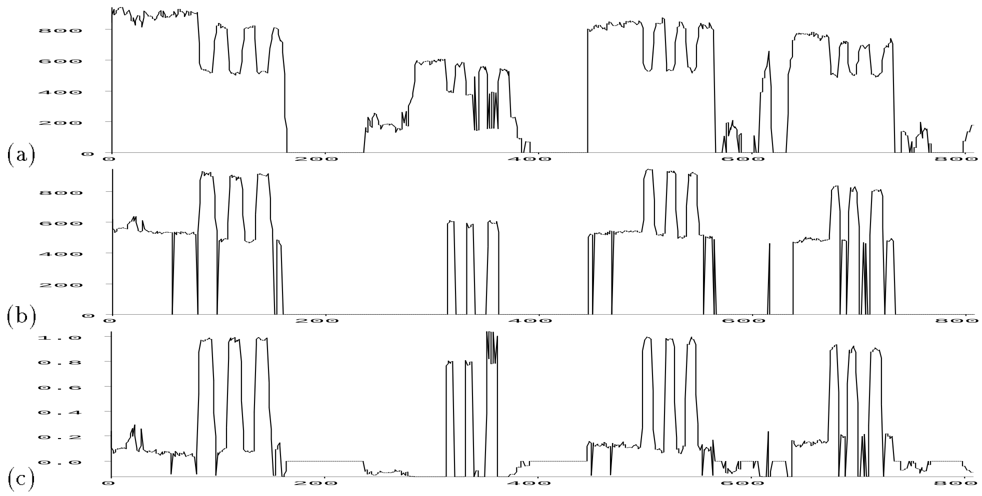

Figure 7: Results from extended run where user moved to multiple locations in scene. (a) Neutral view model score, (b) surprise view model score, and (c) interpolated surprise measure. The active camera followed the users face (scores drop to zero during camera motion), and the surprise measure picks up the three expressions the user performed at each location.

teractive environment. The use of view models and active vision offers performance that is both accurate and real-time, two necessary goals for interactive systems.

References

[1] T. Darrell and A. P. Pentland, Space-Time Gestures. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 335-340, IEEE Comp. Soc. Press, Los Alamitos, CA, 1993

[2] T. Darrell, P. Maes, B. Blumberg, A. P. Pentland, A Novel Environment for Situated Vision and Behavior, *Proc. IEEE Workshop for Visual Behaviors*, IEEE Comp. Soc. Press, Los Alamitos, CA, 1994

[3] T. Darrell, I. Essa, and A. P. Pentland, Correlation and Interpolation Networks for Real-time Expression Analysis/Synthesis, *Advances in Neural Information Processing Systems 7*, Morgan Kaufman 1995

[4] I. A. Essa. *Analysis, Interpretation, and Synthesis of Facial Expressions*, PhD thesis, Massachusetts Institute of Technology, MIT Media Laboratory, Cambridge, MA 02139, USA, 1994.

[5] Turk, M., and Pentland, A., Eigenfaces for Recognition, *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, 1991.

[6] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Computer Vision and Pattern Recognition Conference*, pages 84–91. IEEE Computer Society, 1994.

[7] T. Poggio and F. Girosi, A theory of networks for approximation and learning. MIT AI Lab TR-1140, 1989.