2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

On the Evaluation of Emotion Expressing Robots

Ansgar Bittermann, Kolja Kühnlenz, and Martin Buss
Institute of Automatic Control Engineering (LSR)
Technische Universität München
D-80290 München, Germany
bittermann@tum.de, {kolja.kuehnlenz,m.buss}@ieee.org

Abstract— **Common works on emotion expressing robots are theoretically based on a dimensional (continuous) model of emotions. Nevertheless, performance tests, which are used to evaluate the emotion expressing robots, are based on categorical (discrete) models of emotions. In this paper the use of dimensional-based tests is suggested, e.g. semantic differential approaches like the Pleasure-Arousal-Dominance-Model. By deriving the test from the theory on which the design of an object is based, the validity of the test raises significantly. Major benefits are explicit guidelines for design improvement and the possible integration of arbitratry actuated expressive features for which no common framework as e.g. the Facial Action Coding System (FACS) exists. For illustration purposes, a comparative evaluation study of the robot EDDIE is conducted: one test is based on a categorical model and one test is based on a dimensional model of emotion. A third study based on a dimensional model demonstrates the evaluation of the influence of animallike features on the perceived emotion state.**

I. Introduction

Robot design is an important issue in sociable robotics. The design and control of elements relevant to expressing emotions have a significant impact on how the represented emotion of the robot is perceived by the human. Particularly, the controlled posture of these affective elements is an important aspect and a well investigated issue in the field of human nonverbal communication. A common framework is the Facial Action Coding System (FACS), a comprehensive system which can distinguish facial expressions by the configuration of muscular activations. FACS has become the standard tool in describing facial expressions [1][18]. Hereby, the facial muscles are mapped to 43 action units (e.g. the lip tightener or the brow lowerer). Every facially expressed emotion can be described in this way by a sum of different activated action units. For example fear is displayed by the action units 1,2,4, 20 and 26 [7]. Elements which are not human, i.e. animallike or freely designed, are not covered by FACS.

An issue of equal importance is the choice of concept for the evaluation of displayed facial expressions. A common means are categorical approaches in user studies where test participants may choose best fits from a set, e.g. joy, anger, etc., e.g. [2]. A significant shortcoming of categorical evaluation methods is the low test-theoretical validity and the lack of feedback for design improvements, e.g. guidelines for posture adjustments of expressive elements. In the field of emotion expressing robots the mapping of the three-dimensional affective space to FACS action units is the most common design tool in order to realize smooth continuous transitions between emotional states, e.g. [7]. Hereby, subjective ratings of human emotions are assumed to be transferable to a three-dimensional metric space called affective space. The underlying dimensions are valence, arousal and stance/ dominance) However, this approach is itself a dimensional approach. Thus, designs and evaluations are commonly done in different representations which is not correct by test-theoretical means. In test-theory a test always has to reflect the theoretical foundation of the method it is testing.

In this paper the use of dimensional approaches to evaluate expressive robots is proposed. Common tools are, e.g., semantic differential approaches like the PAD-model of Mehrabian [4][9]. For several reasons stated in this paper it is suggested to prefer the use of a dimensional evaluation model over a categorical model, because the validity of a test can be raised significantly, if it is derived from the same theory on which the design of an object is based. The use of dimensional approaches is a generic tool to integrate actuated expressive elements of arbitrary design in the display of emotional expressions. The evaluation is exemplarily conducted by two different user studies to evaluate the same facial expression robot head: one study uses the categorical approach, the other uses the dimensional approach. Furthermore, a third study shows the use of a dimensional model for evaluating animallike features in emotion expressing robots.

The paper is organized as follows: Section 2 introduces and discusses common dimensional measures exemplified in the field of emotional expressions; Section 3 presents the application of a semantic differential approach to the evaluation of a facial expression robot head; conclusions are given in Section 4.

II. Dimensional Evaluation Approaches

A. Introduction of Quality Measures for Tests

By definition, a test is a specific psychological experiment. The goal of this experiment is to obtain comparative judgements about different subjects and their attitudes, impressions or psychological and physiological variables [3]. The variables, measured in a subject, can be devided into two groups: latent and manifest variables. Manifest variables are easily observable like the height or the weight of a person. Latent variables like attitudes, feelings or personal traits are

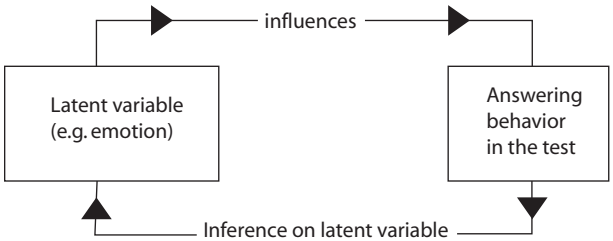

Fig. 1. Correlation between answering behavior in a test and latent variables.

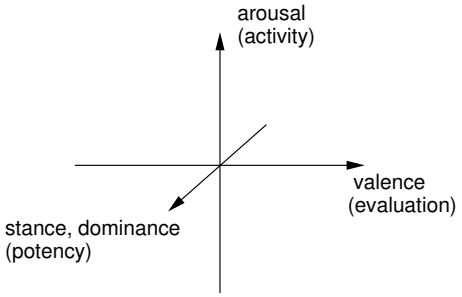

Fig. 2. Transformation from semantic to affective space.

not directly observable and, thus, have to be derived from the answering behavior in a test. The idea of using a test in order to obtain information of a latent variable is depicted in Figure 1.

Thereby, it is assumed that the latent variable influences the answering behavior in a test. After obtaining the answers, the latent variable is deduced from the observed answers of the test (mostly a questionnaire or an interview). This deduction is the most difficult part of the construction of a test since the correlation of the latent variables and the answering behavior in the test cannot be formulated easily. For that reason, it is very important that the deduction of the latent variable shall be highly embedded in a profound theory of how true feelings and attitudes of people can be obtained by using tests. In the following paragraph the most common and used theory of measuring feelings and attitudes is described.

B. The Semantic Differential

Osgood et al. [5] tried to connect scaled measurement of attitudes with the connotative meaning of words. In their classical experiment they discovered that the whole semantic space of words can be described by just three dimensions. The dimensions are evaluation (e.g. good vs. bad), potency (e.g. strong vs. weak), and activity (e.g. aroused vs. calm). The measurement technique they used is called semantic differential (approx. 20-30 bipolar adjectives on a seven-point Likert-scale). By using this method, one can plot a person's attitude in a semantic space or can compare different subjects' attitudes towards a product or object. Due to the wide range of usage, it became a standard tool in marketing, advertising, and attitude research. Applying this knowledge to the field of emotion research, [3] and [4] showed that also emotional adjectives can be reduced to a three-dimensional affective space with the dimensions valence, arousal, and dominance or stance, see Figure 2. Thereby, the three dimensions found by [5] can easily be transformed into the dimensions found by [4].

Valence can be interpreted as the evaluation dimension, arousal as the activity dimension, and potency can be referred to as the dominance or stance dimension. From his dimensional paradigm [4], Mehrabian developed a test system called Pleasure-Arousal-Dominance (PAD) emotion model [9]. The use of Mehrabian's PAD test to measure affective experiences represents a generic way to gather self-report-based user data regarding emotions. In the following paragraph the PAD test is described in more detail.

C. The PAD-Emotion Model by Mehrabian

The test is devided into three scales: a 16-item pleasure-displeasure scale, a 9-item arousal-nonarousal scale, and a 9-item dominance-submissiveness scale. The items of the PAD test are also in the format of a semantic differential. The duration of this test is approximately 7 minutes [10]. Alternatively, a short 12-item version exists. Each emotion expressing robot can, thus, be rated in 2-3 minutes. The reliability (internal consistency) of the full-length version is (α_c: Cronbach's alpha) $\alpha_c = 0.97$ for the pleasure scale, $\alpha_c = 0.89$ for the arousal scale, and $\alpha_c = 0.80$ for the dominance scale [4]. The internal consistency for the abbreviated version is $\alpha_c = 0.95$ for pleasure scale, $\alpha_c = 0.83$ for the arousal scale, and $\alpha_c = 0.78$ for the dominance scale [4].

As a result of the fact that the affective space [4] (with the dimensions valence, arousal and dominance) or the derived two-dimensional version of the circumplex model of emotion [11] are a common theoretical basis for building emotion expressing robots [4], only the PAD model has been introduced as a method to obtain data from affective experiences. There are several other tests measuring attitudes and feelings, e.g. the PANAS [12], which, however, are not as easily transferable into the theoretical model of emotion expressing robots.

D. Advantages of Dimensional Approaches as the Semantic Differential

There are generally two approaches in evaluating emotion expressing robots. On the one hand one can use dimensional approaches as the semantic differential and plot the values of the subjects in the affective space. On the other hand one can treat the emotions as quasi-independent categories, provide a list of possible emotions, and ask the subjects to mark the emotion they mean to see in the face of the emotion expressing robot, e.g. [2].

Dimensional approaches like the PAD-Model [4] are proposed as a generic tool for categorically or dimensionally designed expression robots due to several advantages:

1) Test-Theory: A test is conducted to examine the performance of the emotion expressing robot. Thus, the test also examines the theory in which the robot is embedded.

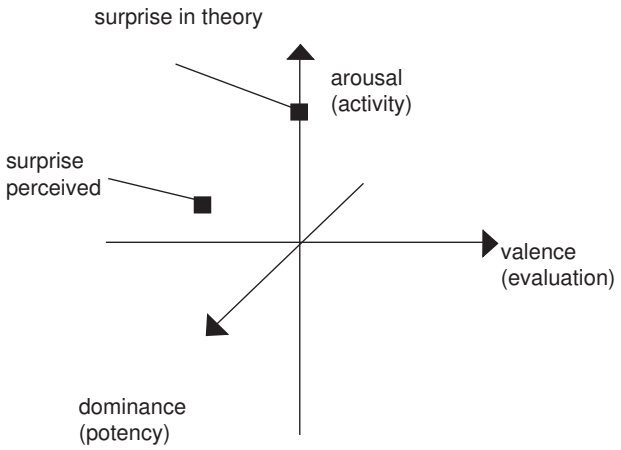

Fig. 3. Plotted results of the semantic differential.

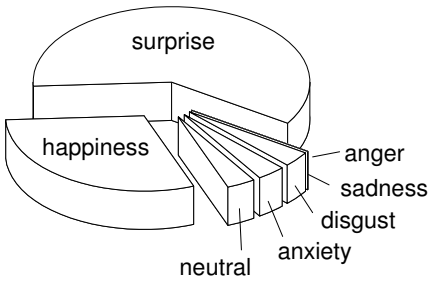

Fig. 4. Results of a test consisting of a list of emotions.

To do so, the test also has to be derived from the same theoretical framework as the robot. This means that if one embeds the robot in the framework of the circumplex model of Russel [11] or the affective space of Mehrabian [4] then the test has to be embedded in the same framework. Since many works use the three-dimensional affective space as a starting point for the development of the emotion expressing robot, the PAD-model fits well in the same theory.

2) Test-Construction: As discussed above, there has to be a found description of the conjunction between latent variable and answering behavior in the test. Since the PAD-model is based on the well proven model of the semantic space [5], it meets these premises.

3) Guidelines for Improvement: The result of an evaluation experiment of an emotion expressing robot should provide concrete instructions how the display can be improved. If the robot is theoretically embedded in a three-dimensional space, the semantic differential provides well interpretable data of the quality of the emotion expression on all three dimensions. Figure 3 shows an example with ficitve data for the perceived emotion 'surprise' and the theoretical values for 'surprise' in the affective space. Contrarily, if a list of categorical emotions is used, only data with 'correct hits' is gathered. Such a categorical test would only state that the emotion 'surprise' is identified correctly by a certain percentage. Furthermore, from these results it would be unclear which affective dimension of the displayed emotion is identified well or poorly.

Figure 4 shows the fictive result of a categorical evaluation method. It can be noted that about 50% of the participants evaluate the facial expression of the robot as surprise, around 30% perceive it as happiness and the rest as other emotions. To this point, no method exists to process the data to gain insight into new improvement guidelines from this evaluation. Furthermore, due to the fact that certain activation units are linked directly to certain areas of the affective space and, moreover, the same activation units are linked to certain artifical facial muscles of the emotion expressing robot, it is possible to conclude the contribution of the artifical facial muscle from the measured position in the affective space. Thus, by using the semantic differential more of the gathered information can be processed and interpreted.

4) Weighting Algorithm: In [2] it is argued that if one provides ten items in a multiple choice test the probability of chance was ten percent for each item to be picked and, thus, the expected chance probability of each emotion to be picked would be the same. However, this is in fact not true. Due to the fact that some emotions share more activation units than others [7] and lie closer together in the affective space, the possibility of each emotion to be picked has to be weighted. Thus, the expected possiblity of 10% would have to be increased for similar emotions and decreased for dissimilar emotions. Yet, an algorithm for weighting the expected percentage has not been developed. Such an algorithm, however, would be needed since these expected percentages are used in statistical tests to analyze whether a displayed emotion is classified by the subjects significantly correctly.

5) Reliability and Validity: The value of a test can be determined by quality measures of the classical test theory (reliability, validity, and objectivity). The PAD test manual provides these measures. Furthermore, it has been used in other studies, e.g. [13] [14], evaluated towards other tests which also test affective experiences [15]. Each test which is used should also provide at least data about its reliability and validity. Otherwise, one cannot judge its test-theoretical quality and should not use this test.

III. Application of the Guidelines – Three Evaluation Studies

In order to exemplarily show the advantages of differential evaluation approaches as the semantic differential

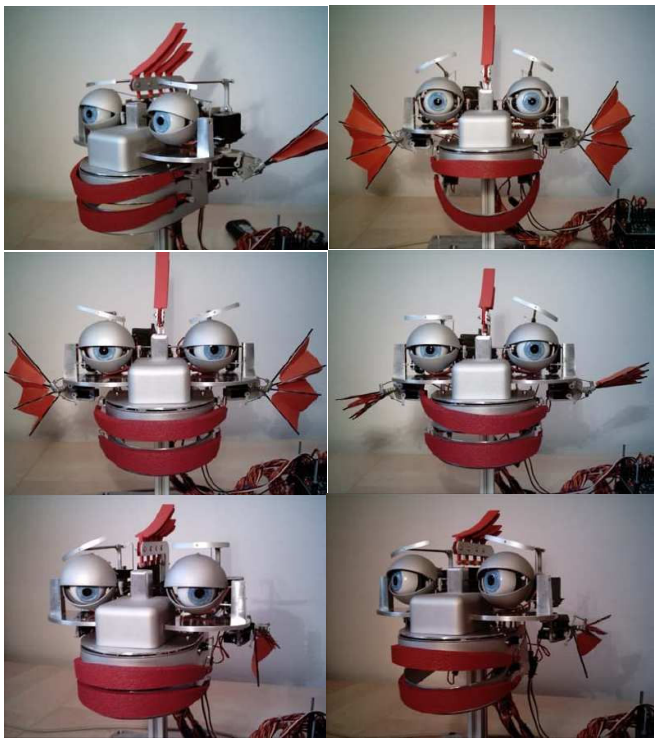

Fig. 5. EDDIE; displayed emotional expressions from top left to lower right: Joy, surprise, anger, disgust, sadness, fear.

for evaluation of expressive robots, two user studies have been conducted based on the mechatronical emotion-display EDDIE developed at the authors' lab [16]. In the following the system setup is presented in brief and the studies are described. The first study uses a test based on a dimensional model of emotions and evaluates the performance of the emotional facial expressions. The second study uses a test based on a categorical model of emotions to evaluate the quality of the emotional expressions of the same robot. The two tests were independently conducted with different participants and their results compared afterwards. The third study is concerned with the integration of arbitrary expressive features for which, yet, no generic design concept is known.

A. System Description

EDDIE is a robot head designed for displaying facial expressions, particularly, emotional ones realizing 13 of the 21 action units of FACS relevant to emotional expressions. The six basic emotions identified by Ekman and Friesen [18] are shown in Figure 5. Additionally, animallike features, the crown of a cockatoo and the ears of a dragon lizard with special folding mechanisms, are integrated.

EDDIE is encapsulated accepting commands from a higher-level decision and control unit via a serial communication protocol. The desired displayed emotion state can be transmitted based on the three-dimensional affective space model and feedback is given in affective and joint space representations. An embedded controller manages the transformation between affective space, action units, and joint space. The basic control architecture is shown in Figure 6.

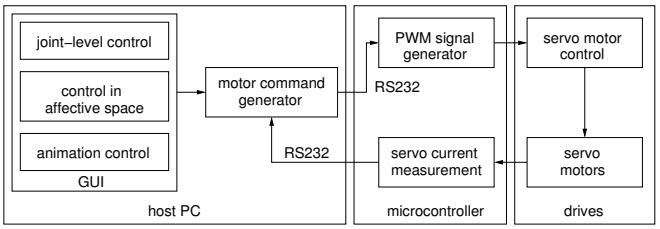

Fig. 6. Control architecture.

More details on design and control of EDDIE can be found in [16].

B. Categorical Evaluation Study

A study with 30 participants (researchers of Technische Universität München and the University of the Armed Forces, München, 11 female, 19 male, age-mean 30 years) has been conducted. Six basic emotions have been presented to each individual. The subjects' ratings of the displayed emotion were rated with a multiple-choice test with a seven-item-scale. The results of this study can be seen in Figure 8. On the abscissa the six displayed emotions are shown. For each displayed emotion the amount of assumed emotions by the subjects is presented. For example, 90% of the participants agree in seeing a sad face if a sad face is displayed. For the displayed emotion anger and anxiety about 50% of the answers were correct. The other 50% devide into other emotions. Evaluating these results, a scientist should be able to draw conclusions in order to improve the robot. Yet, no framework exists in order to formulate new guidelines for robot improvement considering the incorrect answers. Furthermore, it is questionable not to use the data of the incorrect answers as information would be disregarded.

C. Dimensional Evaluation Study

A study with 30 participants (students and researchers of the Ludwig-Maximilians-Universität, Munich, 15 female, 15 male, age-mean 25 years) has been conducted. A number of 30 different facial expressions corresponding to emotion states has been presented to each subject separately in random order. The subjects' impressions of the shown emotions have been acquired by using a German translation of the semantic differential of Mehrabian [17]. The results of the study are presented in Figure 7. The graph consists of the two dimensions valence and arousal (dimension of dominance is not displayed). Each emotion is displayed twice in the affective space: as expected in theory (ground truth) and as found in the study (measurements). The results of the study clearly show how each perceived emotion is empirically located in the affective space.

D. Discussion

From these results and the knowledge of the action units actually needed for each specific emotion [7] one can conclude the quality of the realization of each action unit in the mechatronical robot face. Additionally, steps for improvement can be derived from the results. For example,

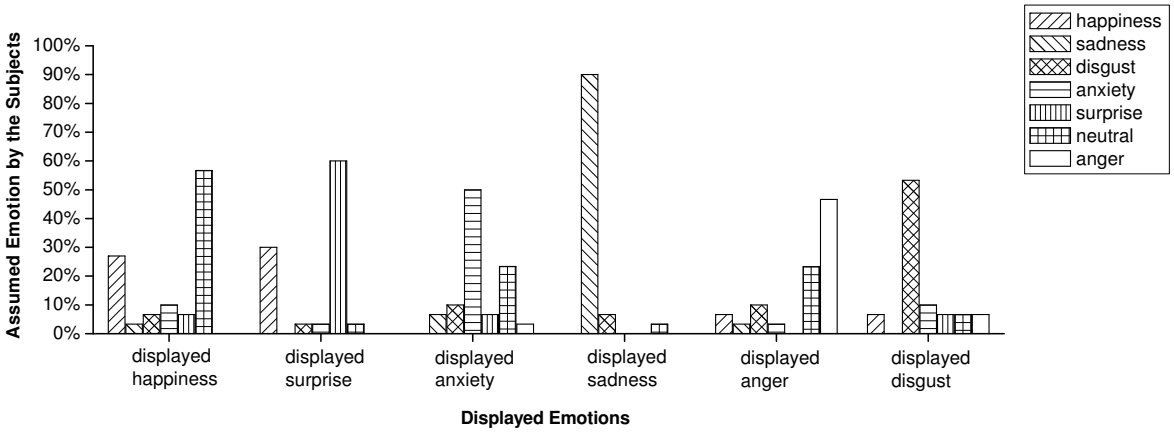

Fig. 8. Detailed results of the categorical evaluation study.

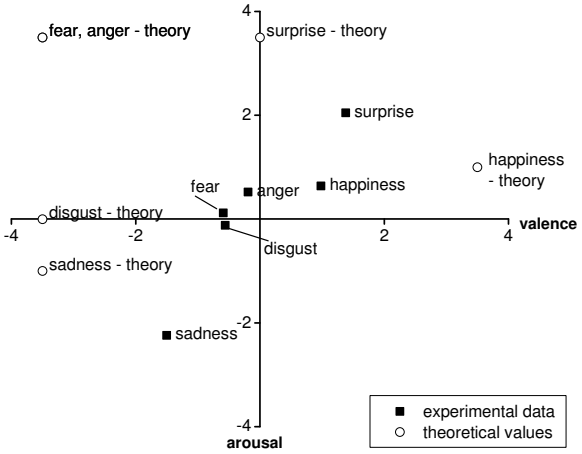

Fig. 7. Results from an evaluation study based on the dimensional semantic differential approach.

the results in Figure 8 show that the displayed emotion 'fear' has been perceived as a nearly neutral emotion with small values on each dimension. By analyzing this problem, one can compare the amount of action units needed for the intended displayed emotion and the amount of action units realized in the robot. Fear consists of action units 1, 2, 4, 20 and 26 [7]. In EDDIE the action units 1 and 2 are combined for technical reasons and cannot work separately. Furthermore, the action unit 4 was not implemented in the display. Yet, action unit 4 reflects the brow lowerer (musculus corrugatro supercilii) and is important for the emotion 'fear'. Furthermore, the range of action unit 20 (lip stretcher) could have been to small to show a motion which people expect from their experience with real human emotions. Based on these conclusions direct tasks of improvement can now easily be derived.

Not only single emotions but also the overall performance of the emotion expressing robot can be assessed. In Figure 8 a "positive valence shift" can be noted. By comparing the theoretical values and the actually found values of the displayed emotions, it is obvious that the displayed emotions are all shifted by one or two units to the positive side of the valence dimension. A possible reason for this shift can be noted in Figure 5. The lips of EDDIE are designed in such a way that it seems to smile in each displayed emotion state. Guidelines for improvement are, thus, clear from this point: the lips have to be redesigned. This example shows how even the summerized results can be used to provide new insight into the overall quality of the robot. This is a clear advantage of this method. Taking all the different aspects into consideration, it can be stated that dimensional evaluation methods as the semantic differential approach provide a powerful tool for evaluation of expressive robots by backprojection of joint space onto affective space via human perception.

E. Integration of Non-Humanlike Expressive Features

In a more general context the semantic differential approach is a generic means to evaluate the influence of actuated expressive elements of an arbitrary kind on the perceived emotion or intention. The knowledge gained from such evaluation procedures can then be used for the derivation of control commands for those expressive elements. Thereby, actuated expressive features of arbitrary design can be systematically controlled in order to intensify or attenuate the displayed intention or emotion in a particular selected dimension, e.g. valence, arousal, stance in case of the affective space.

This is exemplarily shown in an experimental pilot study with 30 participants (15 females, 15 males) evaluating the influence of two animallike features (crown of a cockatoo and ears of a dragon lizard). In a 2x2 ANOVA design with repeated measures (1. Factor crown, 2. Factor ears) it has been analyzed whether these two factors would shift the observed six basic emotions in the affective space. Each

TABLE I

RESULTS OF 2X2 ANOVA, REPEATED MEASURES

Emotion	Dimension	Factor	F-Value	p-value
1,2	V	crown	F(3,12)=4.013	0.034
1,2	V	ears * crown	F(9,36)=3.631	0.003
1,2	A	ears * crown	F(9,54)=3.258	0.003
3,4	V	ears * crown	F(9,18)=5.843	0.001
5,6	V	ears	F(3,6)=4.835	0.048
5,6	V	ears * crown	F(9,18)=4.132	0.005
5,6	A	ears	F(3,6)=67.582	0.000
5,6	A	crown	F(3,6)=11,987	0.006
5,6	D	ears	F(3,62)=46.724	0.000
5,6	D	ears * crown	F(9,18)=9,463	0.000

1=joy, 2=surprise, 3=fear, 4=sadness, 5=anger, 6=disgust
V=valence, A=arousal, D=dominance

factor has been realized in four conditions (from fully stilted to dismantled). All six basic emotions have been displayed with each combination of the two factors. Afterwards, the participants have rated each displayed emotion on the verbal semantic differential scale. Every subject participated in one third of the 96 possible combinations. All data has been tested with a Mauchly-test for sphericity. All values are >0.1. Thus, no Greenhouse-Geisser correction is necessary. Missing values are substituted by linear interpolation. Due to incorrect answering behavior some data has to be excluded. For that reason no F-Test could be calculated for 'joy, surprise'(dimension dominance) and 'fear, sadness' (dimensions arousal, dominance) [8]. The significant results of the ANOVA can be seen in Table I.

The results suggest that the ears and the crown may have an influence especially for the emotions joy, surprise, anger and disgust. As can be seen in Table I, mostly the interaction effect between crown and ears becomes significant. For the emotion anger and disgust the animallike features effect the evaluation of the subjects on all dimensions in the affective space. Surprisingly, these artifical features have a different effect on the evaluation depending on the emotion the robot is expressing. The results of the pilot study, thus, suggest further investigations of the usage of new non-humanlike features in emotion expressing robots.

It has, thus, successfully been shown that arbitrary expressive features can easily be integrated in a control concept of expressive robots using the semantic differential approach.

IV. CONCLUSIONS

In this paper the use of dimensional approaches for evaluations of expressive robots is proposed. It is suggested to prefer the use of a dimensional evaluation model over a categorical model. Main reason is the fact that a test shall be based on the same theory as the method it is testing resulting in significantly increased validity. Thus, dimensional approaches to evaluations of existing emotion expressing robots are suggested as the designs of most state-of-the-art emotion expressing robots are in fact based on dimensional models. Major benefits are guidelines for design improvement obtained from the evaluation results and the possible integration of arbitrary actuated expressive features for which no common framework as, e.g. the facial action coding system (FACS) exists. These benefits have been successfully demonstrated in three user studies evaluating the performance of EDDIE a robot head with facial expression capabilities and animallike expressive features. Dimensional approaches like the semantic differential approach provide a generic and reliable means for design and evaluation of humanlike as well as animallike or fictive robot characters.

V. ACKNOWLEDGMENTS

This work is supported in part within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also `www.cotesys.org`.

REFERENCES

[1] P. Ekman, Methodes for measuring facial action, in K.R. Scherer and P.Ekman, eds., *Handbook of methods in nonverbal behavior research*, Cambridge and New York, Cambridge University Press, UK, pp, 44–90, 1982b.

[2] C. Breazeal, Emotion and Sociable Humanoid Robots, *Int. Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 119–155, 2002.

[3] S. Ertl, Standardization of a semantic differential, *Zeitschrift für Experimentelle und Angewandte Psychologie*, vol. 12, no. 1, pp. 22–58, 1965.

[4] A. Mehrabian and J. A. Russell, An approach to environmental psychology, MIT Press, Cambridge, MA, USA, 1974.

[5] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, The measurement of meaning, Univ. Illinois Press, Oxford, UK, 1957.

[6] J. Rost, Lehrbuch Testtheorie, Testkonstruktion, 1. Aufl., Huber, Bern, Austria, 1996.

[7] K. Grammer and E. Oberzaucher, The reconstruction of Facial Expressions in Embodied Systems: New Approaches to an Old Problem, *ZIF Mitteilungen 2/2006*, pp. 14–31, 2006.

[8] W. Marx, Semantische Dimensionen des Wortfeldes der Gefühlsbegriffe, *Zeitschrift für Experimentelle Psychologie*, vol. 44, pp. 478–494, 1997.

[9] A. Mehrabian, Manual for a comprehensive system of measures of emotional states: The PAD Model, (Available from Albert Mehrabian, 1130 Alta Mesa Road, Monterey, CA, USA 93940), 1998.

[10] —, The PAD Comprehensive Emotion (Affect, Feeling) Tests, `http://www.kaaj.com/psych/scales/emotion.html`, 2006.

[11] J. Russell, Reading emotions from and into faces: resurrecting a dimensional-contextual perspective, in J. Russell and J. Fernandes-Dols, eds., *The Psychology of Facial Expression*, Cambridge University Press, Cambridge, UK, pp, 295–320, 1997.

[12] D. Watson, L. A. Clark, and A. Tellegen, Development and validation of brief measures of positive and negative affect: The PANAS scales, *Journal of Personality and Social Psychology*, vol. 54, pp. 1063–1070, 1988.

[13] P. Valdez and A. Mehrabian, Effects of color on emotions, *Journal of Experimental Psychology*, General, vol. 123, pp. 394–409, 1994.

[14] A. Mehrabian, C. Wihardja, and E. Ljunggren, Emotional correlates of preferences for situation-activity combinations in everyday life. *Genetic, Social, and General Psychology Monographs*, vol. 123, pp. 461–477, 1997.

[15] A. Mehrabian, Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression, *Journal of Psychopathology and Behavioral Assessment*, vol. 19, pp. 331–357, 1997.

[16] S. Sosnowski, A. Bittermann, K. Kühnlenz, and M. Buss, Design and Evaluation of Emotion-Display EDDIE, *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2006)*, to appear.

[17] A. O. Hamm and D. Vaitl, Emotional induction by visual cues, *Psychologische Rundschau*, vol. 44, pp. 143–161, 1993.

[18] P. Ekman and W. V. Friesen, Facial Action Coding Consulting, Psychologist Press, 1977.