Clustering Contextual Facial Display Sequences

Jesse Hoey
Department of Computer Science
University of British Columbia
Vancouver, Canada, V6T 1Z4

Abstract

We describe a method for learning classes of facial motion patterns from video of a human interacting with a computerized embodied agent. The method also learns correlations between the uncovered motion classes and the current interaction context. Our work is motivated by two hypotheses. First, a computer user's facial displays will be context dependent, especially in the presence of an embodied agent. Second, each interactant will use their face in different ways, for different purposes. Our method describes facial motion using optical flow over the entire face, projected to the complete orthogonal basis of Zernike polynomials. A context-dependent mixture of hidden Markov models (cmHMM) clusters the resulting temporal sequences of feature vectors into facial display classes. We apply the clustering technique to sequences of continuous video, in which a single face is tracked and spatially segmented. We discuss the classes of patterns uncovered for a number of subjects.

1 Introduction

Recently, the notion that the primary function of facial displays is the expression of emotion has been challenged by psychologists, who have proposed a model of human facial displays as signals of social intent (the *Behavioral Ecology View*) [7]. They take the position that humans use their face as a way of communicating *through the medium of a social context*. For example, the reasons facial displays are used in normal conversations include both semantic and syntactic support of what the speaker is saying, as well as reactions in the listeners face to offer support of continuation of the dialog [4]. Human computer interaction researchers have also moved towards the ecological view, and have focused on interpreting the human face as a signaling mechanism [3]. That the same conclusions apply to both groups of researchers is not surprising given the media equation [13]: the principles which apply to inter-human communication should apply to human-computer communication.

Analyzing facial displays as communicative mechanisms has important ramifications for the design of facial display recognition systems. Facial displays are dependent on the momentary context in which the display is shown [7]. Context is defined as the circumstances relevant to the display, and may include concurrent or proximate speech, gestures, or environmental factors. For example, eyebrows are sometimes raised during conversation when the speaker is thinking or remembering. However, eyebrows are also raised in backchannel displays of acknowledgment, or when an individual is taking turn in a conversation [3]. In any case, the observed facial display carries little meaning by itself, while the combination of the current context and the observed facial display is meaningful.

The particular configurations and motions observed in facial displays are individual dependent [4]. This individuality of displays is governed in part by cultural, educational, situational and physical factors [16]. Furthermore, although people may use similar facial displays, they use them in dissimilar situations. This implies that pre-defined, user and context independent models of facial displays will not be sufficient. A facial display recognition system must *adapt* to the ways in which a particular human is using their face.

This paper presents a method for adaptive and context dependent recognition of human facial displays. Our system uses a color video camera to track the face of a human interacting with an on-screen embodied agent. The inclusion of an embodied face is important, in order to open a facial display communication channel (to give the human a reason to use their face). The motion in the human's face is described using optical flow over the entire face, projected to the complete orthogonal basis of Zernike polynomials. The trajectories of the resulting feature vector are modeled using a context dependent mixture of hidden Markov models (cmHMM). The cmHMM is a mixture of hidden Markov models [14] augmented with an (observed) context which conditions the mixture model. The cmHMM is trained on a set of unlabeled sequences, and extracts models of salient display patterns and of correlations with the current actions of the embodied agent (the context). This unsupervised clustering method allows for user adaptation and for integration of context. In this work, we examine what motion patterns can be extracted without *any* prior information about facial displays. However our Bayesian approach can be incorporated with prior knowledge of cross-individual statistics.

Most human-computer interaction systems react to a small, pre-defined set of user hand gestures, head motions, and facial displays [3]. Our work differs in that we make no assumptions about the facial displays a particular human will be using. While researchers have examined unsuper-

vised clustering of human gesture sequences [17, 15], to our knowledge no work has focused on facial displays. Wren *et. al* [17] investigate understanding purposeful human gestures using a combination of a kinematic model of human motion, a mixture of hidden Markov models and a high-level classifier, similar to our approach.

Section 2 shows how we obtain feature vectors representative of motion in the face. Section 3 presents the mixture of hidden Markov models as a clustering technique for temporal patterns of feature vectors, and shows how to incorporate context. Section 4 presents results from a number of different subjects engaged in a facial display imitation task.

2 Facial motion representation

Extracting a meaningful and simple feature vector representing the human face and the motion thereof can be approached in many ways. In this work, we focus only on the motion of the face, and use a holistic representation over the entire facial region. We estimate optical flow between successive images, and project this flow field over a tracked facial region to the complete orthogonal basis of Zernike polynomials, yielding a feature vector, $\mathbf{Z} = Z_0 ... Z_T$. The track is updated using both the recovered optical flow and an independent estimate of the face region using skin color segmentation.

The Zernike representation differs from approaches such as Eigen-analysis [11], or facial action unit recognition [10] in that it makes no commitment to a particular type of motion [9]. This is useful for adaptive recognition, and leads to a transportable classification system (e.g. usable for gesture clustering). Although the the recognition of facial action units [10] gives the ability to discriminate between very subtle differences in facial motion, it requires extensive training and domain specific knowledge. We prefer to use a representation which can be extended from simple to complex given the task to be accomplished. For example, a system which only needs to recognize nods of the head could use only the first order ZPs. More complex recognition tasks need only add as much representation as is necessary to distinguish the important facial displays of a particular user.

We estimate optical flow between a successive pair of frames in a video sequence using the robust gradient-based regularization method of [1]. This method yields estimates of flow which are smooth over patches in the human face, preserving important discontinuities but removing high spatial frequency noise arising from violations of the brightness constancy assumption. After the flow is computed, the centroid and scale of the area of interest are estimated (see below) and a feature vector is obtained by projecting the flow onto the basis of Zernike polynomials.

Zernike polynomials are an orthogonal set of complex polynomials defined on the unit disk [12]. The lowest two orders of Zernike polynomials correspond to the standard affine basis. The next order polynomials correspond to extensions of the affine basis, roughly *yaw*, *pitch* and *roll*, as explored in [2]. Higher orders represent motions with higher spatial frequencies. The basis is orthogonal over the unit disk, such that each order can be used as an independent characterization of the flow, and each flow field has a unique decomposition in the basis. Zernike polynomials are expressed in polar coordinates as a radial function, $R_n^m(\rho)$, modulated by a complex exponential in the angle, ϕ:

$$U_n^m(\rho, \phi) = R_n^m(\rho) e^{im\phi} \tag{1}$$

The orthogonality of the basis allows the decomposition of an arbitrary function on the unit disk, $F(\rho, \phi)$, in terms of a unique combination of Zernike polynomials [12]:

$$F(\rho, \phi) \approx \sum_{m=0}^{M} \sum_{n=m}^{N} \left[A_n^m cos(m\phi) + B_n^m sin(m\phi) \right] R_n^m(\rho), \tag{2}$$

The coefficients, A_n^m and B_n^m, of the decomposition of the horizontal and vertical flow estimates, $u(x, y)$ and $v(x, y)$, over the tracked facial region are thus obtained using:

$$\begin{matrix} {}^{u}A_n^m \\ {}^{u}B_n^m \end{matrix} = \frac{\epsilon_m (n+1)}{\pi} \sum_x \sum_y u(x, y) R_n^m(\rho) \begin{matrix} cos(m\phi) \\ sin(m\phi) \end{matrix} \tag{3}$$

where $\phi = \arctan(y'/x')$, $\rho = \sqrt{x'^2 + y'^2} \leq 1$, $x' = (x - x_c)/r_x$, $y' = (y - y_c)/r_y$, and $\{x_c, y_c\}$ and $\{r_x, r_y\}$ are the centroid and scales of the region of interest. The flows can be reconstructed from the coefficients using Equation 2. Feature vectors are sets of the coefficients from Equation 3. The choice of a particular set to represent the flow will depend on the types of flows being modeled [9]. This choice is currently made by the modeler, by removing the $n = 0$ component (translation) and then adding as many orders as can be supported by the data in the modeling process.

The tracking problem is to update the facial region as described by centroid and scale parameters $c = \{x_c, y_c, r_x, r_y\}$ from one frame to the next. We assume that there is only one head present in all frames. We get a first estimate from the first and second order coefficients of the projected flow:

$$x_c' = x_c + {}^{u}A_0^0 \quad y_c' = y_c + {}^{v}A_0^0$$

$$r_x' = r_x + {}^{u}A_1^1 \quad r_y' = r_y + {}^{v}B_1^1$$

Updates using the only the flow are prone to significant drift over any sequence longer than roughly 600 frames (20 seconds). Furthermore, severe permanent tracker failure can be caused by *adaptors*, such as scratching the face. Therefore, we derive a correction term using skin color segmentation. We transform the RGB color images to HSV space, and segment using simple thresholding in hue and saturation. Median filtering removes noisy estimates, and the resulting binary image is projected along horizontal and vertical directions. The region of interest is then estimated by examining where the projected distributions fall below a threshold. The centroid and scale are then updated using a weighted sum of the skin and flow estimates. The scale and centroid

for the initial frame of a sequence is given by the skin segmentation procedure alone. The top row in Figure 1 shows an example track using only updates based on optical flow. The tracker performs well until errors are introduced by the tracked individual's hand, from which the tracker cannot recover. Correction using the skin segmentation (shown in the middle row in Figure 1) yields the track shown in the bottom row in Figure 1.

T=1200 1304 1338 1757

Figure 1: Tracking using only flow (top row). Skin segmentation (middle row) yield corrected track (bottom row).

3 Clustering motion patterns

Once we have recovered a sequence of feature vectors $\mathbf{Z} = Z_0 ... Z_T$ we wish to find and classify the trajectories corresponding to salient facial displays. Simultaneously, we want to model the dependence on the current context. In this work, we will examine facial displays only as *reactions* to context. A person is seated in front of a screen, monitored by a camera mounted atop the display. An on-screen event at time g_τ, C_{g_τ}, is observed by the computer user, who has a reaction, D_{g_τ}, expressed in the face. The measurement of this expression is given by some sub-sequence of $\mathbf{Z}$, $Z_{g_\tau} ... Z_{g_{\tau+1}-1}$. In a particular interaction, there will be a sequence of on-screen events, $\mathbf{C} = C_0 ... C_\mathcal{T}$, which cause a sequence of reactions $\mathbf{D} = D_0 ... D_\mathcal{T}$. The time scales of these events are slower than the time scale of the video frame events. The general case will have further temporal dependencies between the facial display and context events, and may be asynchronous in general [8]. The model we describe implicitly assumes that the context events define time intervals for the facial displays. This assumption may work well in many cases (as it does in Section 4), since facial displays are usually timed with other non-facial events [3]. In general, however, a method for temporally segmenting the input sequence is necessary. Previous authors have approached this temporal segmentation problem by exhaustive search for the most likely time scale [17, 16], or by searching for discontinuities in the temporal trajectories [15]. The complete Bayesian solution involves maximizing over temporal segmentations [6].

Our task is now to cluster the sequences of feature vectors, $Z_{g_\tau} ... Z_{g_{\tau+1-1}}$. We use a mixture of hidden Markov models conditioned on the context variable C. The following describes this model, show how to learn the parameters using the expectation-maximization (EM) algorithm, and describes the initialization procedure we have used. We assume that the number of classes of facial displays is known.

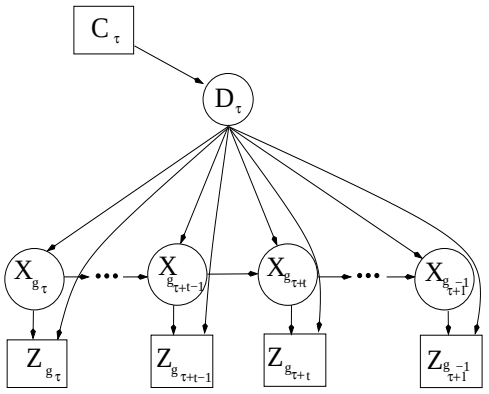


Figure 2: Mixture of hidden Markov models as a dynamic Bayesian network, including context variable, C.

3.1 Mixtures of HMMs

Figure 2 shows a time slice of the context dependent mixture of hidden Markov models (cmHMM) as a dynamic Bayesian network. The description of the entire video sequence ($\mathbf{Z} = Z_0 ... Z_T$) would be a series of such models, each conditioned on an (observable) context variable $\mathbf{C} = C_0 ... C_\mathcal{T}$. As a generative model of the computer user, we can describe Figure 2 as follows. The context C prompts the computer user to perform facial display D. The state of D thus generated is a high level description of the facial display to be performed (such as smile or raise eyebrows), which itself generates a sequence of $T = g_{\tau+1} - g_\tau$ discrete states $\mathbf{X} = X_{g_\tau} ... X_{g_{\tau+1}-1}$. A state X_t generates a specific optical flow description Z_t from a Gaussian distribution over the space of such descriptions. The model is described by four conditional probability distributions: initialization ($P(X_0|D)$), transition ($P(X_t|X_{t-1}D)$ fully connected), observation ($P(Z_t|X_tD)$ parametrized by full covariance Gaussian distributions) and context ($P(D_\tau|C_\tau)$). If we remove the context, C, and the conditioning link, $P(D|C)$, we recover a mixture of hidden Markov models as described in [14]. The context variable, C, is an input to the generative model, which will bias clustering of the input sequences according to the context states, C. This distribution, $P(D|C)$, will be learned simultaneously with the other parameters in the model, and models the effects of the on-screen events on the user.

Given a set of (temporally segmented) feature vectors, $\mathbf{Z}$, and a set of context variables, $\mathbf{C} = C_0 ... C_\mathcal{T}$, we wish to learn the maximum likelihood parameters of the mHMM. That is, we want to find the model parameters, Θ, which maximize the probability

$$P(\mathbf{ZC}\Theta) = \int_{\mathbf{XD}} P(\mathbf{ZXDC}\Theta).$$

We can use the expectation maximization algorithm [5], which lower bounds this probability distribution with a function $q(\mathbf{XD})$, maximizes this bound at the current estimate of the model parameters, Θ', (the 'E' step) giving $q(\mathbf{XD}) = P(\mathbf{XD}|\mathbf{ZC})$ [1], and then maximizes the bound

$$\int_{\mathbf{XD}} P(\mathbf{XD}|\mathbf{ZC}) \log P(\mathbf{ZXDC}) \quad (4)$$

over the model parameters, Θ, (the 'M' step). We factor $P(\mathbf{ZXDC}) = P(\mathbf{ZXD}|\mathbf{C})P(\mathbf{C})$ and the integral in Equation 4 becomes:

$$\int_{\mathbf{XD}} P(\mathbf{XD}|\mathbf{ZC}) \log P(\mathbf{ZXD}|\mathbf{C}) + \int_{\mathbf{D}} P(\mathbf{D}|\mathbf{ZC}) \log P(\mathbf{C}). \quad (5)$$

The first term in Equation 5 is the term that is maximized for a mixture of hidden Markov models, conditioned on the (observed) variable C. Smyth [14] has pointed out that this can be achieved by clustering the variables along each link from D to X variables. The result is a simple hidden Markov model, with hidden states given by the joint variable $\{X, D\}$. The constraint that the variable D does not change over the course of the sequence can be enforced by initializing the the transition matrix for the mixture model, A, as a block diagonal matrix where the blocks are the transition matrices, $A_i = P(X_t|X_{t-1}D_i)$, for each state, i, of the cluster variable, D. The initial state probabilities, $P(X_0|D)$ are then chosen to reflect the weights of the mixture components as given by the conditional distribution $P(D|C)$.

Maximizing the second term in Equation 5 updates the parametrized probability distribution $\Theta_{Dij} = P(D = i|C = j)$ by counting the expected number of times the cluster variable $D = i$ when the context variable $C = j$, N_{Dij}. That is,

$$\Theta_{Dij} = \frac{E_{P(D|\mathbf{ZC})} N_{Dij}}{\sum_i E_{P(D|\mathbf{ZC})} N_{Dij}}$$

where $E_{P(D|ZC)} N_{Dij} = \sum_\tau P(D_\tau = i|\mathbf{ZC})\delta(C_\tau = j)$.

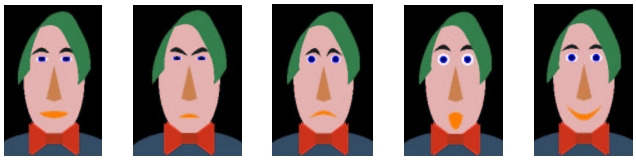
3.2 Initialization

The EM algorithm performs hill climbing on the likelihood surface, and therefore is dependent on the initial choice of the parameters. We use the clustering in log-likelihood space technique described in [14]. It involves fitting a simple HMM to each individual sequence, evaluating the log-likelihood of each sequence given every simple HMM, and then clustering the sequences into K groups using the log-likelihood distance matrix. We use agglomerative clustering with complete linkage (furthest neighbors merging). Simple HMMs are then fit to each of these K clusters, and the results are used to initialize the matrix for the cmHMM. The conditional probabilities of cluster membership, D, given the context variables, C, are initialized by counting the number of observed states C in each cluster. Further details can be found in [8].

[1] We omit explicit representation of Θ, for notational ease

4 Results

To evaluate the model presented in the last section, we asked volunteers to perform a simple facial expression imitation task. They were seated in front of a computer terminal on which an animated cartoon shows facial displays. While many face generation systems use complex 3D graphics, this face is a simple cartoon. This allows for fast rendering, and does not detract from interaction quality, since humans will interact with even the simplest of generated faces as a real human face [13]. Cartoon displays start from a neutral face, as shown in Figure 3(a), then warp to one of the 4 poses shown in Figures 3 for values of $C = 1...4$. These cartoon displays will be referred to as $C_1...C_4$ in the following. The pose is held for roughly a second, and the face then warps back to the neutral pose where it remains for an additional second. Although these displays may be reminiscent of so-called *prototypical facial expressions*, the displays they elicited were clearly *not* expressions of emotion, but only reactions D to contexts C. The subjects were told that their



(a) C_1 C_2 C_3 C_4

Figure 3: (a) neutral face (C=1...4) Faces which subjects were told to imitate

task is to imitate these displays, and were shown each of displays initially and told to practice imitating them. Once they were satisfied with their imitations, they pressed a key, and the system began recording a video sequence through a Sony EVI-D30 color camera mounted above the computer screen. While the subjects were being recorded, the cartoon face performed a series of 40 randomly selected facial displays over a period of 2 minutes.

Zernike feature vectors were recovered for the recorded videos (3600 frames), using a selected basis (specifically ZP coefficients ${}^uA_1^1, {}^vB_1^1, {}^uA_2^2, {}^vB_2^2$). The videos were temporally segmented using the onset times of the cartoon displays and the resulting sequences were input to the HMM clustering and training algorithm described in Section 3, using 3 X states (facial displays are tri-phasic) and 4 clusters (the number of displays the subjects were trying to imitate). The Viterbi algorithm was used to assign cluster membership, D, to each sequence, which were then compared to the known classes of displays the subjects were trying to imitate.

In a cross-validation study, we found that the modeling of the context information did not significantly decrease the likelihoods of test data. The sequences were randomly split into 35 training and 5 test sequences 20 times, and cmHMMs and mHMMs were trained on the training data. The likelihoods of the test data were then evaluated. The results, averaged over the 20 trials and over the 4 subjects were 21.9 ± 2.0 and 22.1 ± 2.0 for the cmHMM and the mHMM, respec-

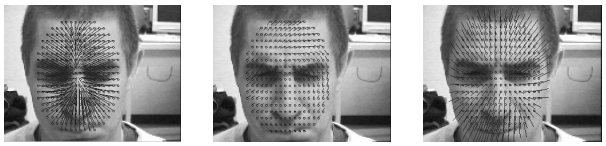
	cluster					cluster			
	D_1	D_2	D_3	D_4		D_1	D_2	D_3	D_4
C_1	2	6	0	0	C_1	4	1	1	0
C_2	0	4	0	2	C_2	5	2	9	0
C_3	0	0	0	12	C_3	0	8	0	0
C_4	7	0	7	0	C_4	0	0	0	10
	$\mathcal{A}$					$\mathcal{B}$			

Table 1: Confusion matrices: subjects $\mathcal{A}$ and $\mathcal{B}$

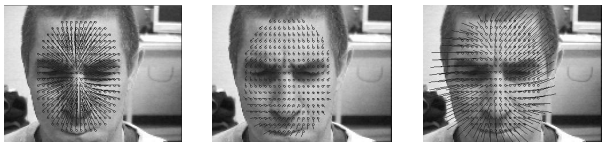
tively. However, qualitative analysis of the clustering over the full data set for each individual showed the addition of context did improve the results.

Only one subject performed significantly different displays in response to C_1 and C_2. After the experiment, most subjects reported either that they did not notice a significant difference between these two cartoon displays, or that they could not find a way to imitate the second one, C_2, due to the extremely down-turned mouth. The one subject who accurately imitated C_2 took more time to practice before starting the experiment. The clustering results for this individual corresponded exactly with the sets of displayed cartoon faces. We will not discuss this one subject further, and show representative results from two of the other four. Complete results can be found in [8]. Table 1 shows the confusion matrix for two subjects, in which each row is one C state (facial display on screen) and each column is one recovered cluster, $D_1 ... D_4$.

Consider subject $\mathcal{A}$. The cmHMM clustered most imitations of C_1 and C_2 together in cluster D_2. Key frames from two sequences in this cluster are shown in Figure 4. Scaled flow fields reconstructed from the feature vectors are shown superimposed. The top sequence was in response to a C_1 cartoon display, while the bottom one was in response to a C_2 cartoon display. Figure 5 shows the feature vector tra-



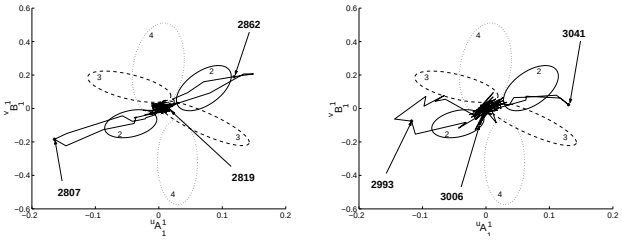
frame= 2807 2819 2862



frame= 2993 3006 3041

Figure 4: Example sequences for subject $\mathcal{A}$. Sequences were of frames 2790-2880 (top row) and 2970-3060 (bottom row)

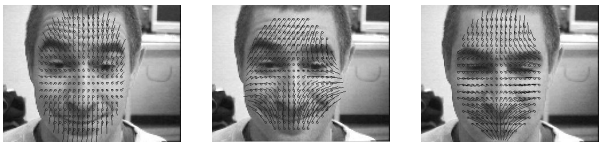
jectories in two of the feature vector dimensions (${}^{u}A_1^1$ and ${}^{v}B_1^1$) for these two sequences. For each value of D=2,3,4, the three Gaussian output distributions (for X=0,1,2) are shown as level curves of the covariance matrix, and are labeled with the D values. The $D = 1$ model (not shown) has a large Gaussian which encompasses most of the others, and seems to be modeling all the sequences which do not fit well into any of the other three models. The remaining three models clearly partition the space evenly, describing vertical expansion and contraction (D=4), horizontal expansion and contraction (D=3), and a combination of both (D=2).



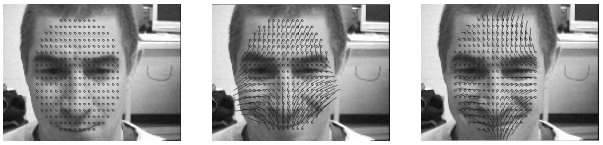
(a) (b)

Figure 5: ZP=1 feature vectors (a), (b) correspond to top and bottom rows in Figure 4

Imitations of C_3 were clustered together, with two imitations of C_2 in the same group These C_2 imitation sequences did not start in a neutral expression (the subject was still reacting to a previous cartoon display). The C_4 imitations were split into two distinct groups (clusters D_1 and D_3), which differed in the amount of eyebrow motion present. Figure 6 shows two sequences which were responses to C_4, but which were clustered into separate groups. The top row sequence contains a significant eyebrow raise (at frame 108), while the bottom one does not. Figure 7 shows the feature vector tra-



frame= 108 114 162

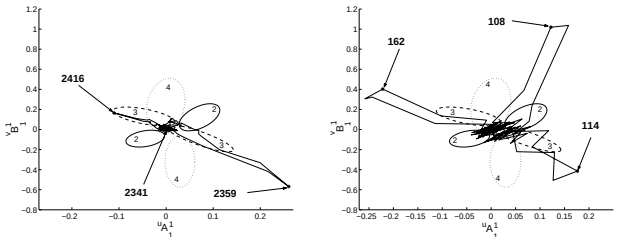


frame= 2341 2359 2416

Figure 6: Example sequences for subject $\mathcal{A}$. Sequences were of frames 90-180 (top row) and 2340-2430 (bottom row).

jectories (again ${}^{u}A_1^1$ and ${}^{v}B_1^1$) for these two sequences. The difference is the additional trajectory in Figure 7(a) which extends into the positive u and v quadrant, and corresponds to the eyebrow raising. The model has uncovered that this is a more important difference in terms of flow fields than any differences between imitations of C_1 and C_2.

Consider now subject $\mathcal{B}$, whose confusion matrix is shown in Table 1. This subject attempted to differentiate between C_1 and C_2, while consistently performing displays C_3 and C_4. The C_2 imitations were split into two clusters: those that were similar to the C_1 imitations, and those that were not. The C_2 imitations grouped together in cluster D_3 contain almost no motion at all. Figure 8 shows key frames



(a) (b)

Figure 7: ZP=1 feature vectors (a), (b) correspond to top and bottom rows in Figure 6

from three example sequences. The top row is a C_1 imitation in cluster D_1, in which the face contracts and then expands. The second row is a C_3 imitation in cluster D_2. The bottom row shows what was supposed to be a C_2 imitation which clustered with the C_3 imitations in cluster D_2. The reason for the apparent mis-classification was a natural *adaptor* which occurred (the subject bit her lip). The subject also closed her eyes (see frame 3405), and seems to have missed the C_2 cartoon display.

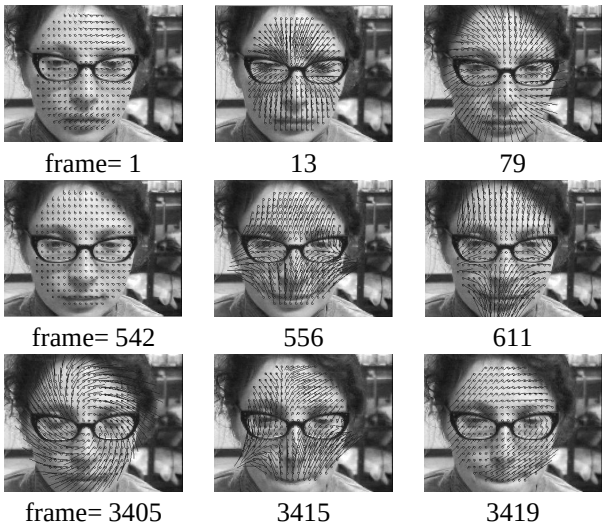


Figure 8: Example sequences for subject $\mathcal{B}$. Sequences are frames 0-90 (top row), 540-630 (middle row), and 3330-3420 (bottom row).

5 Conclusions

We have motivated and presented an approach to adaptive context dependent facial display recognition. We use a holistic optical flow representation projected to a complete orthogonal basis of Zernike polynomials, which is an effective and *a priori* representation of facial motion. The clustering method uses a mixture of hidden Markov models conditioned on a context variable. We have discussed the application of this method to a simple imitation experiment, and have shown how it can uncover classes of facial displays. It is worth noting that the emphasis on context also implies that the recognition of facial expression during speech can not be attempted without an integrated approach [3, 4]. The probabilistic formulation of the models we propose allow them to be integrated with speech recognition, and thus our methods are scalable in this important direction.

References

[1] M. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, January 1996.

[2] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. *International Journal of Computer Vision*, 25(1):23–48, 1997.

[3] J. Cassell. Embodied converstation: Integrating face and gesture into automatic spoken dialogue systems. In S. Luperfoy, editor, *Spoken Dialogue Systems*. MIT Press, in press.

[4] N. Chovil. Social determinants of facial displays. *Journal of Nonverbal Behavior*, 15(3):141–154, Fall 1991.

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

[6] S. Fine, Y. Singer, and N. Tishby. The hierarchical Hidden Markov Model: Analysis and applications. *Machine Learning*, 32:41, 1998.

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

[8] J. Hoey. Clustering facial displays in context. Technical Report TR-01-17, University of British Columbia, Vancouver, BC, November 2001.

[9] J. Hoey and J. J. Little. Representation and recognition of complex human motion. In *Proc. IEEE CVPR*, Hilton Head, SC, June 2000.

[10] Y. li Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), February 2001.

[11] M.Turk and A.Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[12] A. Prata and W. Rusch. Algorithm for computation of Zernike polynomials expansion coefficients. *Applied Optics*, 28(4):749–754, February 1989.

[13] B. Reeves and C. Nass. *The media equation*. Cambridge University Press, 1996.

[14] P. Smyth. Clustering sequences with hidden Markov models. In *NIPS 10*, 1997.

[15] M. Walter, A. Psarrou, and S. Gong. Data driven gesture model acquisition using minimum description length. In *Proc. BMVC*, Manchester, UK, September 2001.

[16] A. D. Wilson, A. F. Bobick, and J. Cassell. Temporal classification of natural gesture and application to video coding. In *Proc. CVPR*, Puerto Rico, June 1997.

[17] C. R. Wren, B. P. Clarkson, and A. P. Pentland. Understanding purposeful human motion. In *Proc. Face and Gesture Recognition*, Grenoble, France, 2000.