Affect Valence Inference From Facial Action Unit Spectrograms

Daniel McDuff
MIT Media Lab
MA 02139, USA
djmcduff@mit.edu

Rana El Kaliouby
MIT Media Lab
MA 02139, USA
kaliouby@mit.edu

Karim Kassam
Harvard University
MA 02138, USA
kskassam@fas.harvard.edu

Rosalind Picard
MIT Media Lab
MA 02139, USA
picard@mit.edu

Abstract

The face provides an important channel for communicating affect valence, the positive or negative emotional charge of an experience. This paper addresses the challenging pattern recognition problem of assigning affect valence labels (positive, negative or neutral) to facial action sequences obtained from unsegmented videos coded using the Facial Action Coding System (FACS). The data were obtained from viewers watching eight short movies with each second of video labeled with self-reported valence and hand coded using FACS. We identify the most frequently occurring Facial Actions and propose the usefulness of a Facial Action Unit Spectrogram. We compare both generative and discriminative classifiers on accuracy and computational complexity: Support Vector Machines, Hidden Markov Models, Conditional Random Fields and Latent-Dynamic Conditional Random Fields. We conduct three tests of generalization with each model. The results provide a first benchmark for classification of self-report valences from spontaneous expressions from a large group of people (n=42). Success is demonstrated for increasing levels of generalization and discriminative classifiers are shown to significantly outperform generative classifiers over this large data set. We discuss the challenges encountered in dealing with a naturalistic dataset with sparse observations and its implications on the results.

1. Introduction

Affect valence is the positive or negative emotional charge of an event or experience [17]. The need to quantify affect valence arises in numerous domains, ranging from medical or psychological studies of well-being to market research, advertising, usability and product evaluation [8]. Self-report is the current gold-standard measure of affect valence, where people are interviewed, asked to rate their feeling on a Likert scale or turn a dial to quantify valence.

While convenient and inexpensive, self-report is problematic because it is subject to biasing from the interviewer, the context and other factors of little interest [19]. The act of introspection is challenging to perform in conjunction with another task and may in itself alter that state [13]. Finally, like most affective states, affect valence is a dynamic phenomena that unfolds and changes over time. Self-report measures do not capture these dynamics, unless solicited frequently over the course of an event; however, that interrupts the experience and disturbs emotional engagement with the object of interest.

Unlike self-report, facial expressions are implicit and do not interrupt a person's experience. In addition, facial expressions are continuous and dynamic, allowing for a representation of how valence changes over time. However, there are thousands of combinations of facial and head movements that occur in natural situations and it is a challenging problem to learn which ones are reliable and relevant predictors of affective valence. For example, "true smiles" can occur with negative experiences [18]. The Facial Action Coding System (FACS) [4, 21] is a catalogue of unique action units (AUs) that correspond to each independent motion of the face. FACS enables the measurement and scoring of facial activity in an objective, reliable and quantitative way, and is often used to discriminate between subtle differences in facial motion. FACS does not depict what the underlying affective state is. Thus, "expert" labelers are often employed to label video or AU sequences into corresponding affective states. For example, two or more labelers may be shown video segments and asked to rate each segment's valence as positive or negative based on the observed AUs. Inter-coder agreement is then computed to determine the reliability of that valence label.

The current work examines whether head and facial action unit labels provide a reliable measure of affect valence. Our work addresses four new challenges that, to the best of our knowledge, have not been addressed before. First, unlike existing studies that rely on "expert labels" for acted or "posed" videos, here training and testing are performed on spontaneous videos with affect valences self-reported by the viewers whilst they watched movie clips. In our work self-report measures were used as the labels. While we expect

1

that in most cases self-report reflects the observed facial expressions, there are cases where there may be a mismatch between the two. Also, the dataset presents highly variable head and facial activity depending on the viewer and the movie clip. Together these variables result in an ecologically valid, albeit noisy and challenging dataset from a machine learning perspective. Secondly, we are interested in the dynamics of affect valence, and how these dynamics change over time in relation to the person's head and facial activity. Thus in our experiments, we model the temporal progression of affect valence with every second of video and compare that to self-report valence. We then compare the accuracy of static and dynamic classifiers in mapping head and facial action unit sequences to affect valences. We use a 'trinary' valence classification system with negative and positive at the extremes and a neutral state to represent states that are neither strongly positive or strongly negative. Third, unlike existing facial datasets our naturalistic dataset is characterized by sparse facial and head actions; we explore the implications of the sparsity of data on the performance of generative versus discriminative models. We test their ability to generalize to an unseen viewer and an unseen context. Finally we examine the possibility of an inherent lag between the expression of an AU, or set of AUs, and the corresponding valence report. These four challenges addressed in this paper help to advance understanding of how AUs relate to self-report valence in naturalistic situations.

Affect expression and appraisal is subject to subtle nuances that vary from person to person [14]. Training a model on data from a self-report study such as ours reveals subtleties of spontaneous naturalistic behavior that are discriminative for a large group. We learned that a slight lip part was highly discriminative for positive affect something we would not have predicted using only posed expressions.

In the remainder of this paper, we demonstrate feature selection to identify discriminative AUs. We present a novel spectrogram for facial and head action units to represent the unfolding of action units over time and their association with self-report of affect valence. This is a powerful visualization tool that can be applied in many areas that involve automated facial and gesture analysis. The temporal dependency between feature observation and relevant affect report is analyzed. We then conduct a series of experiments comparing static and dynamic, generative and discriminative classifiers, over three generalization tests, to map sequences of head and facial actions into affect valences, providing the first rigorous baseline for labeling dynamically-changing valence on spontaneous expressions for a large group of people (n=42.)

2. Related Work

Affect recognition from naturalistic behavior, in particular sequential valence recognition, is relatively unexplored.

This is partly due to the many challenges inherent in collecting naturalistic data and obtaining "ground-truth" labels needed to train and test affective models. To date, most efforts train on data of actors demonstrating deliberate affective states or cull the naturalistic data, chopping it into short, segmented clips based on which chunks have labels agreed to by experts. El Kaliouby [9] demonstrates inference of complex mental states from facial expression and head gesture sequences using a dynamic Bayesian network; however she trains on segmented video clips of actors. Similarly, Bartlett et al. [1] present automated facial expression recognition by training on data that has been segmented and subjected to agreement by expert labelers. This approach limits the generalization to naturalistic data, and we know there is a clear distinction between spontaneous and deliberate facial behavior [16].

In this paper, we investigate facial valence recognition on civilian movie viewers whose facial expressions were FACS coded every second and matched with their self report valence for the corresponding second. Our training and testing data are of spontaneous facial behavior that is continuous (unsegmented) and not "expert" labeled; we use all the video, not just the segments agreed to be in certain categories based on a panel of experts. Zeng et al. [22] also consider classification of spontaneous facial expression into positive, negative and neutral classes. The authors report promising results on two persons: an accuracy of 87% on a female and 79% on a male. However, these results are obtained using person-dependent k-fold cross validation. We provide the first benchmark for spontaneous facial affect using a group of 42 people. We present person-independent as well as movie-independent experiments, comparing state of the art generative and discriminative models.

This paper addresses labeling continuous sequences of noisy high-dimensional unsegmented data. Much work has been done on the labeling of segmented and unsegmented videos, particularly in the area of gesture recognition [11, 15]. There are many similarities to visual gesture recognition systems in our problem. Most gesture recognition systems employ Hidden Markov Models (HMMs). However recently other models have been demonstrated for this task. Conditional Random Fields (CRFs) and a variant, latent dynamic Conditional Random Fields (LDCRFs), have been demonstrated to perform well on unsegmented gesture classification [20, 15]. Chang et al. [3] present an application of partially-observed hidden conditional random fields for facial expression recognition. However this is the first application of CRFs and LDCRFs to naturalistic facial affect recognition. We contrast the success of these models and compare them to multi-class Support Vector Machines and a set of Hidden Markov Models.

3. Data Collection

We obtained 336 FACS-coded movies of 42 viewers each watching eight different movie clips. The viewers (20 females, mean age = 25.6, SD = 10.1) were instructed to watch clips (11 to 204 seconds) while simultaneously rating their feelings on a slider bar on the screen that ranges from extremely unpleasant (-100) to extremely pleasant (100). The movie clips were previously identified as strong elicitors of the six basic emotions [7]: Robin Williams Live (amusement, 81s), When Harry Met Sally (amusement, 154s), Cry of Freedom (anger, 155s), Pink Flamingos (disgust, 55s), Silence of the Lambs (fear, 204), Capricorn One (surprise, 47s), Sea of Love (surprise, 12s), Lion King (sadness, 131s). The viewers did not need to change their line of sight to see the scale relative to the movie. To obtain the highest psychological validity of our findings we use the validated human-coded AUs and use automated methods only for the mapping to valence. FACS coding was performed by two certified FACS coders. Any discrepancies were revisited and a common agreement found. This yielded a binary label and intensity (A to E) for each of 65 action units (AUs) and action dynamics (ADs), per second. For the remainder of the paper we refer to AUs and ADs as AUs. Only FACS action codes of intensity C or higher were considered as indicators of an AU. Existing work shows successful detection of even the most subtle AUs such as AU11 [6].

Self-report data is inherently noisy and particularly prone to high levels of variance. To reduce the effect of variations in the reference levels from viewer to viewer the valences for each viewer watching each clip were normalized to a -100 to 100 scale. The normalized valences were classed into three categories: positive (>10), negative (< -10) and remainder as neutral. This gave priors of: positive 45.5%, negative 44.8% and neutral 9.71%. This class system with a majority positive and negative labels was chosen as we did not wish to have unnecessary loss of information about positive and negative indicators.

4. Feature Selection

A total of 65 AUs were coded. We observed that a subset of these AUs were not appropriate for our study, such as obscured facial features (70-75), or appeared in too many cases to be discriminative, such as blinking (45). These were removed from the feature set.

We wanted to identify the most discriminative AUs and how they unfolded over time in relation to self-reported valence across viewers of a certain movie. We devised a novel way to aggregate and visualize AU labels across viewers and over time: the Facial Action Unit Spectrogram (FAUS), shown in Figure 1 encodes the frequency of each AU per second across all viewers of a specific movie clip. Each
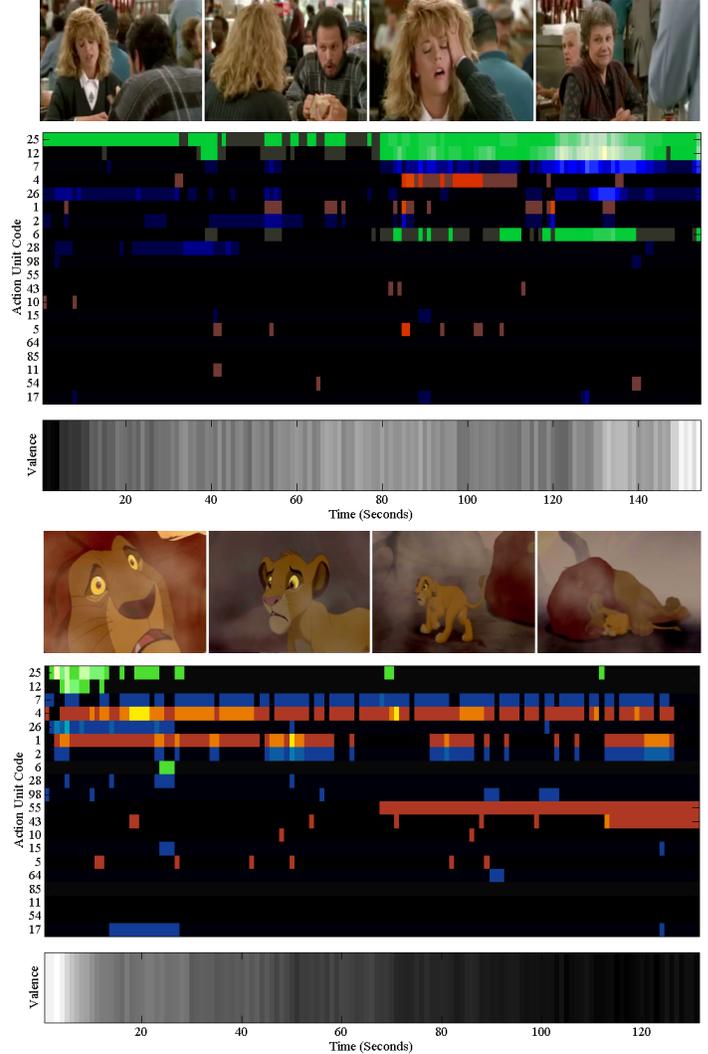

Figure 1. Facial Action Unit Spectrogram across viewers showing the frequency of head and facial action units over time. High intensity pixels denote relatively high frequency occurrence. The color represents the association of the AU computed from the discriminative power. The valence plot shows the average normalized valence of the viewers. Positive valence is denoted by light areas and negative valence by dark areas.

row represents an AU, starting at the top with the AU that occurred most frequently throughout the entire set of data (AU25). The intensity of a pixel across each row represents the relative frequency: the greater the intensity, the greater the number of viewers that exhibited that AU at that second.

Each row is assigned a color to reflect its association with positive (green) or negative (red) valence labels (computed from the discriminative power). The blue row correspond to AUs with a discriminative power, $|P_i| < 0.1$. We computed the discriminative power P of each AU_i as follows:

$$P_i = P(Positive|AU_i) - P(Negative|AU_i) \quad (1)$$

The spectrogram plot is matched with an aggregated valence plot that gives an indication of the valence rating across the pool shown in Figure 1. The valence map shows the mean normalized valence, higher intensity corresponding to more positive reports on average, relative to the reference for that movie. Figure 1 (top) shows the FAUS for the cafe scene in 'When Harry Met Sally' [7] and (bottom) the Mufasa death scene from 'Lion King'.

From these spectrograms we were able to easily identify significant features associated with different valence trends. Of the eight movie clips there were clearly a sub-set of more expressive clips when comparing the FAUS. As a result the data set was constrained to samples where the viewers were watching one of the five more expressive clips: Robin Williams Live, Cry of Freedom, Pink Flamingos, Lion King and When Harry Met Sally.

It was possible through the visualization to identify specific features that appeared consistently in a single class and those that were present in both. In particular, we were able to identify differences in the AU combinations that occurred during movies labelled as negative. Commonly occurring combinations during the 'Lion King' (sadness) and 'Cry of Freedom' (anger) were noticeably different from those during 'Pink Flamingos' (disgust). This suggests in certain cases inference about the type of negative label may be possible. However this distinction is not explored further here.

It was not possible to select features solely by a discriminative power criteria. This is because many of these examples occurred in relatively few of the 24,000 samples. The AUs were subsequently ordered by frequency of occurrence across the constrained data set and the 20 most frequently occurring selected. Those discarded occurred in less than 0.15% of the samples. A full list of the 20 AUs and their codes is given in Table 1. Discriminability analysis was then performed on these.

Figure 3 plots the magnitude of the discriminative power of the 20 most frequently occurring AUs, given the constrained data set. The magnitude of P_i quantifies the discriminative power of an AU for affect valence; the color depicts whether its presence signals a positive or negative valence. The most discriminative AU was AU11 Nasolabial deepener, followed by AU85 or head nod, then by AU55, head tilt left. AU28 had a discriminative power of 0.001, implying that this AU for this study did not encode any important valence information. Although 10 of the features were stronger negative discriminators, compared to four positive discriminators, the positive discriminators occurred relatively frequently.

The FAUS in Figure 1 also shows that the action units for the videos were relatively sparse. Where the inputs of

AU Code	Description	AU Code	Description
25	Lips Part	43	Eyes Closed
12	Lip Corner Pull	98	Head Scratch
7	Lid Tightener	15	Lip Corner Depressor
4	Brow Lowerer	64	Eyes Down
26	Jaw Drop	10	Upper Lip Raiser
6	Cheek Raiser	5	Upper Lid Raiser
1	Inner Brow Raiser	85	Head Nod
2	Outer Brow Raiser	11	Nasolabial Deepener
28	Lip Suck	17	Chin Raiser
55	Head Tilt Left	54	Head Down

Table 1. Top 20 AUs in discriminability and frequency of occurrence.

Figure 2. The five most frequently occurring AUs in our data set. From left to right: 25 - Lips Part, 12 - Lip Corner Pull, 7 - Lid Tightener, 4 - Brow Lowerer, 26 - Jaw Drop [10].
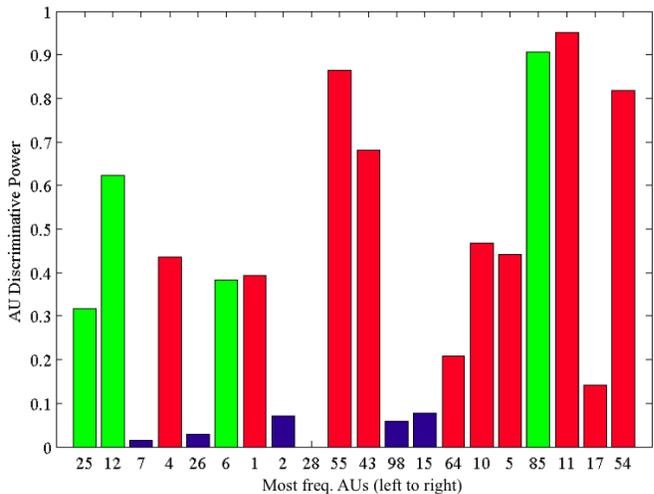

Figure 3. Chart showing the discriminative power of each AU computed as in Equation (1). The 20 most common action units across the entire data set are shown. Green bars denote AUs with a majority positive associated labels and red bars denote AUs with a majority negative labels. Blue denotes an AU with a discriminative power < 0.1. None of these AUs were associated with a majority neutral labels.

the model were all zeros the example was removed from the validation, training and testing sets. It was not assumed that the model could infer labels for incomplete data where inputs at particular instances were zero.

5. Classification

We initially analyze the temporal dependence between observed AUs, the features, and reported labels using an SVM classifier. Three further experiments were carried out
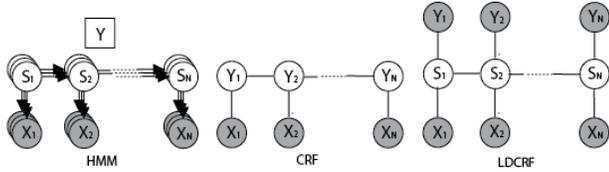
Figure 4. Structure of models. X_j represents the jth observation, S_j the jth hidden state and Y^j the jth label where appropriate. The HMM model requires a chain to be trained for each class label. The CRF [12] and LDCRF [15] models produce labels for each second of the sequence. We use a moving window, selecting the final label each time. With the CRF and LDCRF long range dependencies are possible, with the HMM they are not [15].

on the data in order to evaluate the success of the classifiers with different dependence assumptions between the training data and the test data. We evaluated the success of several classifiers in these tests, investigating static vs. dynamic and generative vs. discriminative models. The experimental methodology is described in Section 6. Classifiers designed for this task must be robust to new data. Expressions are person dependent and context dependent in this work. Unlike the one existing study in this area [22], we test the classifiers on their ability to generalize to unseen viewers and also to unseen contexts. As described above CRFs and their variants have been shown to perform favorably relative to HMMs on sequential labeling tasks [12]. There are distinct benefits when considering labeling of unsegmented sequences.

The window size was varied between zero and two seconds. A window of zero seconds corresponds to taking the current data point in each instance. Varying the window size allows for different length dependencies to be modeled. Dependencies longer than three seconds were not considered.

The SVMs were implemented using LIBSVM [2]. The HMMs were implemented using the HMM toolbox[1] for MATLAB. The CRF and LDCRF classifiers were all implemented using the HCRF toolbox[2]. All prototyping was carried out in MATLAB. In all cases parameter selection was performed using a Bayesian Information Criterion (BIC).

5.1. Support Vector Machine

A Support Vector Machine classifier was used as the first benchmark. This is a static approach to classification and therefore the data set was separated into a sample per second. The multi-class classifier was trained with one label for each valence level. A Radial Basis Function (RBF) kernel was used. Where the model was tested for a window size > 0 the observations were appended to one another. During validation the penalty parameter, C, and the RBF kernel

parameter, γ, were each varied from 10^k with k=-3,...3.

5.2. Hidden Markov Model

HMMs are commonly used in sequential labeling and are used here to provide a benchmark. An HMM was trained for each of the three valence classes. This is a generative approach to modeling the data. HMMs require segmented sequences associated with the relevant class for training, the data was therefore segmented into subsequences where the final labels all belonged to the same class. In testing the class label associated with the highest likelihood HMM was assigned to the final frame in the sequence.

5.3. Conditional Random Fields

In contrast to HMMs, CRFs and CRF variants are discriminative approaches. CRFs allow for modeling of extrinsic dynamics advantageous in labeling unsegmented sequences, as we are considering. The model removes the independence assumption made in using HMMs and also avoids the label biasing problem of Maximum Entropy Markov Models (MEMMs)[12]. The dynamics of expressions are significant in distinguishing between them [16], as such we hypothesized a potential benefit in removing the assumption that current features are solely dependent on the current valence label. A single CRF was trained with a state label corresponding to each valence class. As a result no segmentation of the data sets was required. Regularization terms from 10^k with k=-3,...,3 were compared.

5.4. Latent-Dynamic Conditional Random Fields

Morency et al. [15] demonstrate the use of LDCRFs in continuous gesture recognition. A similar technique is used here to combine the advantages of CRFs and HCRFs [20] capturing the dynamics of both. As a result intrinsic and extrinsic dynamics can be modeled. Unsegmented data as used for training the CRF model was used for training the LDCRF model. During validation the regularization factor (10^k with k=-3,...,3) and hidden states (0,...,9) were varied.

6. Results and Discussion

The facial action unit spectrograms in five of the eight movie clips showed significantly more facial expressions. The resulting set of data included 42 viewers and five movies giving a set of over 24,000 examples. We conduct several experiments that 1. explore time-delays between facial expressions and self-report labels; 2. compare discriminative to generative models, 3. compare models that explicitly model dynamics with classifiers that don't on three different observation lengths or window sizes (0, 1 and 2) and 4. test the models under three generalization conditions.

The HMM models required no more than 100 iterations of the training scheme. The CRF model needed a maxi-
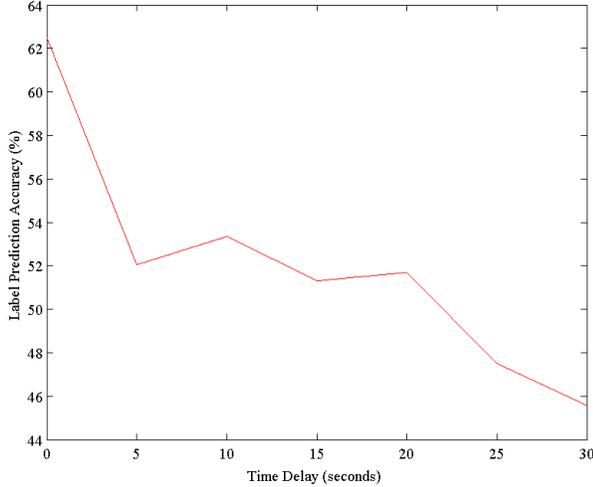
---

[1]K. Murphy - Hidden Markov Model (HMM) Toolbox for MATLAB
[2]L-P. Morency - Hidden-state Conditional Random Field Library

Figure 5. Plot of accuracy of SVM predicting labels at time t+delay from observations at time t, where delay is {0, 5, 10, ...,30}.

Window Size (s)	Accuracy (%)		
	0	1	2
SVM	60.56	59.19	60.10
HMM	34.45	38.19	26.89
CRF	60.66	62.22	59.65
LDCRF	59.76	59.38	57.25

Table 2. Comparison of Accuracy of Models for Viewer and Movie Dependent Case.

Window Size (s)	Accuracy (%)		
	0	1	2
SVM	58.79	54.77	52.81
HMM	32.76	29.13	35.55
CRF	57.42	55.57	54.72
LDCRF	55.66	55.95	55.48

Table 3. Comparison of Accuracy of Models for Viewer Independent Case.

Window Size (s)	Accuracy (%)		
	0	1	2
SVM	47.58	52.99	48.89
HMM	48.35	26.71	35.86
CRF	58.41	57.58	52.99
LDCRF	57.10	55.20	55.55

Table 4. Comparison of Accuracy of Models for Movie Independent Case

mum of 200 iterations in training and the LDCRF 300 iterations. When using large window sizes the computational cost training the CRF and LDCRF models became a significant issue. K-fold validation in such a case required considerable computational power. At most, validation and training of the model took 20 minutes on a machine with an Intel 3.00GHz processor and 3.25 GB of RAM.

6.1. Observation-Label Time Delay Experiment

To establish if there was a systematic delay between the time at which features were expressed and the self-reported valence, we tested a static SVM classifier while varying the time delay between AU observations and self-report labels. We experimented with one-second delays starting from zero (i.e. no delay) to ten seconds, and then every five seconds up to a 30 second delay. For each delay, we randomly pick 10 viewers and set aside their data for validation. From the remaining data of 32 viewers, we hold-out one video for testing and train on the rest, using the SVM parameters obtained from model selection. Using K-fold cross validation, we repeat this process 336 times, once for each viewer-movie combination. Figure 5 shows the results at delays of 0, 5, 10 up to 30 seconds. The best performance was obtained when there were no delays between observations and labels. In all remaining experiments, we assume that there is no time delay between the observation and label.

6.2. Viewer and Movie Dependent Experiment

In the next three sections we compare the SVM to three dynamic models (HMM, CRF and LDCRF) on three different observation lengths (0, 1 and 2) and three generalization conditions (viewer-movie dependent, viewer-independent and movie independent). In the first experiment, validation, training and testing criteria are as described in Section 6.1 where a single video (a specific viewer-movie combination) is held-out for testing, meaning that there will be examples of that viewer and that movie in the training set. Table 2 shows the accuracy of the models for window sizes, 0, 1 and 2 seconds. The SVMs, CRFs and LDCRFs yield comparable results (average accuracy of 62%), while the HMM is clearly worse (average accuracy of 35%). Window size does not have a significant impact on accuracy.

6.3. Viewer-Independent Experiment

In this generalization experiment, we held-out data pertaining to one viewer (i.e. all five movies of that person), randomly selected 10 viewers from the remaining set and used their data (i.e. all 50 movies) for validation, and then trained on the data from the remaining 31 viewers. Using K-fold cross validation, this process was repeated for all 42 viewers. This is a challenging test for the models since people vary a lot in how expressive they are as well as in the specific way with which they express positive or negative experiences. Table 3 shows the accuracy of the models for window sizes, 0, 1 and 2 seconds. As in the first ex-
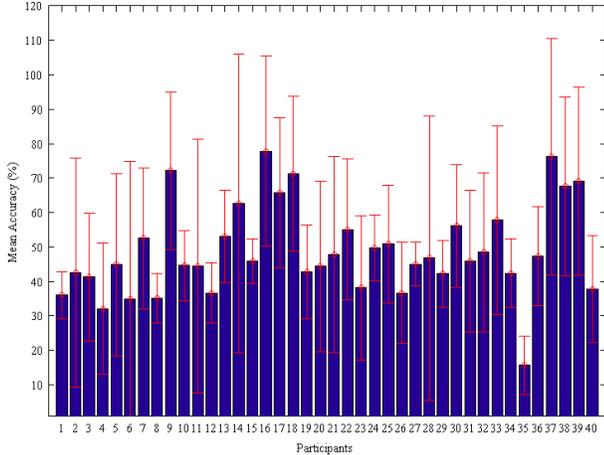
Figure 6. Viewer-independent experiment: the average accuracy and variance of the SVM, CRF and LDCRF per viewer, the error bars show the standard deviation either side of the mean. Only 40 bars are shown as 2 of the 42 participants did not exhibit any AUs.

periment, the SVMs, CRFs and LDCRFs outperform the HMM and exhibit similar performance across all windows sizes. Figure 6 shows the average accuracy and variance of the discriminative models (SVM, CRF, LDCRF) per viewer. Viewers 16 and 37 exhibit the highest accuracy; viewer 35 shows the lowest accuracy.

6.4. Movie-Independent Experiment

In the third and arguably most challenging generalization condition, we held-out data pertaining to a movie (i.e. all 42 viewers of that movie). Validation was performed on data from 10 viewers and trained on the data from the remaining 32 viewers. In both cases the movie being tested was excluded from these sets. This was repeated for all five movies using K-fold cross validation. Table 4 shows the accuracy of the models for window sizes, 0, 1 and 2 seconds. The SVMs, CRFs and LDCRFs continue to outperform the HMM. Again, the window size does not appear to have any impact on the results, except for the HMM. Figure 7 shows the average accuracy and variance of the discriminative models (SVM, CRF, LDCRF) per movie. The amusement movies ('When Harry Met Sally' and 'Robin Williams') yielded higher accuracy than the anger movie ('Cry Freedom') and the sad movie ('Lion King').

6.5. Discussion

In the first experiment we considered the possibility of a lag between a person's facial expression and his/her cognitive appraisal of valence through self-report, expecting to find cases where a person's expression preceded their self-report label. We found that within a 20 second delay, the accuracy of the SVM classifier was not affected much by the
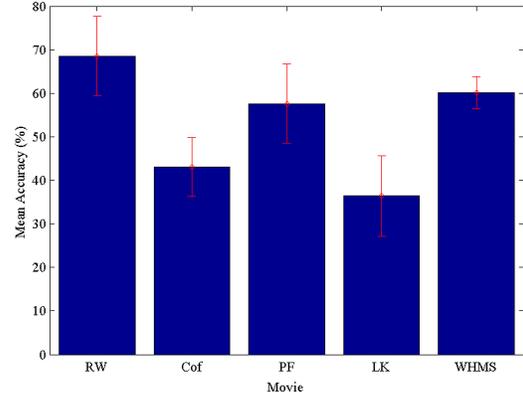

Figure 7. Movie-independent experiment: the average accuracy of the SVM, CRF and LDCRF per movie, the error bars show the standard deviation either side of the mean.

time delay, ranging from 52% to 62.5% with the highest accuracy yielded when AUs at a specific moment in time were compared to self-report labels at that same instance. This general insensitivity to the time delay could be explained by the fact that, in this specific dataset, there were no dramatic changes in valence within any one movie clip.

As expected, and consistent with similar generalization experiments in pattern recognition and machine learning [5], the accuracy of the models decreased with the increasing difficulty of the generalization condition. This is especially a challenge in our case because our dataset is spontaneous (whereas in posed datasets, "actors" receive clear instructions on how to act). The viewer-movie dependent condition performed the best (62.5% on average) followed by the viewer-independent condition (55% on average) then by the movie-independent condition where the average accuracy was 45%. It was particularly interesting to see that the amusement movies yielded a better accuracy compared to sad movies. At the same time, the viewer-independent results underscore differences between people in how they express an affective state.

The discriminative models (SVM, CRF and LDCRF) outperformed the generative models (HMMs) on this dataset. The were no significant results between window sizes. As such the benefits of the CRF in capturing extrinsic dynamics and the LDCRF in capturing intrinsic and extrinsic dynamics was not demonstrated. The impact of increased window size may be seen for larger window sizes.

7. Conclusions

This work provides several new insights into valence classification from spontaneous naturalistic facial actions. First, a large amount of naturalistic data was collected from 42 people, and all AU's labeled by human experts to give

maximal accuracy. We then trained a variety of static and dynamic models, both generative and discriminative, to see how well these "ideal AU's" mapped to self-reported valence. The results definitively show that the discriminative models give greater performance than generative models, in part due to the sparsity of the input features.

No significant lag between facial expressions and self-report response was found. New facial action unit spectrograms were constructed, showing a fast way to visualize a lot of AU information related to valence (or other desired labels) in an easy graphical way. Using these spectrograms, it was identified that in certain cases there existed a distinct time delay between common occurrence of a particular action unit and a local maxima in the mean valence report. This suggests that there may be specific temporal dynamics, perhaps associated with groups of AU's, that could further improve performance.

The models were evaluated over a series of training and test sets with increasing generalization. The performance was greatest with a common viewer and context in the training and testing sets. Significant performance was also demonstrated when generalizing to an unseen viewer and also when generalizing to a new context.

Our results show learnable mappings from naturalistic expressions to self report valence using AUs.

References

[1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, page 568. Citeseer, 2005. 2

[2] C. Chang and C. Lin. LIBSVM: a library for support vector machines, 2001. 5

[3] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:533–540, 2009. 2

[4] P. Ekman and W. V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists, 1978. 1

[5] R. El Kaliouby and P. Robinson. Generalization of a vision-based computational model of mind-reading. *Proceedings of First International Conference on Affective Computing and Intelligent Interaction*, pages 582–589, 2005. 7

[6] Y. Gizatdinova and V. Surakka. Automatic detection of facial landmarks from au-coded expressive facial images. *Image Analysis and Processing, International Conference on*, 0:419–424, 2007. 3

[7] J. Gross and R. Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995. 3, 4

[8] R. Hazlett and J. Benedek. Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies*, 65(4):306–314, 2007. 1

[9] R. el Kaliouby and P. Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *2004 IEEE Workshop on Real-Time Vision for Human-Computer Interaction at the 2004 IEEE CVPR Conference*, 2004. 2

[10] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the fourth IEEE International conference on Automatic Face and Gesture Recognition*, page 46, 2000. 4

[11] A. Kapoor and R. Picard. A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5. ACM New York, NY, USA, 2001. 2

[12] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 282–289. Citeseer, 2001. 5

[13] M. Lieberman, N. Eisenberger, M. Crockett, S. Tom, J. Pfeifer, and B. Way. Putting feelings into words: affect labeling disrupts amygdala activity in response to affective stimuli. *PSYCHOLOGICAL SCIENCE-CAMBRIDGE-*, 18(5):421, 2007. 1

[14] D. Matsumoto and C. Kupperbusch. Idiocentric and allocentric differences in emotional expression, experience, and the coherence between expression and experience. *Asian Journal of Social Psychology*, 4(2):113–132, 2001. 2

[15] L. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*. Citeseer, 2007. 2, 5

[16] M. Pantic. Machine analysis of facial behavior: Naturalistic and dynamic behavior. *Phil. Trans. Royal Society B*, 2009. 2, 5

[17] D. Sander. Oxford Companion to Emotion and the Affective Sciences. pages 401–402, 2009. 1

[18] K. Schneider and I. Josephs. The expressive and communicative functions of preschool children's smiles in an achievement-situation. *Journal of Nonverbal Behavior*, 15(3):185–198, 1991. 1

[19] N. Schwarz and F. Strack. Reports of subjective well-being: Judgmental processes and their methodological implications. *Well-being: The foundations of hedonic psychology*, pages 61–84, 1999. 1

[20] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2006. 2, 5

[21] T. Wehrle and S. Kaiser. Emotion and Facial Expression. In J. A. Russell and J.-M. Fernandez-Dols, editors, *The Psychology of Facial Expression*, pages 49–63. Cambridge University Press, 1997. 1

[22] Z. Zeng, Y. Fu, G. Roisman, Z. Wen, Y. Hu, and T. Huang. Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5):1–8, 2006. 2, 5