BIOLOGICALLY VS. LOGIC INSPIRED ENCODING OF FACIAL ACTIONS AND EMOTIONS IN VIDEO

M.F. Valstar and M. Pantic

Computing Department, Imperial College London, UK
{M.F.Valstar, M.Pantic}@imperial.ic.ac.uk

Abstract

Automatic facial expression analysis is an important aspect of Human Machine Interaction as the face is an important communicative medium. We use our face to signal interest, disagreement, intentions or mood through subtle facial motions and expressions. Work on automatic facial expression analysis can roughly be divided into the recognition of prototypic facial expressions such as the six basic emotional states and the recognition of atomic facial muscle actions (Action Units, AUs). Detection of AUs rather than emotions makes facial expression detection independent of culture-dependent interpretation, reduces the dimensonality of the problem and reduces the amount of training data required. Classic psychological studies suggest that humans consciously map AUs onto the basic emotion categories using a finite number of rules. On the other hand, recent studies suggest that humans recognize emotions unconsciously with a process that is perhaps best modeled by artificial neural networks (ANNs). This paper investigates these two claims. A comparison is made between detection of emotions directly from features vs a two-step approach where we first detect AUs and use the AUs as input to either a rulebase or an ANN to recognize emotions. The results suggest that the two-step approach is possible with a small loss of accuracy and that biologically inspired classification techniques outperfrom those that approach the classification problem from a logical perspective, suggesting that biologically inspired classifiers are more suitable for computer-based analysis of facial behaviour than logic inspired methods.

1. INTRODUCTION

The ability to detect and understand facial expressions and other social signals of someone with whom we are communicating is the core of social and emotional intelligence. Human Machine Interaction systems capable of sensing stress, inattention and heedfulness and are able to adapt and respond to these affective states of users are likely to be perceived as more natural, efficacious and trustworthy.

But what exactly is an affective state? Traditionally the terms “affect” and “emotion” have been used synonymously. Following Darwin, discrete emotion theorists propose the existence of six or more basic emotions that are universally displayed and recognized [8]. These include emotions such as happiness, anger, sadness, surprise, disgust and fear. Data from both modern Western and traditional societies suggest that non-verbal communicative signals (especially facial expressions) involved in these basic emotions are displayed and recognized cross-culturally [8]. However, in real life people show far more expressions, such as ’boredom’ or ’I don’t know’. There is much less evidence that these subtler expressions are universally displayed and interpreted as well.

Table 1. Rules for mapping Action Units to emotions, according to the FACS investigators guide. A||B means “either A or B”.

Emotion	AUs	Emotion	AUs
Happy	{12}	Fear	{1,2,4}
	{6,12}		{1,2,4,5,20, 25||26||27}
Sadness	{1,4}		
	{1,4,11||15}		{1,2,4,5,25||26||27}
	{1,4,15,17}		{1,2,4,5}
	{6,15}		{1,2,5,25||26||27}
	{11,17}		{5,20,25||26||27}
	{1}		{5,20}
Surprise	{1,2,5,26||27}		{20}
	{1,2,5}	Anger	{4,5,7,10,22,23,25||26}
	{1,2,26||27}		{4,5,7,10,23,25||26}
	{5,26||27}		{4,5,7,17,23||24}
Disgust	{9||10,17}		{4,5,7,23||24}
	{9||10,16,25||26}		{4,5||7}
	{9||10}		{17,24}

Instead of directly classifying facial expressions into a finite number of basic emotion classes, we could also try to recognize the underlying facial muscle activities and then interpret these in terms of arbitrary categories such as emotions, attitudes or moods [11]. The Facial Action Coding System (FACS) [4] is the best known and the most commonly used system developed for human observers to describe facial activity in terms of visually observable facial muscle actions (i.e., Action Units, AUs). Using FACS, human observers uniquely decompose a facial expression into one or more of in total 44 AUs that produced the expression in question.

Classic psychological studies like the EMFACS (emotional FACS), suggest that it is possible to map AUs onto the basic emotion categories using a finite number of rules (as suggested in the FACS investigators guide [4], table 1). This effectively suggests that facial expressions are decoded at a conscious level of awareness. Alternative studies, like the one on “the thin slices of behaviour” [1], suggest that human expressive nonverbal cues such as facial expressions are neither encoded nor decoded at an intentional, conscious level of awareness. In turn, this finding suggests that biologically inspired classification techniques like artificial neural networks (ANNs) may prove more suitable for tackling the problem of (basic) emotion recognition from AUs as such techniques emulate human unconscious problem solving processes in contrast to rule-based techniques, which are inspired by human conscious problem solving processes.

Recent work on emotion detection using biologically inspired algorithms has used ANNs [5], SVMs [2], Bayesian Networks [3, 16] and Hidden Markov Models (HMMs) [3]. Recent work on facial

AU detection using biologically inspired algorithms has used similar techniques: ANNs [13], SVMs [2, 14], and Bayesian Networks [16]. Recent work on AU and emotion detection that used algorithms inspired by human conscious problem solving processes includes rule-based systems [9], case-based reasoning [10], and latent semantic analysis [5]. For a survey of past work in the field, see [11].

The goal of this paper is a twofold. First we want to investigate whether a two-step approach to emotion recognition, where both the facial feature extraction and the AU recognition precede the emotion prediction, attains similar recognition rates as a single-step approach in which the recognition of emotions is conducted based directly upon the extracted facial features. Detection of AUs as a first step in facial expression detection has several advantages. AUs are independent from high level interpretations in terms of emotions or moods. They also cause a dimensionality reduction, as all expressions can be described using only 44 attributes, namely 44 AUs. In contrast to one-step expression detection, AU detectors can be trained independently of the facial expression shown. Hence, in order to train an AU detector the training data set does not need to contain examples of all 7000 frequently occurring facial expressions. On the other hand, the main reason to presuppose that a single-step approach could perform better is the error accumulation inherent in multiple-step approaches. The second goal is to investigate the suggestion implicitly made by recent alternative studies in psychology that biologically inspired classification techniques like ANNs are more suitable for tackling the problem of emotion recognition from AUs than logic inspired classifiers such as rule-based systems.

The organisation of the paper is as follows: section 2 describes the methodology used to investigate the research issues. In section 3 we present and discuss the results of the conducted experiments. Final remarks and recommendations for further research conclude the paper.

2. METHODOLOGY

2.1. Facial Point Tracking

To track 20 facial characteristic points illustrated in Fig. 1, in an input image sequence, we use the method proposed in [14]. The facial points are automatically detected in the first frame of the input image sequence using the method proposed in [15] that employs individual feature patch templates to detect points in specific facial regions such as the regions around the mouth corners, the eyebrows' corners, etc. These feature models are GentleBoost learned templates of Gabor wavelet features derived from 13x13 pixel image patches. After 20 fiducial points are localized in the first frame of the input face image sequence, windows positioned around each of the facial points define a number of color templates. We subsequently track each color template for the rest of the image sequence with the particle filtering with factorized likelihoods algorithm [12]. We use the same observation model as proposed in [14], which is both insensitive to variations in lighting and can cope with small deformations of the template. This polymorphy aspect is necessary as many areas around facial points change their appearance when a facial action occurs (i.e. the mouth corner when smiling). The particle filtering scheme results for every image sequence in a set of points P with dimensions $n * 20$, where n is the number of frames of the input image. For all points p_i, where $i = [1:20]$ denotes the facial point, we compute two features for every frame n:

$$\begin{aligned} f_1\left(\boldsymbol{p_i}\right) &= \boldsymbol{p_{i,y,n}} - p_{i,y,1} \\ f_2\left(\boldsymbol{p_i}\right) &= \boldsymbol{p_{i,x,n}} - p_{i,x,1} \end{aligned} \tag{1}$$

that correspond to the deviation of respectively the y and the x coordinate from the related coordinates at the first (expressionless) frame. Then, for all pairs of points $\boldsymbol{p_i}$, $\boldsymbol{p_j}$,$i \neq j$ we compute in each frame the features

$$\begin{aligned} f_3\left(\boldsymbol{p_i}, \boldsymbol{p_j}\right) &= \|\boldsymbol{p_i} - \boldsymbol{p_j}\| \\ f_4\left(\boldsymbol{p_i}, \boldsymbol{p_j}\right) &= f_3\left(\boldsymbol{p_i}, \boldsymbol{p_j}\right) - \|\boldsymbol{p_{i,1}} - \boldsymbol{p_{j,1}}\| \end{aligned} \tag{2}$$

where the norm in equation (2) is the L_2 norm. Finally, we compute the first time derivative df/dt of all features defined above, resulting in a set F_n of 840 features per frame.

2.2. One-step Emotion Recognition

Our one-step approach to emotion recognition from face image sequences is based on support vector machines (SVMs). SVMs are very suitable for the task in question because, in general, the high dimensionality of the feature space does not affect the training time, which depends only on the number of training examples. To solve our six emotion detection problem we used a one-versus-one multi-class SVM classifier.

We also experimented with training the SVM classifiers on the features selected by GentleBoost [6]. In feature selection by GentleBoost, each feature is treated as a weak classifier. GentleBoost selects the best of those classifiers and then boosts the weights using the training examples to weight the errors more. The next feature is selected as the one that gives the best performance on the errors of the previously selected features. At each step, it can be shown that the chosen feature is uncorrelated with the output of the previously selected features. In our study we use GentleBoost to rank the features in order of importance, based on the strong classifier outputs. Then we use SVMs in a cross validation routine to determine the optimal number of features to use. On average, 7 of a total of 840 features were picked as the most informative features in one-vs-one emotion classification. As shown in [2], when SVMs are trained using the features selected by a boosting algorithm, they perform better. In our case, the average recall increase is 18.9% (see Table 4). This is mainly due to an imbalance between the number of features and positive examples present in our dataset (840 features vs. an average of 26 positive samples per class).

2.3. Two-step emotion recognition

In the two-step approach to emotion recognition from face image sequences, both the facial feature extraction (and selection) and the AU recognition precede the emotion prediction. To detect 15 different AUs occurring alone or in combination in an input image sequence, we used 15 separate SVMs to perform binary classification using one-versus-all partitioning of data resulting from the feature extraction and selection stages described in sections 2.1 and 2.2. The choice of 15 AU categories (Table 2) has been influenced by both the AUs that can be encoded based upon the utilized features defined in section 2.1 and the components of expressions (i.e., micro-events) that seem to be hardwired to emotions (Table 1, [4]).

To perform emotion classification based on AU predictions (the second stage of our two-step emotion recognition approach), we experimented with both the biology and the logic inspired classification engines. The logic inspired recognition engine is a rule-based system that maps the 15 AUs onto the 7 emotion categories. The utilized rules are the EMFACS rules suggested by Ekman and colleagues in the FACS investigators guide (Table 1, [4]). The biologically inspired recognition engine that we have experimented with was an ANN with 3 hidden layers, each of which had 27 nodes, and

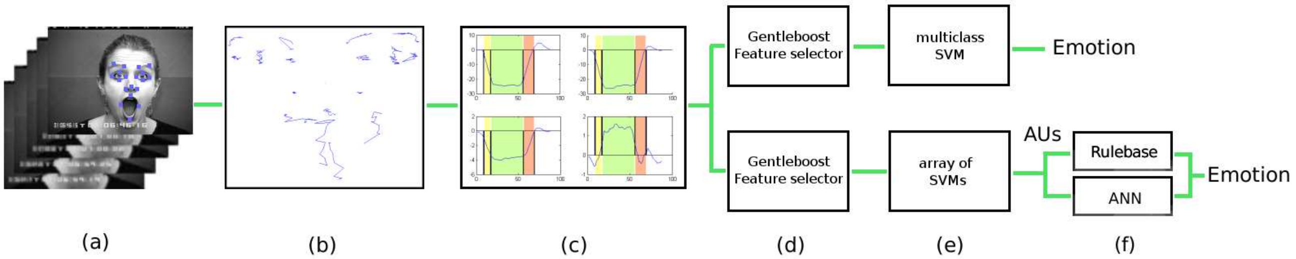

Fig. 1. Overview of the automatic AU and emotion detection system: (a) input image sequence with tracked facial points, (b) tracking results, (c) the four most important features for recognition of AU2 shown over time, (d) GentleBoost is used to select the most important features, (e) which are subsequently fed to (mc-)SVMs, (f) for two step approach: emotion detection from AUs by Artificial Neural Networks or a rulebase.

Table 2. Classification results for automatic AU detection. *clr* is the classification rate.

AU	clr	AU	clr	AU	clr	AU	clr
1	0.882	6	0.954	12	0.837	25	0.902
2	0.935	7	0.771	15	0.882	26	0.876
4	0.863	9	0.948	20	0.902	27	0.908
5	0.863	10	0.765	24	0.867		

one output layer containing 6 nodes, one for every emotion. All neurons used the log sigmoid evaluation function.

3. EXPERIMENTS

For this study we have used data from the most commonly used Cohn-Kanade database [7]. This FACS-coded database consists of recordings of subjects who display one of the six basic emotions on command. The part of the database that is available from the authors upon request consists of a total of 487 recordings of 97 subjects. From this database, we selected 153 image sequences of 66 subjects. Image sequences were included in this validation set if two experts decided by consensus on the basic emotion displayed in the samples in question.

The first goal of our study is to investigate the performance of a two-step approach in which AUs are detected automatically in the first step and the complex expressions, in this case six basic emotions, in the second step. We compare this two-step approach with the one-step emotion detection where we feed the features directly into a multiclass SVM to detect emotions. Both two step approaches use binary SVMs to detect AUs first. In the second step we use either an ANN or a rulebase to map the AUs to emotions. This second step in the two-step approach also gives us the data needed for the second goal of our study: evaluating whether the classic or alternative psychological studies made a correct assumption, i.e., whether biologically inspired techniques indeed outperform logic inspired ones.

For both the one-step emotion recognition approach and the automatic AU detection method we apply a leave-one-subject-out cross validation scheme. Within every validation fold, we apply an inner cross validation loop, randomly splitting the training data in two sets which we use for feature selection and kernel parameter optimisation. With this approach, we ensure in the outer cross validation loop that our results are person-independent and by means of the inner cross validation loop we avoid over-fitting of the classifier to our data. We train our SVMs using a radial basis frequency kernel $k(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right)$. Thus, the two parameters to optimize are the kernel width σ and the penalty parameter C. Results for the one-step emotion detection and AU detection are shown in table 3 part B and table 2 respectively. The ANN used to detect emotions from AUs in the two-step approach was evaluated using a standard leave-one-subject-out cross validation. The rulebase used to detect emotions from AUs in the two-step approach was directly applied to the output of the automatic AU detector as the rules are fixed and do not need any training. Part C and D of table 3 show the two-step approach results for the logic inspired rulebase and the biologically inspired ANN, respectively. To test the goodness of our rulebase, we also applied the rulebase on manually coded AUs.

The two-step approach performs worse than the one-step approach. However, the difference using ANNs is not that big. The interpretation free description of expressions in terms of AUs and the decreased dimensionality of the classification problem may be considered more valuable then the small increase in accuracy.

Table 3 part C shows that results for the logic inspired approach deteriorate significantly when the rules are applied on automatically detected AUs instead of manually labeled AUs. Obviously, the rulebase is a very rigid, nonadaptive system that is sensitive to noise in the input. The fact that the rule based classifier cannot learn and has no means to compensate for known weaknesses in the AU detector (such as the bad classification of AU15 which influences sadness recall) contributes to a performance decrease. This is not the case for the biologically inspired ANN, as part D of table 3 clearly shows. This suggests that the alternative studies in psychology offer a model for high-level facial behaviour interpretation that is more suitable for computer-based analysis than the model offered by classic studies.

4. CONCLUSION

This paper shows that a two-step approach to emotion recognition, where the AU recognition precede the emotion prediction, attains recognition rates similar to those of a single-step approach in which the recognition of emotions is conducted directly from the extracted facial features. We like to stress at this point the benefits of the two step approach. The most important aspect is that an AU description of a facial expression is absolutely interpreta-

Table 3. Results of emotion recognition, *clr* is classification rate, *rec* is recall and *pr* is precision: A) classification of manually labeled AUs to emotions by rules, B) classification of features to emotions by a multi-class SVM, C) automatically detected AUs classified to emotions by rules, D) Neural Networks classifying automatically detected AUs into emotions E) one-step emotion classification without feature selection.

	A			**B**			**C**			**D**			**E**		
Emotion	**clr**	**rec**	**pr**	**clr**	**rec**	**pr**	**clr**	**rec**	**pr**	**clr**	**rec**	**pr**	**clr**	**rec**	**pr**
Anger	0.967	0.714	0.909	0.948	0.643	0.750	0.889	0.214	0.333	0.915	0.500	0.539	0.923	0.549	0.608
Disgust	0.980	0.960	0.923	0.948	0.920	0.793	0.856	0.920	0.535	0.935	0.760	0.826	0.926	0598	0.652
Fear	0.980	0.960	0.923	0.922	0.760	0.760	0.889	0.680	0.654	0.895	0.720	0.667	0.901	0.495	0.662
Happy	1.00	1.00	1.00	0.967	0.972	0.897	0.941	0.917	0.846	0.948	0.917	0.868	0.907	0.757	0.822
Sadness	0.980	0.952	0.909	0.935	0.619	0.867	0.882	0.191	0.800	0.889	0.571	0.600	0.903	0.307	0.613
Surprise	1.00	1.00	1.00	0.967	0.938	0.909	0.941	0.844	0.871	0.974	0.938	0.938	0.966	0.829	0.938
Average	**0.985**	**0.931**	**0.944**	**0.948**	**0.809**	**0.829**	**0.900**	**0.626**	**0.673**	**0.926**	**0.734**	**0.740**	**0.903**	**0.620**	**0.681**

tion free as it encodes muscle activations. Secondly, it is impossible to collect training data to train a system that is able to detect each of the 7000 different facial expressions. This makes the one-step approach impractical in real world applications. Although tables 2 and 3 clearly show the error accumulation often encountered in two-step approaches, this sacrifice is well worth the dimensionality reduction, reduction of required training data and enhanced abstraction obtained by means of the two-step approach. Concerning the discussion of logic vs. biologically inspired decision making algorithms, we have shown that in a two-step emotion recognition approach logic inspired methods are outperformed by biologically inspired methods. In our case the difference is 10% in recall and 6% in precision. This suggests that alternative studies in psychology offer a model for high-level facial behaviour interpretation that is more suitable for computer-based analysis than the model offered by classic studies.

Acknowledgements

The authors would like to thank Jeffrey Cohn of the University of Pittsburgh for providing the Cohn-Kanade database. The work of M.F. Valstar has been supported by the Netherlands BSIK-MultimediaN-N2 Interaction project. The work of M. Pantic has been supported by the Netherlands Organization for Scientific Research (NWO) Grant EW-639.021.202. The work has been conducted while the authors were at Delft University of Technology.

5. REFERENCES

[1] N. Ambady and R. Rosenthal, "Thin slices of behavior as predictors of interpersonal consequences: A meta-analysis", *Psychological Bulletin,* vol. 11, pp. 256-274, 1992

[2] M.S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and actions", *in SMC'04,* vol. 1, pp. 592-597, 2004

[3] I. Cohen, N. Sebe, A. Garg, L.S. Chen and T.S Huang "Facial expression recognition from video sequences - temporal and static modeling", *J. CVIU,* vol. 91, pp. 160-187, 2003

[4] P. Ekman, W.V. Friesen and J.C. Hager, "The Facial Action Coding System: A Technique for the Measurement of Facial Movement", San Francisco: Consulting Psychologist, 2002

[5] B. Fasel, F. Monay and D. Gatica-Perez, "Latent semantic analysis of facial action codes for automatic facial expression recognition", *In MIR'04, pp. 181-188, 2004*

[6] J. Friedman, T. Hastie, and R. Tibshirani. "Additive logistic regression: a statistical view of boosting", *The Annals of Statistics*, vol. 28, no. 2, pp. 337-374, 2000

[7] T. Kanade, J. Cohn and Y. Tian, "Comprehensive database for facial expression recognition", *In FG'00,* pp. 46-53, 2000

[8] D. Keltner and P. Ekman, "Facial expression of emotion", *Handbook of emotions,* M. Lewis and J.M. Haviland-Jones, Eds. Guilford Press, New York, pp. 236-249, 2000

[9] M. Pantic and I. Patras, "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences", *IEEE T. SMC - B*, vol. 36, no. 2, pp. 433-449, 2006

[10] M. Pantic and L.J.M. Rothkrantz, "Case-based reasoning for user-profiled recognition of emotions from face images", *in ICME'04*, vol. 1, pp. 391-394, 2004

[11] M. Pantic and L.J.M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-computer Interaction", *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390, 2003

[12] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features", *In FG'04,* pp. 97-102, 2004

[13] Y. Tian, T. Kanade and J.F. Cohn, "Recognizing action units for facial expression analysis", *IEEE T. PAMI,* vol. 23, no. 2, pp. 97-115, 2001

[14] M.F. Valstar, I. Patras and M. Pantic, "Facial Action Unit Detection using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data", *In CVPR'05,* vol. 3, pp. 76-84, 2005

[15] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using gabor feature based boosted classifiers", *in SMC'05,* pp. 1692-1698, 2005

[16] Y. Zhang and Q. Ji. "Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences", *IEEE T. PAMI,* vol. 27, no. 5, pp. 699-714, 2005