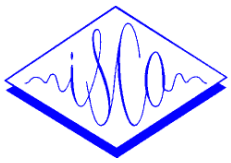

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Visual perception of human bodies and faces for multi-modal interfaces

Alex P. Pentland and Trevor Darrell
{sandy,trevor}@media.mit.edu
MIT Media Lab
20 Ames Street
Cambridge MA 02139 USA

1 Abstract

In this paper we describe recent work in our laboratory on the use of computer vision techniques for real-time multi-modal interfaces. The methods described here allow the non-invasive perception of human users; no special markers or identifying features are assumed. Both user-independent and user-dependent algorithms for gesture recognition are used, depending on the context. We apply the same techniques used for recognition to the problem of generation of animated forms to accompany spoken language. Both real-time recognition and animation of facial gestures (e.g., a lip-synched "talking head") have been implemented within our framework.

2 Introduction

Ideally, good user interfaces should be multi-modal; they should integrate information from a variety of senses. Vision and speech offer two modalities which are especially well suited for such integration, as the information provided by each is often complementary. This is evident both in the domain of low-level signal processing, where visually guided lip-reading can disambiguate phonetic classification [10], and in high-level processing, where the semantic information conveyed in the acoustic signal is often ambiguous without accompanying information from a physical body gesture, such as pointing [1]. Here we present work performed in our lab towards the real-time perception of human bodies and faces, and intersection of those techniques with issues in spoken language perception and generation.

In the following sections we overview a system which can isolate and track a user's body, recognize gestures performed by his/her hands and face, and animate a 3-D facial model which matches his/her expression, all in real-time. We employ a three stage framework in our system. The first stage is comprised of a simple figure-ground segmentation mechanism, which isolates a user from other objects in the scene by color or motion classification. The second stage implements an attention/tracking mechanism which localizes the user in 3-D and finds his/her head and hands. The third stage performs analysis and recognition of gestures and expressions using the localized hand and head information. The segmentation, attention, and tracking stages are user-independent and have been run on hundreds of users. The recognition and animation stages require some user-dependent training to learn the pattern of poses in particular gestures and expressions. Our system assumes a relatively unconstrained office environment, in which a user is seen by a camera in a arbitrary, but known, setting.

3 Low-level processing

The first stage in the analysis of body form and gesture is to isolate and size normalize the body from the background, using images acquired from a video camera. To do figure ground segregation, we employ simple chromatic and/or motion differencing methods, combined with connected-component "blob" analysis that finds simple shape statistics of the isolated form.

In general we model the background as an arbitrary, but static, pattern. Mean and variance information about the background pattern is computed using samples collected over a sliding time-window. By using these statistics to determine thresholds for pixel class membership, accurate figure-ground membership is possible [3]. If the scene conditions are such that static background subtraction is inappropriate, for example if there are multiple moving objects which cover large portions of the scene, then more sophisticated clustering methods are needed. In these cases motion-based grouping methods could be applied to find regions which are moving with coherent motions [4].

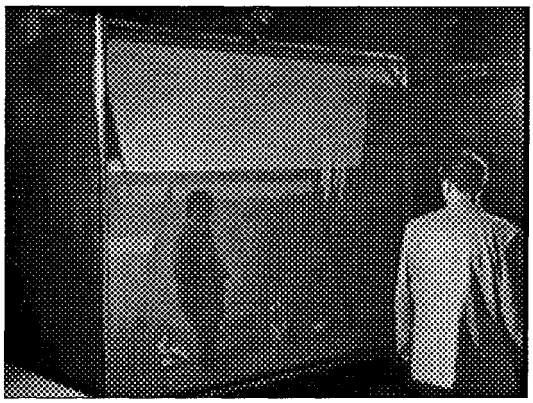

Figure 1: The ALIVE system: full-body wireless interaction with a virtual world

Once a difference signal has been computed we apply connected components analysis to find a foreground region. We binarize the difference image, find connected regions, and compute the image bounding box and first-order moments of the largest connected region. (For a review of connected components and binary image processing methods see [2] and [8].) When the figure of the user has been isolated from the background, we can compute its rough location in the world. If we assume the user is indeed sitting or standing on the ground plane, and we know the calibration of the camera, then we can compute the location of the bounding box in 3D using a straight-forward geometric projection.

Next, an attention mechanism localizes certain key features, such as the hands and head of a user, given certain assumptions about the imaging geometry and the possible poses of the user (Figure 2). The hand-tracking algorithm we have developed for this task is comprised of several different context-dependent search heuristics. In general, a normalized correlation search is done along the sides of the upper-torso bounding box to find a strong horizontal edge. The upper-torso bounding box is defined in such a way that it discounts the effect of shadows and feet in the horizontal dimension: it takes the vertical dimensions from the real bounding box and computes horizontal dimensions from the top 66% of the user's image.

Depending on the current context, different search windows and search patterns are used. The main contextual cue is the size of the bounding box, which provides a rough estimate of overall pose. If the box is narrow, we infer that the user's hands are at their side, and we do not attempt to find them in the silhouette. In this case we return the hand position to be located on the side of the bounding box, at the same relative height along the bounding box as it was last reliably seen.

This simple level of processing can produce interesting and entertaining vision-based interactive systems. The ALIVE system, demonstrated at SIGGRAPH 93, used this technology to allow a user to interact with an artificial life world in a completely wireless manner [9]. In this system a user saw him/herself on a large video screen, walking among the virtual creatures. Through the vision mechanism, the creatures could also "see" the location and pose of the user, and react and/or be acted upon accordingly. The system ran in real-time and had over 500 users during the week-long demonstration; most had successful (e.g. interesting or enjoyable) interactions with the creatures in the artificial world.

4 Hand/face models

For more sophisticated interface tasks, we need to have a model of a user's head/hand, beyond just a point location. Precise, 3-D models of human forms are possible to formulate, but in practice are difficult to estimate with any stability. Moreover the amount of complexity they provide is often unnecessary for the desired task. We have found that efficient and accurate characterization of pose is possible through the use of a representation which uses a distributed set of "views", 2-D canonical poses.

In our view-based approach, a set of views which covers the space (set) of poses used in a given task is constructed, and a signal represented by its distance to each view (using normalized correlation). (For domains where poses span a linear space, an eigenvector decomposition provides an optimal set of views [11].) A simple clustering algorithm can pick a set of prototype poses given a training sequence containing gestures or expressions of interest [5]. One example set of poses for a particular user is shown in Figure 4.

5 Gesture Recognition

Recognizing facial features can greatly aid recognition of spoken messages. In addition to the lip-reading results mentioned above, which provide information on labial articulation which is often obscured in the acoustic signal but is clear in the visual signal, several other facial cues are relevant for speech. In particular for the analysis of prosody there may be correlated visual expression cues, for example raising eyebrows in conjunction with a particular intonation structure. Our system provides real-time analysis and recognition of these features. Integrating expression cues from both speech and vision can provide for much more robust recognition rates than from either cue alone.

We employ two levels of gesture recognition; one that is user-independent and relies only on the point-locations provided by the attention/tracking method, and one that exploits the higher resolution, view-based representation that requires

Figure 2: Binary silhouettes of users after figure-ground processing in the ALIVE system. Task-dependent vision routines to find hands and pose information use this representation as input.

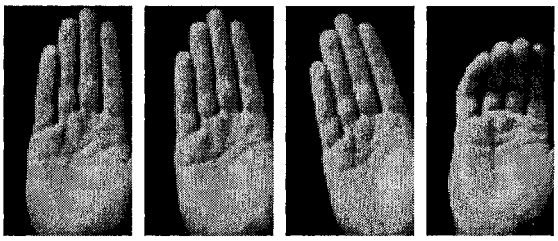

Figure 3: Four view models used to recognize a user's waving hand gesture. These models were acquired using an unsupervised clustering procedure given a training sequence containing the gesture.

user-dependent training as described above.[1]

With the user-dependent representation, we model gestures as spatio-temporal patterns, and cascade two sets of view models: first we extract spatial views and compute a spatial feature vector, and then "temporal" views whose domain is the spatial feature vector over time. Figure 3 shows the spatial view models acquired for a user performing a waving gesture. The dynamic time warping method is used in conjunction with the computation of distance for temporal views, so that variation in the temporal sampling and/or rate of production of a gesture does not effect recognition performance. As reported in [5] we are able to achieve recognition performance of over 90% for detecting different types of waving gestures performed by a particular user.

6 Expression Generation

Finally, we can use the mechanisms described above as a tool for expression generation as well as recognition. Often a formidable problem in expression generation is determining the precise time-course of actuation parameters to create a realistic expression. Realistic face models have too many muscle control parameters to be able to hand-program animation sequences. However, vision-based analysis can help solve this problem, by allowing a real person to generate the precise expression timing using his/her own face. We have implemented this type of system, animating a physically based skin-and-muscle model of a human face using parameters in real-time computed from a real user's face using non-invasive visual analysis.

Our system operates by mapping the view-based distance scores for a person's face to a particular motor configuration in the generation model. Training examples of this mapping are given for a set of canonical expressions and define a multi-dimensional mapping using the Radial Basis Function technique, which is then used to drive the model face based on the view distance scores of newly acquired images. Figure 4 shows an example of our technique; for full details see [6].

We are currently applying this to the task of generating lip shapes during speech. By modeling a small number of lip shapes and acquiring view-models from a user for each shape, we can mimic and interpolate the users lip shape in real time. Thus when the user performs an utterance, we can construct a synthetic "talking-head" whose lip shape and motion is synchronous with the utterance. This system can either watch a user in a real-time interactive loop, which would be useful in teleconferencing applications, or be used to observe and store the time course of lip shape for given phonemes or phoneme pairs (to capture co-articulation effects), allowing for face generation from a purely synthetic utterance. Just as computer recognition of certain phonemes is aided by information from the vision sense, human recognition of spoken language will benefit by "seeing" the speech via an animated face.

7 Conclusion

A user's body pose and hand/face gestures are an important part of any human dialog. Here we have shown some first steps at detecting and interpreting these interface modalities. Two systems, the Alive system and the facial expression mimic, have been implemented and achieve real-time performance, and thus are useful examples of interactive vision-based systems. Combined with information from other interface channels, such as spoken language, the techniques exploited in these systems have the potential to greatly aid multi-modal interfaces.

[1] Currently these implemented as separate systems, but they are being merged, through the use of a foveated camera set-up.

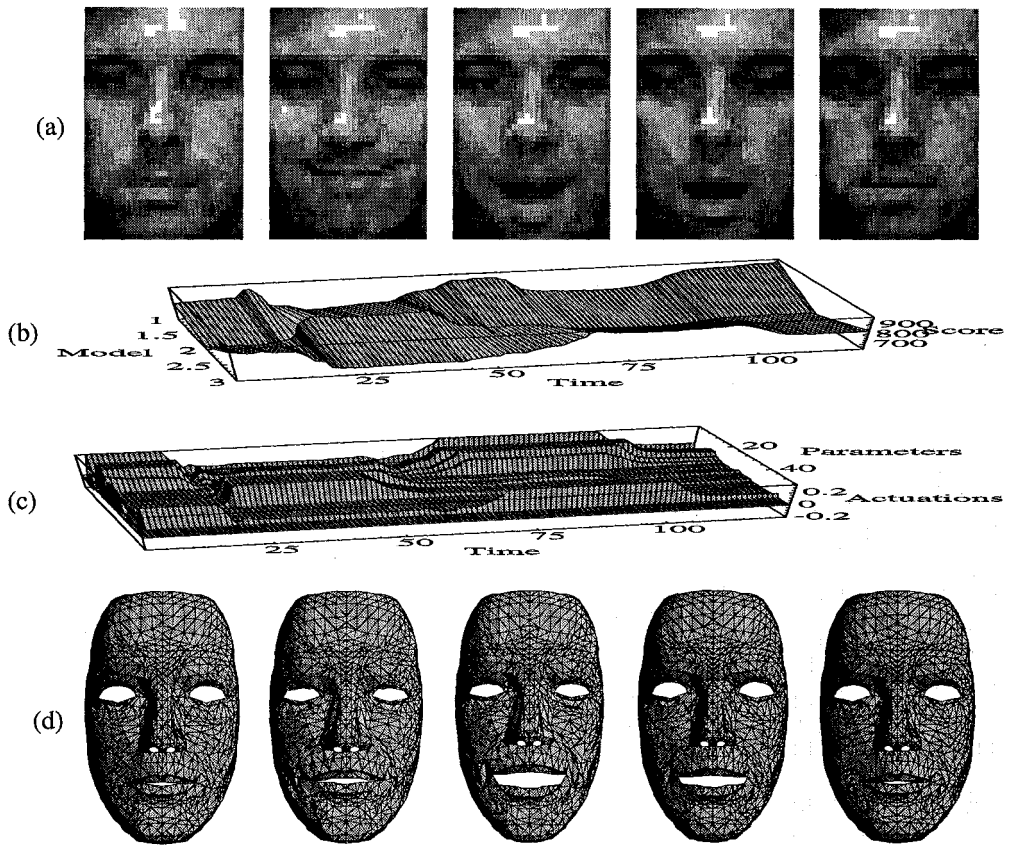

Figure 4: (a) Input sequence of face images (b) distance of face images to three canonical views (c) interpolated muscle control parameters (d) model faces animated with these muscle parameters.

References

[1] Bolt, R. (1984) "The Human Interface, where people and computers meet," *Lifetime Learning Publications*, Belmont, California.

[2] Ballard, D., and Brown, C., *Computer Vision*, Prentice-Hall, Englewood, 1982.

[3] Bichel, M. and Pentland, A., Topological Matching for Human Face Recognition, in *Proc. Looking at People Workshop*, IJCAI '93, Chamberry, France, August 1993.

[4] Darrell T., and Pentland A. P., Robust Estimation of a Multi-Layer Motion Representation, in *Proc. IEEE Motion Workshop*, Princeton, 1991.

[5] Darrell, T., and Pentland, A. P., Classification of Hand Gestures using a View-Based Distributed Representation in *Proc. Neural Information Processing Systems (NIPS)-6*, Denver 1993.

[6] Darrell, T., Essa, I., and Pentland, A. P., Correlation and Interpolation Networks for Real-time Expression Analysis/Synthesis, MIT Media Lab TR-284

[7] Darrell, T., Maes, P., Blumberg, B., and Pentland, A. P., A novel environment for situated vision and behavior, in *Proc. IEEE workship on Visual Behaviors*, Seattle, 1994.

[8] Horn, B.K.P.S., *Robot Vision*, M.I.T. Press, Cambridge, MA, 1991.

[9] Maes, P., Darrell, T., Blumberg, B., and Pentland, A. P., "The Alive system", SIGGRAPH 93 Visual Proceedings, Anahiem, 1993

[10] Mase, K., (1991) "Recognition of Facial Expression from Optical Flow", *ICICE Transactions*, Vol. E 74, No. 10, pp. 3474-3483

[11] Turk, M., and Pentland, A. P., "Eigenfaces for Recognition", *J. Cog. Neurosci.*