Audio-Visual Emotion Challenge 2012: A Simple Approach

Laurens van der Maaten*
Pattern Recognition and Bioinformatics Group, Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
lvdmaaten@gmail.com

ABSTRACT

The paper presents a small empirical study into emotion and affect recognition based on auditory and visual features, which was performed in the context of the Audio-Visual Emotion Challenge (AVEC) 2012. The goal of this competition is to predict continuous-valued affect ratings based on the provided auditory and visual features, *e.g.*, local binary pattern (LBP) features extracted from aligned face images, and spectral audio features.

Empirically, we found that there are only very weak (linear) relations between the features and the continuous-valued ratings: our best linear regressors employ the offset-feature to exploit the fact that the ratings have a dominant direction (more increasing than decreasing). Much to our surprise, only exploitation of this bias already leads to results that improve over the baseline system presented in [10]. The best performance we obtained on the AVEC 2012 test set (averaged over the test set and over four affective dimensions) is a correlation between predicted and ground-truth ratings of 0.2255 when making continuous predictions, and 0.1921 when making word-level predictions.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Applications

Keywords

Social signal processing; Affective computing.

1. INTRODUCTION

Social signal processing aims to automatically identify social cues by inspecting modalities such as facial expressions, human body pose, gestures, non-verbal auditory information, and speech [14]. Whilst the development of systems that automatically recognize basic emotions (see [1] for an overview) or facial action units (*e.g.*, [4, 2, 12]) has seen great improvements over the last decade, the automatic recognition of higher-level social signals such as *arousal*, *expectancy*, *power*, and *valence* is still very difficult. To foster the development of new techniques for the identification of such social signals, Schuller *et al.* [11, 10] recently started organizing the Audio-Visual Emotion Challenge (AVEC). The goal of this competition is to predict ratings of four affective dimensions that were obtained by soliciting human raters (viz. of arousal, expectancy, power, and valence). The ratings are continuous and specified for each frame in the videos, *i.e.* they are measured at 25 frames per second. The competition consists of two main parts, viz. the fully continuous sub-challenge (FCSC) and the word-level sub-challenge (WLSC). In the FCSC challenge, a prediction needs to be made for every frame in the videos, *i.e.* predictions need to be made at 25 frames per second. In the WLSC challenge, word timings are provided and predictions are made of the average rating during the utterance of a single word.

The paper describes our submission to the AVEC 2012 competition, which comprises a very simple system that nonetheless appears to outperform the baseline system described by Schuller *et al.* [10]. Specifically, we use very simple linear regressors because we surmise these are most robust against the large amounts of noise in the input features and because it is straightforward to train on massive databases such as the AVEC 2012 data. We did experiment with more complex models including conditional random fields [3, 7, 13], linear dynamical models [9], and Gaussian Process regressors [8] in preliminary experiments; but we could not establish any performance improvements using these more complicated models. In part, this may be because the relations between the data and the target values appear to be very weak: our best predictors largely ignore the observations when making predictions (see Section 4). In future work, we intend to capture as much of the learning signal as possible using ensemble methods.

In our study, we did not focus on extracting better video or audio features, and we also did not try to use the word annotations in our predictions; we merely tried to maximize the performance whilst using the pre-computed features as provided by the organizers of the challenge [10]. We surmise that the extraction of more elaborate features will, however, be necessary to obtain truly good performances on the AVEC 2012 competition data.

We describe our system in Section 2, and present the results of our experiments in Section 3. We conclude the paper with a discussion in Section 4.

*The website of Laurens van der Maaten can be found at: `http://homepage.tudelft.nl/19j49`

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

2. SYSTEM

Our system comprises two main parts: (1) a part that performs feature extraction and (2) a part that regresses the resulting features onto the four target dimensions. The two parts are described separately below.

2.1 Features

We adopted the audio and video features that were extracted by Schuller *et al.* [10] from the SEMAINE corpus [5]. The audio features comprise 25 features that are related to the overall energy and the spectrum of the audio signal (*e.g.*, energy per band, entropy, and harmonicity) and 9 features that are voice-related (*e.g.*, F0, jitter, and harmonics-to-noise ratio). The video features were obtained by aligning the face images with the help of a standard Viola-Jones face detector, and densely extracting 8-bit local binary patterns (LBPs; [6]) from the aligned face images in 10×10 pixel blocks. For the WLSC challenge, word timings were measured and the audio and visual features were averaged over the full temporal span of the word [10].

Because previous studies in, *e.g.*, automatic action-unit labeling have found that changes of features over time may be very informative in social signal processing [12], we also compute temporal differences between features. Specifically, on the FCSC challenge, we compute the differences between a feature and its corresponding feature 100 frames earlier (*i.e.* using a differencing window of 4 seconds). On the WLSC challenge, we compute the differences between features and the corresponding features measured during the previous word. In preliminary experiments, we found these window sizes to produce good results.

2.2 Regression

As our regressor of choice, we used a simple l_2-regularized linear least-squares regression. Given a training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}$ represents a feature vector and y the corresponding target, a linear least-squares regressors minimizes the following regularized least-squares criterion:

$$\mathcal{L}(\mathbf{w}; \mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n - y_n \right) + \lambda \|\mathbf{w}\|^2, \qquad (1)$$

where $\mathbf{w}$ represents a linear combination of the input features, and λ represents a regularization parameter that needs to be set by the user. It is well-known that the weights $\mathbf{w}^*$ that minimize the regularized quadratic loss are given by:

$$\mathbf{w}^* = \left(\lambda \mathbf{I}_D + \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] \right)^{-1} \mathbb{E}[\mathbf{x}y] \qquad (2)$$

$$= \left(\lambda \mathbf{I}_D + \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right)^{-1} \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n y_n \right), \qquad (3)$$

where $\mathbf{I}_D$ represents the $D \times D$ identity matrix. The key advantage of training a linear least-squares regressor is that it scales linearly in N and that it does not require the full data set to be read into memory; the expectations $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ and $\mathbb{E}[\mathbf{x}y]$ under the data distribution can be computed on-the-fly using a single sweep through the data (*i.e.* requiring only one full read of all the data from disk). This gives us an important advantage over support vector regressors and Gaussian process regressors, which scale quadratically in N, and over gradient-descent learners that require multiple sweeps over the training data. This makes it possible to train a linear least-squares regressor in about 10 minutes on a simple laptop computer (most of which is spent on reading data from disk).

We did notice that there are correlations between some of the target dimensions, *e.g.*, between arousal and valence. We did not exploit such correlations in our predictors, but it is straightforward (and may be helpful) to do so: a simple whitening transform may be applied on the target dimensions. After predicting the whitened targets, the reverse whitening transform needs to be applied to obtain the final prediction. Indeed, it is also possible to incorporate (Gaussian) temporal smoothness constraints on the targets in the quadratic loss, which leads to standard linear dynamical system. We did not use such a Kalman smoother in our study, because the rating sequences are not modeled well by Gaussian dynamics. Instead, the target sequences appear to have been generated by a switching distribution: the targets are approximately constant most of the time, but sporadically, the target values change drastically.

Since the evaluation criterion of the competition is the correlation between the predicted and the true ratings, it is not important that one accurately predicts the actual ratings. Because the target values only sporadically change drastically, it is much more important to accurately predict when a rating is rapidly increasing, when it is rapidly decreasing, or when it is roughly keeping the same value. A simple approach to achieve this is to train the regressors on the derivative of the rating signal instead of on the rating signal itself. In our experiments, we estimate the derivative of the rating signal using the first-order differences of the signal. (Indeed, it may have been helpful to use a smoother derivative estimate that is obtained by, *e.g.*, applying a derivative-of-Gaussian filter.)

On the WLSC challenge, we use regularized linear least squared to train a single regressors that predicts the affective dimensions from the average features for a word (the averaging is over the temporal span of the word). On the FCSC challenge, we follow [10] and use the word timings to train two separate regressors: one regressor for the parts of the video that do contain speech, and one regressor for the parts of the video that do not contain speech.

3. EXPERIMENTS

To evaluate the performance of our approach, we performed experiments on the AVEC 2012 data set. The setup of these experiments is described in 3.1. The results of the experiments are presented in 3.2.

3.1 Experimental setup

In our experiments, we used the pre-processed version of the AVEC 2012 data set as described in [10]. The data is annotated per frame by human raters for arousal, expectancy, power, and valence; the ratings are continuous-valued. The data is subdivided into a of 31 training, 32 validation, and 32 test videos. In total, the data contains $1,358,123$ frames (comprising 7.5 hours of video). We used the fixed division of the data in all our experiments. In the paper, we present only results obtained by (1) training on the training data and testing on the development data and (2) training on the combined training and development data and testing on the test data (in which case the organizers measured the performance of our predictors).

Features	Ratings	Arousal	Expect.	Power	Valence	Mean
Normal	*Normal*	0.1550	**0.1599**	0.1193	0.2047	0.1597
Normal	*Derivative*	**0.2733**	0.0695	0.4193	0.1199	0.2205
Derivative	*Normal*	0.0936	0.0914	0.1477	0.0152	0.0870
Derivative	*Derivative*	0.2686	0.0864	**0.4216**	0.1434	0.2300
Baseline system [10]		0.181	0.148	0.084	**0.215**	0.1570
Our best system		**0.2733**	**0.1599**	**0.4216**	0.2047	**0.2649**

Table 1: Correlations between predicted and ground-truth ratings on the AVEC 2012 development set for the FCSC competition (higher is better; best performance is boldfaced).

Features	Ratings	Arousal	Expect.	Power	Valence	Mean
Normal	*Normal*	0.0702	0.1060	0.0219	**0.1833**	0.0953
Normal	*Derivative*	**0.2463**	**0.2662**	0.3501	0.0497	0.2281
Derivative	*Normal*	0.2366	0.2642	0.3477	0.0440	0.2231
Derivative	*Derivative*	0.2407	0.0481	**0.4088**	0.1030	0.2001
Baseline system [10]		0.018	0.009	0.001	0.002	0.0075
Our best system		**0.2463**	**0.2662**	**0.4088**	**0.1833**	**0.2762**

Table 2: Correlations between predicted and ground-truth ratings on the AVEC 2012 development set for the WLSC competition (higher is better; best performance is boldfaced).

We use cross-validation to determine the optimal regularization parameter λ. In all experiments, we allowed our linear predictors to predict an offset by adding a constant feature with a value of 1 to the data; the offset was not regularized.

The quality of the predicted targets is measured via its Pearson correlation with the ground-truth targets. This means that the exact value of the targets does not need to be correct, but that the *direction* of the predicted target signal needs to be correct over time: high correlations are achieved when the predicted target signal goes up if the ground-truth target signal goes up, and the predicted target signal goes down if the ground-truth target signal goes down. (This makes the performance measure invariant under rating offset variations between movies.) The correlations are computed per movie, and subsequently, averaged over all movies in the test data to obtain the final quality measure.

3.2 Results

Table 1 and 2 provide an overview of the results on the development set, obtained using regressors that were trained on the training set. The results are presented separately for the FCSC and WLSC sub-challenges. The results presented in the tables reveal that predicting the derivative of ratings may be beneficial for the prediction of arousal and power, which suggests that human perception of these affective dimensions is for a large part relative to earlier observations of the same individual. For expectancy and valence, the situation is much less clear: these two affective dimensions are much harder to predict, which is reflected in lower average scores and higher variances in the scores between experiments. On the development set, the best mean correlations we obtain are 0.2649 for the FCSC sub-challenge and 0.2762 for the WLSC sub-challenge, which is quite a bit better than the performances of the baseline system (which are 0.157 and 0.039, respectively [10]).

Table 3 and 4 present the results on the AVEC 2012 test set, obtained using regressors that were trained on the combined training and development set (the performance on the test set was measured by the AVEC 2012 competition organizers). Again, the results are presented separately for the FCSC and WLSC sub-challenges. The results on the test set are in line with those on the development set: predicting arousal and power is relatively easy compared to predicting expectancy and valence. In addition, it also appears that predicting rating derivatives is more effective for the former two affective dimensions, although the results are less pronounced than on the development set. For valence, it appears to be best to predict the actual ratings using the standard features, whereas for expectancy, the results vary inexplicably between the two sub-challenges. The best average performances we obtained are 0.2255 for the FCSC sub-challenge and 0.1921 for the WLSC sub-challenge. Whilst these performances are noticeably lower than the results on the development set – despite the fact that they were obtained using regressors that were trained on twice as much data – the performances are still better than the performances of the baseline system (which are 0.112 and 0.027, respectively [10]).

4. DISCUSSION

A curious observation we made is that – when we predict rating derivatives – cross-validating over λ frequently leads to the selection of *extremely* high values of λ (in both competitions). In particular, the value of λ is frequently chosen to be so high that all feature weights are effectively pushed to zero. In this way, the predictor essentially ignores the observations altogether and only uses the offset to make the predictions. The learning then exploits a bias in the data, viz. that ratings are more likely to go up over time than down. Since the offset is in the rating-derivative domain, it can be used to predict a linearly increasing rating. (If we had included a time stamp as a feature with the normal features, we could have obtained a similar effect.). Such a "constant" predictor actually already outperforms the baseline system on the development set.

Features	Ratings	Arousal	Expect.	Power	Valence	Mean
Normal	*Normal*	0.1524	**0.1659**	0.1278	**0.1453**	0.1479
Normal	*Derivative*	0.2809	0.0341	0.2680	0.0517	0.1578
Derivative	*Normal*	0.2056	0.1095	0.1669	0.0684	0.1376
Derivative	*Derivative*	**0.2990**	0.0798	**0.2918**	0.0409	0.1779
Baseline system [10]		0.141	0.101	0.072	0.136	0.1125
Our best system		**0.2990**	**0.1659**	**0.2918**	**0.1453**	**0.2255**

Table 3: Correlations between predicted and ground-truth ratings on the AVEC 2012 test set for the FCSC competition (higher is better; best performance is boldfaced).

Features	Ratings	Arousal	Expect.	Power	Valence	Mean
Normal	*Normal*	0.0861	0.1269	0.0817	**0.1375**	0.1080
Normal	*Derivative*	0.1281	0.0528	0.2123	0.0295	0.1057
Derivative	*Normal*	**0.1669**	0.1269	0.2160	0.0096	0.1299
Derivative	*Derivative*	0.1452	**0.2414**	**0.2225**	0.0255	0.1587
Baseline system [10]		0.021	0.028	0.009	0.004	0.0155
Our best system		**0.1669**	**0.2414**	**0.2225**	**0.1375**	**0.1921**

Table 4: Correlations between predicted and ground-truth ratings on the AVEC 2012 test set for the WLSC competition (higher is better; best performance is boldfaced).

This observation, together with the observations that (1) prediction performances for expectancy and valence are very low and unstable and (2) more complex regression models such as Gaussian process regressors did not produce performance improvements in our preliminary experiments, suggests that there is very little information in the data that is relevant to the prediction of affective dimensions. Next to extracting better features, it thus seems like a sensible approach to combine the results of different approaches via simple ensemble methods such as linear blending to maximally exploit the little bit of learning signal in the data. (Similar approaches have produced good results in other competitions in which the learning signal was very weak, such as the Netflix and Heritage Health Prize competitions.)

5. ACKNOWLEDGMENTS

The author thanks Michel Valstar and David Tax for helpful discussions. Laurens van der Maaten is supported by the EU-FP7 NoE on Social Signal Processing (SSPNet).

6. REFERENCES

[1] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Patt. Recogn.*, 36(1):259–275, 2003.

[2] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1940–1954, 2010.

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the Int. Conf. on Machine Learning*, pages 282–289, 2001.

[4] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[5] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. on Affective Computing*, 3:5–17, 2012.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[7] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10), 2010.

[8] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[9] H. Rauch, F. Tung, and C. Striebel. Maximum likelihood estimates of linear dynamical systems. *AIAA Journal*, 3:1445–1450, 1965.

[10] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. Avec 2012 – the continuous audio/visual emotion challenge. In *Proceedings of the Second International Audio/Visual Emotion Challenge and Workshop*, 2012.

[11] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011 – the first international audio/visual emotion challenge. In *Proc. of the First AVEC Workshop*, 2011.

[12] L. van der Maaten and E. Hendriks. Action unit classification using active appearance models and conditional random fields. *Cognitive Processing*, 2012.

[13] L. van der Maaten, M. Welling, and L. Saul. Hidden-unit conditional random fields. In *JMLR W&CP 15*, pages 479–488, 2011.

[14] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27:1743–1759, 2009.