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

Abstract

We present a vision-based, adaptive, decision-theoretic model of human facial displays. 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 learn the *meaning* of a facial display by learning its relationship to actions, outcomes and utilities. The model is a partially observable Markov decision process, or POMDP, the parameters of which are learned from training data using an *a-posteriori* constrained optimization technique based on the expectation-maximization algorithm. One of the most significant advantages of this type of learning is that it does not require labeled data from expert knowledge about which behaviors are significant in a particular interaction. Rather, the learning process *discovers* clusters of facial motions and their relationship to the context automatically. We present an results from an experiment in which we record two humans playing a collaborative game, and learn their behaviors. We use the resulting model to predict human actions.

1 Introduction

Recent research on the communicative function of the face has concluded that facial displays are often purposeful communicative signals [7], that the purpose is dependent on both the display and the context of its emission [19], and that the signals vary widely between individuals [19]. 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 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. Video observations are integrated into the POMDP using a dynamic Bayesian network, which creates spatial and temporal abstractions amenable to decision making at the high level. 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. One of the most significant advantages of this type of learning is that it does not require labeled data from expert knowledge about which behaviors are significant in a particular interaction. Rather, the learning process *discovers* clusters of facial motions and their relationship to the context automatically. As such, it can be applied to any situation in which non-verbal gestures are purposefully used in a task. 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.

2 Prior Work

There are many examples of work in computer vision analysing facial displays [20], and human motion in general [3, 18]. However, this work is usually supervised, in that models of particular classes of human motion are learned from labeled training data. There has been some recent research in unsupervised learning of motion models [1, 9], but few have attempted to explicitly include the modeling of actions and utility, and none have looked at facial displays. Action-Reaction Learning [15] is a system for analysing and synthesising human behaviours. It is primarily reactive, however, and does not learn models conducive for high level reasoning about the long term effects of actions.

Our previous work on this topic has led to the development of many parts of the system described in this paper. In particular, the low-level computer vision system for instantaneous action recognition was described in [12], while the simultaneous learning of the high-level parameters was explored in [10]. This paper combines this previous work, explicitly incorporates actions and utilities, and demonstrates how the model is a POMDP, from which policies of action can be extracted. Complete details can be found in [11].

POMDPs have become the semantic model of choice for decision theoretic planning in the artificial intelligence (AI) community. While solving POMDPs optimally is intractable for most real-world problems, the use of approximation methods have recently enabled their application to substantial planning problems involving uncertainty, for example, card games [8] and robot control [17]. POMDPs were applied to the problem of active gesture recognition in [5], in which the goal is to model unobservable, non-foveated regions. This work models some of the basic mechanics underlying dialogue, such as turn taking, channel control, and signal detection. Work creating embodied agents has led to much progress in creating agents that interact using verbal and non-verbal communication [4]. These agents typically only use a small subset of manually specified facial expressions or gestures. They focus instead primarily on dialogue management and multi-modal inputs, and have not used POMDPs.

3 POMDPs for non-verbal display understanding

A POMDP is a probabilistic temporal model of an agent interacting with the environment [16], shown as a Bayesian network in Figure 1(a). A POMDP is similar to a hidden

Markov model in that it describes observations as arising from hidden states, which are linked through a Markovian chain. However, the POMDP adds actions and rewards, allowing for decision theoretic planning. A POMDP is a tuple $\langle S, A, T, R, O, B \rangle$, where S is a finite set of (possible unobservable) states of the environment, A is a finite set of agent actions, $T : S \times A \rightarrow S$ is a transition function which describes the effects of agent actions upon the world states, O is a set of observations, $B : S \times A \rightarrow O$ is an observation function which gives the probability of observations in each state-action pair, and $R : S \rightarrow \mathcal{R}$ is a real-valued reward function, associating with each state s its immediate utility $R(S)$. A POMDP model allows an agent to predict the long term effects of its actions upon his environment, and to choose actions based on these predictions. Factored POMDPs [14] represent the state, S, using a set of variables, such that the state space is the product of the spaces of each variable. Factored POMDPs allow conditional independencies in the transition function, T, to be leveraged. Further, T is written as a set of smaller, more intuitive functions.

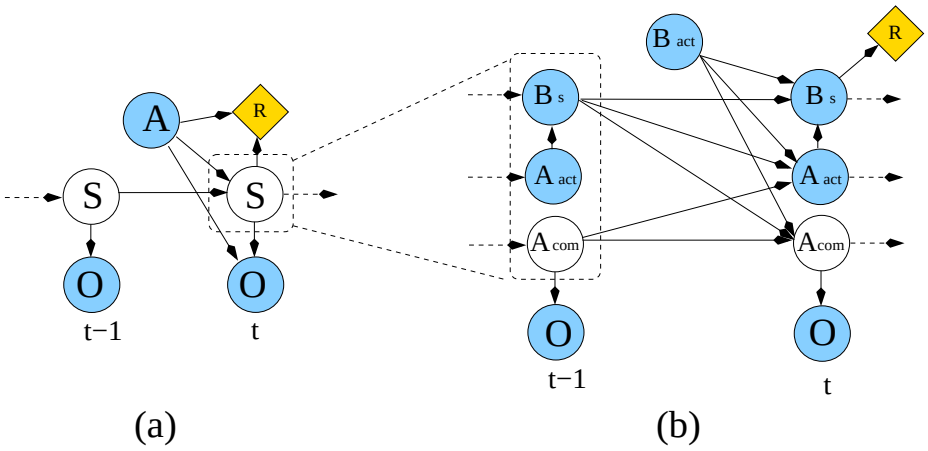

Figure 1: (a) Two time slices of general POMDP. (b) Two time slices of factored POMDP for facial display understanding. The state, S, has been factored into $\{Bs, Aact, Acom\}$, and conditional independencies have been introduced: Ann's actions do not depend on her previous actions and Ann's display is independent of her previous action given the state and her previous display. These independencies are not strictly necessary, but simplify our discussion, and are applicable in the simple game we analyse.

Purposeful facial display understanding implies a multi-agent setting, such that each agent will need to model all other agent's decision strategies as part of its internal state [1]. In the following, we will refer to the two agents we are modeling as "Bob" and "Ann", and we will discuss the model from Bob's perspective. Figure 1(b) shows a factored POMDP model for facial display understanding in simple interactions. The state of Bob's POMDP is factored into Bob's private internal state, Bs, Ann's action, $Aact$, and Ann's facial display, $Acom$, such that $S_t = \{Bs_t, Aact_t, Acom_t\}$. While Bs and $Aact$ are observable, $Acom$ is not, and must be inferred from video sequence observations, $\mathbf{O}$. In general, both $Aact$ and Bs may also be unobservable. However, we wish to focus on learning models of facial displays, $Acom$, and so we will use games in which $Aact$ and Bs are fully observable. For example, in a real game of cards, a player must model the suit of any played card as an

[1] This is known as the *decision analytic* approach to games, in which each agent decides upon a strategy based upon his subjective probability distribution over the strategies employed by other players.

unobservable variable, which must be inferred from observations of the card. In our case, games will be played through a computer interface, and so these kinds of actions are fully observable.

The transition function is factored into four terms. The first involves only fully observable variables, and is the conditional probability of the state at time t under the effect of both player's actions: $\Theta_S = P(Bs_t|Aact_t, Bact, Bs_{t-1})$. The second is over Ann's actions given Bob's action, the previous state, and her previous display: $\Theta_A = P(Aact_t|Bact, Acom_{t-1}, Bs_{t-1})$. The third describes Bob's expectation about Ann's displays given his action, the previous state and her previous display: $\Theta_D = P(Acom_t|Bact, Bs_{t-1}, Acom_{t-1})$. The fourth describes what Bob expects to see in the video of Ann's face, $\mathbf{O}$, given his high-level descriptor, $Acom$: $\Theta_O = P(\mathbf{O}_t|Acom_t)$. For example, for some state of $Acom$, this function may assign high likelihood to sequences in which Ann smiles. This value of $Acom$ is only assigned meaning through its relationship with the context and Bob's action and utility function. We can, however, look at this observation function, and interpret it as an $Acom$ = 'smile' state. Writing $C_t = \{Bact_t, Bs_{t-1}\}$, $A_t = Aact_t$, and $D_t = Acom_t$, the likelihood of a sequence of data, $\{\underset{1,T}{\mathbf{OCA}}\} = \{O_1 \dots O_T, C_1 \dots C_T, A_1 \dots A_T\}$, is

$$P(\{\underset{1,T}{\mathbf{OCA}}\}|\Theta) = \sum_k P(\mathbf{O}_T|D_{T,k}) \sum_l \Theta_A \Theta_D P(D_{T-1,l}, \{\underset{1,T-1}{\mathbf{OCA}}\}|\Theta) \quad (1)$$

where $D_{t,k}$ is the k^{th} value of the mixture state, D, at time t. The observations, $\mathbf{O}$, are temporal sequences of finite extent. We assume that the boundaries of these temporal sequences will be given by the changes in the fully observable context state, C and A. There are many approaches to this problem, ranging from the complete Bayesian solution in which the temporal segmentation is parametrised and integrated out, to specification of a fixed segmentation time [18].

3.1 Observations

We now must compute $P(\mathbf{O}|Acom)$, where $\mathbf{O}$ is a sequence of video frames. We have developed a method for generating temporally and spatially abstract descriptions of sequences of facial displays from video [12, 13]. We give a brief outline of the method here. Figure 2 shows the model as a Bayesian network being used to assess a sequence in which a person smiles.

We consider that spatially abstracting a video frame during a human facial display involves modeling both the current configuration and dynamics of the face. Our observations consist of the video images, I, and the temporal derivatives, f_t, between pairs of images. The task is first to spatially summarise both of these quantities, and then to temporally compress the entire sequence to a distribution over high level descriptors, $Acom$. We assume that the face region is tracked through the sequence by a separate tracking process, such that the observations arise from the facial region in the images only. We use a flow-based tracker, described in more detail in [13].

The spatial abstraction of the derivative fields involves a projection of the associated optical flow field, v, over the facial region to a set of pre-determined basis functions. The basis functions are a complete and orthogonal set of 2D polynomials which are effective for describing flow fields [13]. The resulting feature vector, Z_x, is then conditioned on a set of discrete states, X, parametrised by normal distributions. The projection is accomplished by analytically integrating the observation likelihood, $P(f_t|X)$, over the space of optical flow fields and over the feature vector space. This method ensures that all observation noise is consistently propagated [12]. The abstraction of the images also uses projections of the raw (grayscale) images to the same set of basis functions, resulting in a feature vector, Z_w, which is also modeled using a mixture of normal distributions with mixture coefficients W.

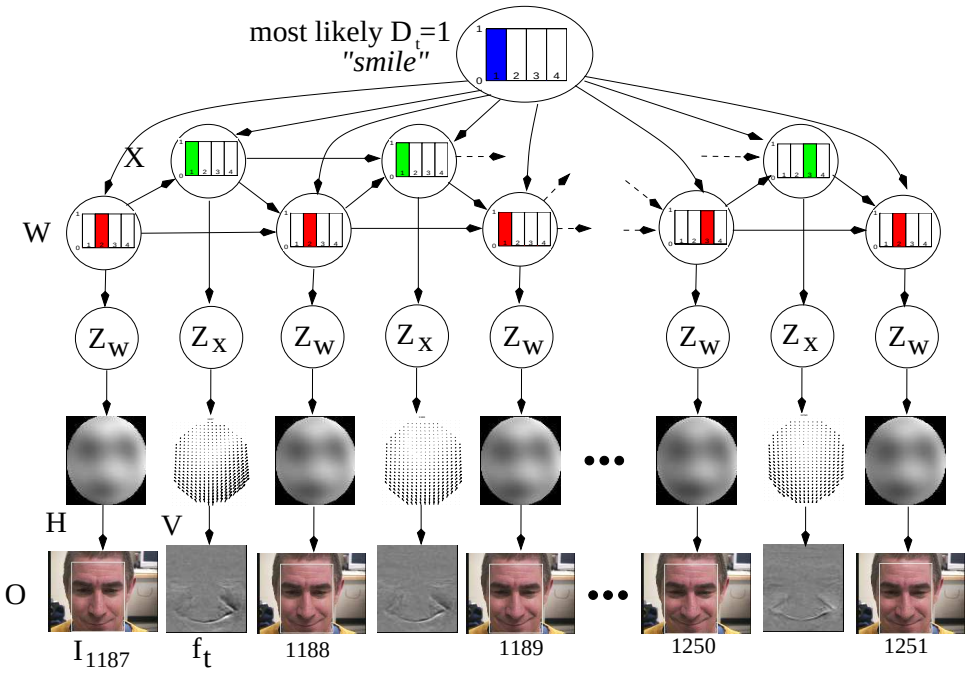

Figure 2: A person smiling is analysed by the mixture of CHMMs. Observations, O, are sequences of images, I, and image temporal derivatives, f_t, both of which are projected over the facial region to a set of basis functions, yielding feature vectors, Z_x and Z_w. The image regions, H, are projected directly, while it is actually the optical flow fields, V, related to the image derivatives which are projected to the basis functions [12]. 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, D, of coupled Markov chains. Probability distributions over X, W and D are shown for each time step as bar charts. H and V nodes show their expected value, e.g. v is actually $\langle v \rangle = \int_v v P(v|O)$.

The basis functions are a complete and orthogonal set, but only a small number may be necessary for modeling any particular motion. We use a feature weighting technique that places priors on the normal means and covariances, so that choosing a set of basis functions is handled automatically by the model [12].

At each time frame, we have a discrete dynamics state, X, and a discrete configuration state, W, which are abstract descriptions of the instantaneous dynamics and configuration of the face, respectively. These are temporally abstracted using a mixture of coupled hidden Markov models (CHMM), in which the dynamics and configuration states are interacting Markovian processes. The conditional dependencies between the X and W chains are chosen to reflect the relationship between the dynamics and configuration. This mixture model can be used to compute the likelihood of a video sequence given the facial display descriptor, $P(\mathbf{O}|Acom)$:

$$P(\underset{1,T}{\{\mathbf{O}\}}|D_T) = \sum_{ij} P(f_t|X_{T,i}) P(I_t|W_{T,j}) \sum_{kl} \Theta_{Xijk} \Theta_{Wjkl} P(X_{T-1,k}, W_{T-1,l}\, \underset{1,T-1}{\{\mathbf{O}\}}|D_T) \tag{2}$$

Where Θ_X, Θ_W are the transition matrices in the coupled X and W chains, and $P(f_t|X_{T,i}), P(I_t|W_{T,j})$ are the associated observation functions [13]. The mixture components, D, are a set of discrete abstractions of facial behavior.

3.2 Learning POMDPs

We use the expectation-maximization (EM) algorithm [6] to learn the parameters of the POMDP. It is important to stress that the learning takes place over the *entire* model simultaneously: both the output distributions, including the mixtures of coupled HMMs, and the high-level POMDP transition functions are all learned from data during the process. The learning classifies the input video sequences into a spatially and temporally abstract finite set, $Acom$, and learns the relationship between these high-level descriptors, the observable context, and the action. We only present some salient results of the derivation here. We seek the set of parameters, Θ^*, which maximize

$$\Theta^* = \arg\max_{\Theta} \left[\sum_{\mathbf{D}} P(\mathbf{D}|\mathbf{O}, \mathbf{C}, \mathbf{A}, \theta') \log P(\mathbf{D}, \mathbf{O}, \mathbf{C}, \mathbf{A}|\Theta) + \log P(\Theta) \right] \quad (3)$$

subject to constraints on the parameters, Θ^*, that they describe probability distributions (they sum to 1). The "E" step of the EM algorithm is to compute the expectation over the hidden state, $P(\mathbf{D}|\mathbf{O}, \mathbf{C}, \mathbf{A}, \theta')$, given θ', a current guess of the parameter values. The "M" step is then to perform the maximization which, in this case, can be computed analytically by taking derivatives with respect to each parameter, setting to zero and solving for the parameter.

The update equation for the D transition parameter, $\Theta_{Dijk} = P(D_{t,i}|D_{t-1,j}C_{t,k})$, is then

$$\Theta_{Dijk} = \frac{\alpha_{Dijk} + \sum_{t \in \{1 \ldots N_t\} | C_t = k} P(D_{t,i} D_{t-1,j} | \mathbf{O}, \mathbf{A}, \mathbf{C}\theta')}{\sum_i \left[\alpha_{Dijk} + \sum_{t \in \{1 \ldots N_t\} | C_t = k} P(D_{t,i} D_{t-1,j} | \mathbf{O}, \mathbf{A}, \mathbf{C}\theta') \right]}$$

where the sum over the temporal sequence is only over time steps in which $C_t = k$, and α_{Dijk} is the parameter of the Dirichlet smoothing prior. The summand can be factored as

$$P(D_{t,i} D_{t-1,j} | \mathbf{O}, \mathbf{A}, \mathbf{C}\theta') = \beta_{t,i} \Theta_{A*i*} P(\mathbf{O}_t | D_{t,i}) \Theta_{Dijk} \alpha_{t-1,j}$$

where $\alpha_{t,j} = P(D_{t,j}\{\underset{1,t}{\mathbf{OAC}}\})$ and $\beta_{t,i} = P(\{\underset{t+1,T}{\mathbf{OAC}}\}|D_{t,i})$ are the usual forwards and backwards variables, for which we can derive recursive updates

$$\alpha_{t,j} = \sum_k P(\mathbf{O}_t|D_{t,j}) \Theta_{A*j*} \Theta_{Djk*} \alpha_{t-1,k} \qquad \beta_{t-1,i} = \sum_k \beta_{t,k} \Theta_{A*k*} P(\mathbf{O}_t|D_{t,k}) \Theta_{Dki*}$$

where we write $\Theta_{A*j*} = P(A_t = *|D_{t,j}C_t = *)$ and $P(\mathbf{O}_t|D_{t,i})$ is the likelihood of the data given a state of the mixture of CHMMs (Equation 2). The updates to $\Theta_{Aijk} = P(A_{t,i}|D_{t,j}C_{t,k})$ are $\Theta_{Aijk} = \sum_{t \in \{1 \ldots N_t\} | A_t = i \vee C_t = k} \xi_j$, where $\xi_j = P(D_{t,j}|\mathbf{OAC}) = \beta_{t,j}\alpha_{t,j}$. The updates to the j^{th} component of the mixture of CHMMs is weighted by ξ_j, but otherwise is the same as for a normal CHMM [2]. The complete derivation, along with the updates to the output distributions of the CHMMs, including to the feature weights, can be found in [13].

3.3 Solving POMDPs

If observations are drawn from a finite set, then an optimal policy of action can be computed for a POMDP [16] using dynamic programming over the space of the agent's belief about the state, $b(s)$. However, if the observation space is continuous, as in our case, the problem becomes much more difficult. In fact, there are no known algorithms for computing optimal policies for such problems. Nevertheless, approximation techniques have been developed, and yield satisfactory results [17]. Since our focus in this paper is to learn POMDP models, we use the simplest possible approximation technique, and simply consider the POMDP as a fully observable MDP: the state, S, is assigned its most likely value in the belief

state, $S = \arg\max_s b(s)$. Dynamic programming updates then consist of computing value functions, V^n, where $V^n(s)$ gives the expected value of being in state s with a future of n stages to go, assuming the optimal actions are taken at each step. The actions that maximize V^n are the policy with n stages to go. We use the SPUDD solver to compute these policies [14].

4 Experimental Results

We trained the POMDP model on videos of two humans playing a cooperative card game. In each round of the game, players attempt to play matching cards after an initial phase in which they can communicate with each other through a real-time video link (with no audio). 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 to help them win the game: nodding and shaking. The facial regions of the players were tracked in the video using an optical flow based tracker, with corrections from an exemplar database [13].

The data was split into training and test sets, and our POMDP model with $N_a = 4$ display states was learned with the training set. The learning discovered appropriate motion sequence models for each of the head gestures the players were using. Two of the learned display states (d_1 and d_2) described neutral displays with little or no motion, while one described head nods, and the other head shakes. An approximate two-stage policy of action was computed using the MDP approximation. The value function, $V(s)$, assigns nearly equal values to the null displays (d_1 and d_2). This indicates that making the distinction between these two behaviors is not useful for determining value, and we can merge them, resulting in a three-state model [13].

The computed policy was consulted, and the recommended actions were compared to Bob's actual actions taken in the game. The model correctly predicted 6/7 actions in the testing data, and 19/20 in the training data. The error in the testing data was due to the subject glancing at something to the side of the screen, leading to a classification as d_4. This error demonstrates the need for dialogue management, such as monitoring of the subject's attention [17].

Figure 3 shows example frames from a sequence in which the subject shook her head. The entire sequence was classified as facial display state d_3 by the final merged model with three states. Figure 4 shows example frames from a sequence in which the subject nodded her head, classified as facial display state d_2 by the final merged model.

5 Conclusions

We have presented a method for learning decision theoretic models of purposeful human non-verbal displays using partially observable Markov decision processes. It discovers spatially and temporally abstract categories of motion sequences and their relationship with actions, utilities and context automatically from video. No prior knowledge about the types of displays expected in an interaction is needed to train the model. The learned values of states are used to discover the number of display classes which are important for achieving value in the context of the interaction. This type of value-directed structure learning allows an agent to only focus resources on necessary distinctions.

Acknowledgements: Supported by the Institute for Robotics and Intelligent Systems (IRIS), a Canadian Network of Centers of Excellence, and by a Precarn scholarship. The authors thanks Nicole Arksey, Don Murray, and Pascal Poupart.

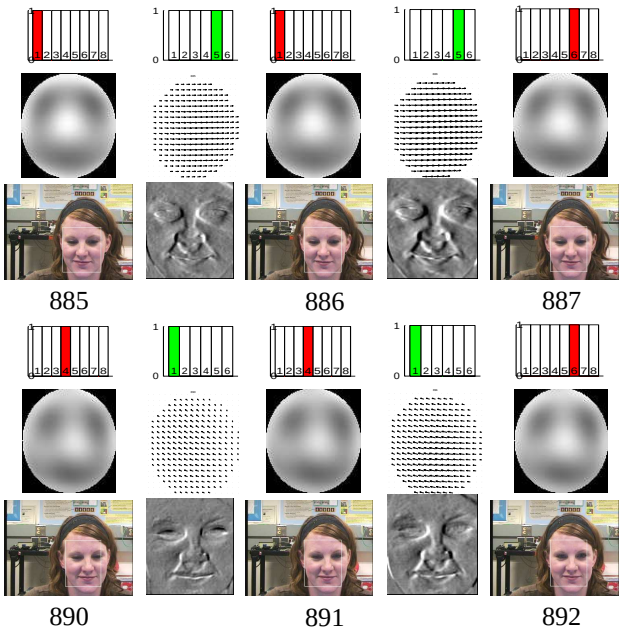

Figure 3: Part of a sequence of subject shaking her head, classified as model d_3.

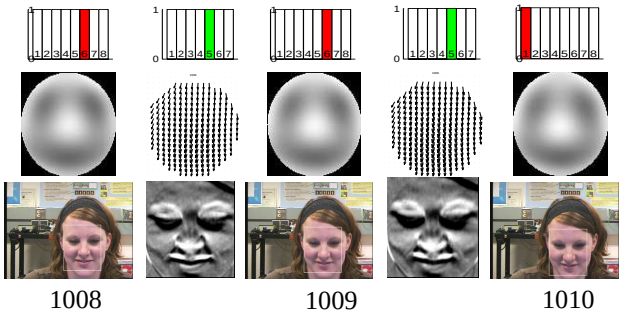

Figure 4: Part of a sequence of subject nodding, classified as model d_2.

References

[1] Matthew Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11:1155–1182, 1999.

[2] Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden Markov models for complex action recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 994–999, Puerto Rico, 1997.

[3] Chris Bregler. Learning and recognising human dynamics in video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 568–574, 1997.

[4] Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors. *Embodied Conversational Agents*. MIT Press, 2000.

[5] Trevor Darrell and Alex P. Pentland. Active gesture recognition using partially observable Markov decision processes. In *13th IEEE Intl. Conference on Pattern Recognition*, Vienna, Austria, 1996.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data using the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[7] Alan J. Fridlund. *Human facial expression: an evolutionary view*. Academic Press, San Diego, CA, 1994.

[8] Hajime Fujita, Yoichiro Matsuno, and Shin Ishii. A reinforcement learning scheme for a multi-agent card game. *IEEE Trans. Syst., Man. & Cybern*, pages 4071–4078, 2003.

[9] Aphrodite Galata, Anthony G. Cohn, Derek Magee, and David Hogg. Modeling interaction using learnt qualitative spatio-temporal relations and variable length Markov models. In *Proc. European Conference on Artificial Intelligence*, July 2002.

[10] Jesse Hoey. Clustering contextual facial display sequences. In *Proceedings of IEEE International Conference on Face and Gesture*, Washington, DC, May 2002.

[11] Jesse Hoey. *Decision Theoretic Learning of Human Facial Displays and Gestures*. PhD thesis, University of British Columbia, 2004.

[12] Jesse Hoey and James J. Little. Bayesian clustering of optical flow fields. In *Proc. International Conference on Computer Vision*, pages 1086–1093, Nice, France, October 2003.

[13] Jesse Hoey and James J. Little. Decision theoretic modeling of human facial displays. Technical Report TR-04-02, Department of Computer Science, 2004.

[14] Jesse Hoey, Robert St-Aubin, Alan Hu, and Craig Boutilier. SPUDD: Stochastic planning using decision diagrams. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 279–288, Stockholm, 1999.

[15] Tony Jebara and Alex P. Pentland. Action reaction learning: Analysis and synthesis of human behaviour. In *IEEE Workshop on The Interpretation of Visual Motion*, 1998.

[16] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[17] Michael Montemerlo, Joelle Pineau, Nicholas Roy, Sebastian Thrun, and Vandi Verma. Experiences with a mobile robotic guide for the elderly. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, Edmonton, Canada, 2002. AAAI.

[18] Nuria Oliver, Eric Horvitz, and Ashutosh Garg. Layered representations for human activity recognition. In *Proceedings of International Conference on Multimodal Interfaces*, Pittsburgh, PA, October 2002.

[19] James A. Russell and Jose Miguel Fernández-Dols, editors. *The Psychology of Facial Expression*. Cambridge University Press, Cambridge, UK, 1997.

[20] Yingli Tian, Takeo Kanade, and Jeffrey F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), February 2001.