

Decision Theoretic Learning of Human Facial Displays

Jesse Hoey and James J. Little

Department of Computer Science, University of British Columbia
2366 Main Mall, Vancouver, BC, CANADA V6T 1Z4
{jhoey,little}@cs.ubc.ca

Changes in the human face occur due to many factors, including communication, emotion, speech, and physiology. Most systems for facial expression analysis attempt to *recognize* one or more of these factors, resulting in a machine whose inputs are video sequences or static images, and whose outputs are, for example, basic emotion categories. Our approach is fundamentally different. We make no prior commitment to some particular recognition task. Instead, we consider that the *meaning* of a facial display to an observer is contained in its relationship to actions and outcomes. Agents must distinguish facial displays according to their *affordances*, or how they help an agent to maximize utility. We show how an agent can learn relationships between observations of a person’s face, the context in which that person is acting, and its own actions and utility function. The agent can then base its decisions on these learned decision-theoretic models, allowing it to make value-directed action choices based, in part, upon observed facial displays.

Recent research on the communicative function of the face supports our approach. Psychologists have concluded that facial displays are purposeful communicative signals [1], that the purpose is dependent on both the display and the context of its emission [4], and that the signals vary widely between individuals [4]. These considerations imply that a rational communicative agent must learn the relationships between facial displays, the context in which they are shown, and its own utility function: it must be able to compute the utility of taking actions in situations involving purposeful facial displays. The agent will then be able to make value-directed decisions based, in part, upon the “meaning” of facial displays as contained in these learned connections between displays, context, and utility. Learning these relationships will further allow an agent to adapt to new interactants and new situations.

The model we propose is a partially observable Markov decision process, or POMDP, which realises the design constraints suggested by the psychology literature, combining the recognition of facial signals with their interpretation and use in a consistent utility-maximization framework. The parameters of the model are learned from training data using an *a posteriori* constrained optimization technique, such that an agent can learn to act based on the facial signals of a human through observation. The training is *unsupervised*: we do not train classifiers for individual facial displays, and then combine them in the model. Rather, the learning process *discovers* clusters of facial motions and their relationship to the context automatically. The advantage of this approach is threefold. First, we do not need expert knowledge about which facial motions are important. Second, since the system learns categories of motions, it will adapt to novel gestures or displays without modification. Third, resources can be focused on tasks that will be useful for the agent. It is wasteful to train complex classifiers for the recognition of fine facial motion if only simple displays are being used in the agent’s context.

A POMDP is a probabilistic temporal model of an agent interacting with the environment [3], shown as a Bayesian network in Figure 1. A POMDP describes how an agent’s actions, A , affect the state of the world, S , which may be only observable through its effects on observations, O .

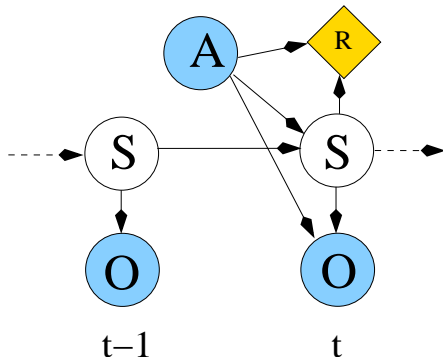


Figure 1: A POMDP.

A reward function, $R(S, A)$, defines the utility of performing actions in world states. The parameters of the model include a transition function, $T \equiv P(S_t|A_tS_{t-1})$ and an observation function, $B \equiv P(O_t|S_tA_t)$. The agent’s actions, A , are chosen from a finite set of discrete possibilities. The states, S , combine both observable descriptors of context, and unobservable high-level descriptors of facial display sequences, O . The observations, O , are sequences of video frames, and so the observation function, B , defines temporal and spatial abstractions of input video data, allowing decision making to occur over a small set of states which are summarizations of the continuous sensory signals. For example, one such state may correspond

with the wink of an eye, another may correspond to a smile, and decisions may be made based upon distributions over these high-level motion states. The models for building these abstractions are multi-level dynamic Bayesian networks, which couple dynamics and configuration information about the human face. The model is not restricted to learning facial displays: the same structure could learn any visually observable temporally variable feature, such as gestures. Figure 2 shows an example of our model explaining part of a video sequence in which a person smiles.

We trained the POMDP model on videos of a two-player cooperative card game. The game is structured such that communication is crucial to play, but the players can only see (not hear) their teammate through a real-time video link. There are no game rules concerning the video link, so there are no restrictions placed on communication strategies the players can use. The players naturally came up with simple head gestures and facial expressions to help them win the game. Our POMDP model was trained on a portion of the data, and an approximate policy of action was computed. When applied to the remaining portion of the data, the POMDP inferred the facial displays that the players were using, and correctly predicted the human player’s actions in the test data. An autonomous game playing agent using this model, therefore, would have played as well as the human player. Our current work involves finding redundant states in the learned POMDP, leading to value-directed structure learning of the observation function: the model learns which displays are useful to distinguish.

References

- [1] Alan J. Fridlund. *Human facial expression: an evolutionary view*. Academic Press, San Diego, CA, 1994.
- [2] Jesse Hoey and James J. Little. Bayesian clustering of optical flow fields. In *Proc. International Conference on Computer Vision (ICCV)*, Nice, France, October 2003.
- [3] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.
- [4] James A. Russell and Jose Miguel Fernández-Dols, editors. *The Psychology of Facial Expression*. Cambridge University Press, Cambridge, UK, 1997.

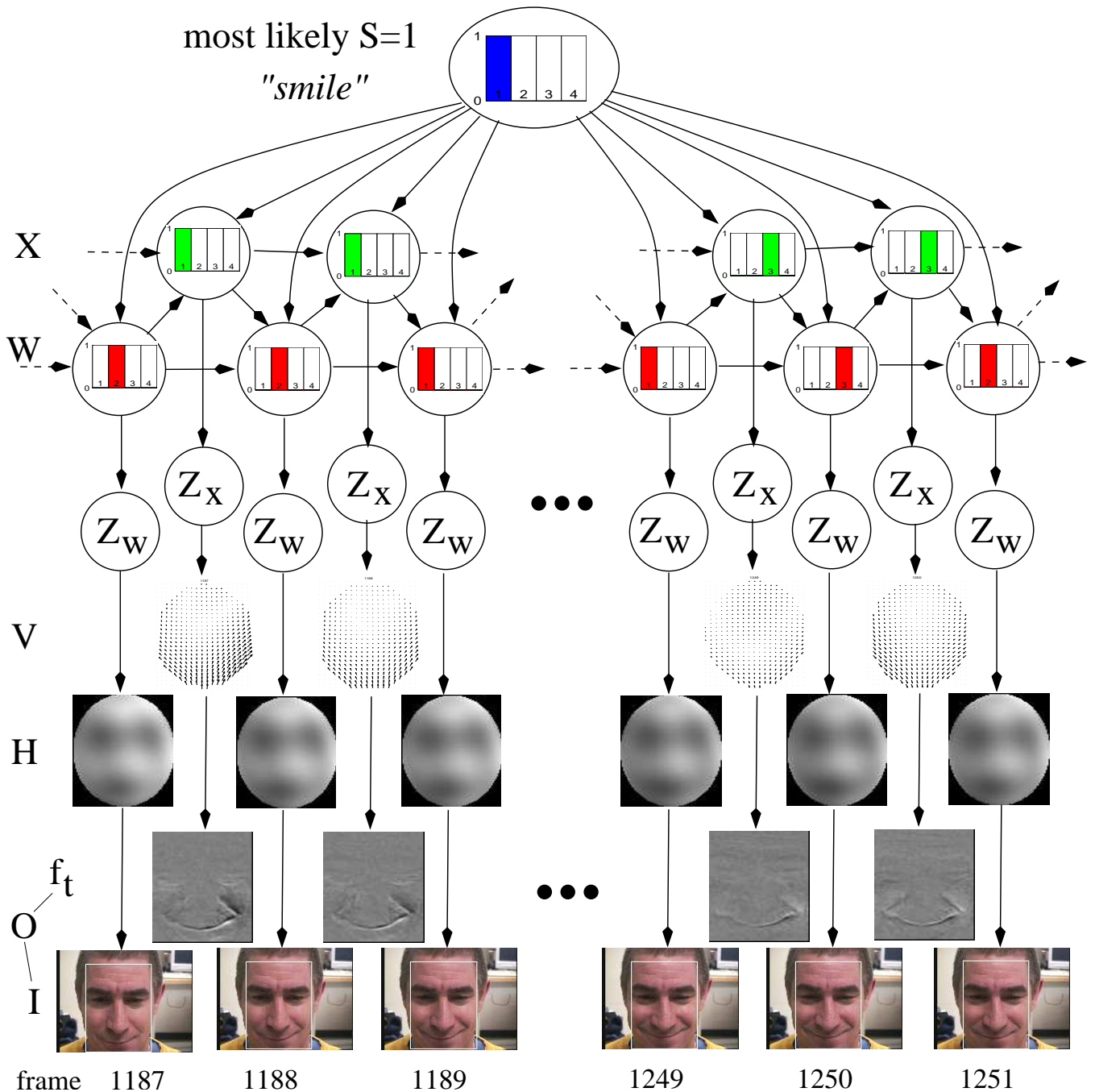


Figure 2: Observation function for the POMDP. Observations, O , are sequences of images, I , and image temporal derivatives, f_t , both of which are projected over the facial region to a set of basis functions, yielding feature vectors, Z_x and Z_w . The image regions, H , are projected directly, while it is actually the optical flow fields, V , related to the image derivatives which are projected to the basis functions [2]. Z_x and Z_w are both modeled using mixtures of Gaussians, X and W , respectively. The class distributions, X and W , are temporally modeled as mixture, S , of coupled Markov chains. The probability distribution over S is at the top. The most likely state, $S = 1$, can be associated with the concept “smile”. Probability distributions over X and W are shown for each time step. All other nodes in the network show their expected value given all evidence. Thus, the flow field, v , is actually $\langle v \rangle = \int_v v P(v|O)$.