Same or Different? Recollection Of or Empathizing With an Emotional Event From The Perspective of Appraisal Models

Gert-Jan de Vries   Paul Lemmens   Dirk Brokken
Philips Research Europe
High Tech Campus 34, NL-5656 AE Eindhoven
GJ.de.Vries@philips.com   Paul.Lemmens@philips.com   Dirk.Brokken@philips.com

Abstract

The same stimulus can evoke different emotions for different individuals. Incorporating personalized construal of stimuli is how appraisal models differ from dimensional models of emotion. Scherer formulated a model of the cognitive antecedents of emotions and analyzed recollections of events and emotions that his participants provided. In the present study, we were interested whether Scherer's appraisal model also applies to a situation in which participants have to empathize with a photo by building a story around the event depicted. Our results show that recollecting a personal emotional event and empathizing with a photo seem to be similar processes. Results of machine-learning techniques, however, show lacking performance of classifications to express discrete emotion labels in terms of appraisal data.

1. Introduction

In the present study, we approach emotions from a cognitive perspective. In this area of emotions research, emotion is in its briefest format usually defined as an affective reaction with a positive or negative connotation that occurs as a result of an appraisal, which is a personal interpretation of a stimulus [5]. Our work is done in the context of Scherer's component-process model (CPM, [8,9,10]). This paper reports our effort to investigate whether the process of recollecting a personal emotional event is similar (in terms of emotional outcome) to the process of empathizing with the story around a photo depicting a generic emotional event.

We conducted two experiments to gather sufficient data to compare our empathizing approach with Scherer's results of recollecting events [8]. We applied principal component analysis to study the structure of the data. We investigated whether one or more principal components mapped onto dimensions like valence, arousal, or dominance to study whether dimensional models [7] would be useful reductions in situations where less accuracy is required. Scherer et al. [10] already conducted such an experiment and in this paper we explored whether their analyses on our data resulted in similar conclusions.

Finally, we explored whether Scherer's 15 appraisal dimensions operated as predictors in a classification task. We used machine-learning techniques to study whether these techniques improved upon the approach used by Scherer [8] who used the averages per emotion as prototypes for classification of appraisal vectors to emotion. We considered classification using prototypes as well as non-prototype classifiers.

2. Appraisal theory

An appraisal (sometimes also called a construal) is an interpretation of the causes and consequences of a stimulus given the perceiver's personal goals, standards, and norms, etc. Because of the involvement of personalized cognitions like goals, such stimuli are often interpreted as being beneficial or a hindrance in achieving a goal. Thus, the element of valence, a positive or negative connotation of a stimulus, is introduced. This element is what distinguishes an emotional from a non-emotional response that can both result from the appraisal of a stimulus. Currently, two theories dominate the field: Scherer's Component-Process Model (CPM) [9] versus the OCC model of Ortony et al. [5]. The latter of the two is particularly focused on describing the concepts and their inter-relations that are needed to explain how the construal of stimuli results in an emotion.

The former model, Scherer's CPM, is more elaborate and describes the relation between stimulus and emotion on different levels of granularity. Particularly because the CPM model provides a detailed description of steps to establish a relation between a stimulus and an emotion, we chose that model to try and build a predictive computational model of the appraisal process.

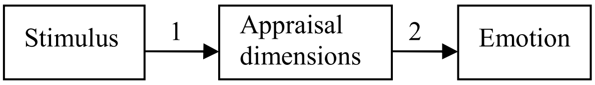


Figure 1: Appraisal model represented as a two-stage process.

We reduced the complex CPM to a two-stage strategy to cover the steps that are between the perception of a stimulus and the (overt) display of an emotion (see Figure 1). This simplification obviously removes many elements that are characteristic for the CPM but enables us to formulate a computational model.

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

The first stage builds up a representation of the stimulus in terms of the appraisal dimensions, as specified by the model. This stage results in the specification of what we call an appraisal vector which is a collection of values for each of the fifteen appraisal dimensions. This vector is used in the second stage to map the appraisal dimensions onto an emotion. The experiments described in this paper focus on the second stage. That is, given an appraisal, determine the corresponding emotion.

For the experiments, we used the full set of 15 appraisal dimensions comprising the appraisal objectives Relevance, Implications, Coping potential and Normative Significance. The emotions covered by this model are Enjoyment, Joy, Disgust, Contempt, Sadness, Desperation, Anxiety, Fear, Irritation, Anger, Indifference, Shame, Guilt, and Pride (see [8]).

3. Method

We carried out two experiments to gather sufficient data to apply machine-learning techniques. Most of the data was gathered via a web-based experiment, but an experiment in the lab served as a precursor for developing the web-based version.

3.1. Participants

In the lab experiment, data was collected from 33 voluntary participants. The data of one participant was left out because the debriefing revealed that this participant had not correctly interpreted multiple questions. Therefore the remaining data set contained data from 32 participants (18 were male; mean age 27 years). Each participant rated 5 pictures, yielding 160 appraisals of which one instance was disregarded because all questions were answered with “not applicable” (yielding 159 appraisals in total).

The web-based experiment resulted in the acquisition of 79 completed surveys (out of 436 invitations). Another 129 respondents completed the survey only partially. Often the incomplete sets of answers contained complete subsets corresponding to the appraisals of individual pictures. We compared these subsets with the appraisals from completed surveys and found that they were sufficiently similar to be included in the dataset. Therefore we selected all (picture-wise) complete appraisals with the restriction that not all answers were indicated as “not applicable”. This resulted in 261 appraisals, originating from 102 participants (74 male; overall age range 25-64 years, mean age 37) who appraised on average 2.6 pictures.

3.2. Stimuli

For each of the 14 emotions covered by the appraisal model we selected two pictures. Approximately half the pictures we used originate from a validated picture set of Overbeek et al. [6]. Because not all emotions covered by the appraisal model were present in this picture set, we added pictures found on the web, targeting at the remaining emotions of Anger, Anxiety, Contempt, Desperation, Indifference, Pride, Shame, and Guilt. Out of this set of 28 photos, for each subject a number of photos were chosen randomly (with the restriction that no photo should be chosen more than once per subject) in such a way that over all participants all photos were presented an approximately equal number of times.

3.3. Procedure

The experiment started with general demographic questions (i.e., gender and age). Subsequently, one of the five randomly selected photos was shown full screen for 6 seconds. For each photo, the participants were asked to imagine being in the event during which the photo was taken, to assume a role in that event, and to fill in an 18-item questionnaire. Questions 1 and 2 probed for the most prominent emotion in a photo and how strong it was experienced. The third question was to give a description of the assumed role of the participant and the situation s/he imagined. This question was added to help a participant stick to a single role while filling in the remaining 15 items of the questionnaire. These fifteen questions were those used by Scherer [8]. Note that our procedure is dissimilar from that employed by Scherer who asked only confirmation of a predicted emotion after a participant filled in the 15 questions regarding the appraisal dimensions. During the questionnaires a thumbnail of the corresponding photo was shown in the upper left corner of the screen. We used the answers to the first three questions to study the range of interpretations of (a particular) photo(s) and to verify that answers on the appraisal questions that followed could be expected.

In the lab experiment, a single participant was invited to take place in front of a pc located in a lab room, and got a short introduction, by the host, on what was to be expected. The host remained available for questions, but retreated to another part of the lab room such that the participant could fill in the questionnaire privately. In this experiment, each participant had to answer the 18 questions for five different, randomly chosen photos. The experiment took 25-30 minutes to complete. At the end of the experiment a debriefing was done.

With the exception of the number of photos shown (three in the web-based experiment versus five in the lab experiment), the procedure from the lab experiment was copied as much as possible to the web-based experiment. The instructions were presented on the screen and the debriefing was omitted. In total the survey took 15-20 minutes to complete. LimeSurvey, as a backend of the www.simplicitylabs.net environment, was used as environment to perform the survey in.

4. Results

Before we started with the analyses, we assessed whether the results of both experiments were sufficiently similar to be grouped into a single data set. We manually compared randomly selected samples of answers of each experiment for similarity. Then, we compared statistical data properties like means and standard deviations. Finally, several of the analyses described below were also done on each data set separately. These comparisons showed that the two data sets were highly similar and in fact similar to such an extent that they could be merged.

The analyses proceeded as follows. First, we compared the results from our study with the results obtained by Scherer [8]. Then we analyzed the structure of the data and the discriminative strength of appraisal in terms of different emotions by applying Principal Component Analysis (PCA). Finally, we applied the data in a classification task.

4.1. Comparison to Scherer [8]

With small exceptions in setup, our study was quite similar to Scherer's [8] work. We determined whether this is also reflected in the results by applying *t*-tests on the means and standard deviations from our experiment and those in [8] that were based on approximately 200 participants scoring the 15 appraisal dimensions based on recounting a personal emotional memory.

For each combination of emotion and appraisal dimension we computed the *t*-value from the means and standard deviations and from that the accompanying *p*-values, which are graphically represented in Figure 2. This figure shows that only in one case the *p*-value approaches a significance level of 0.05.

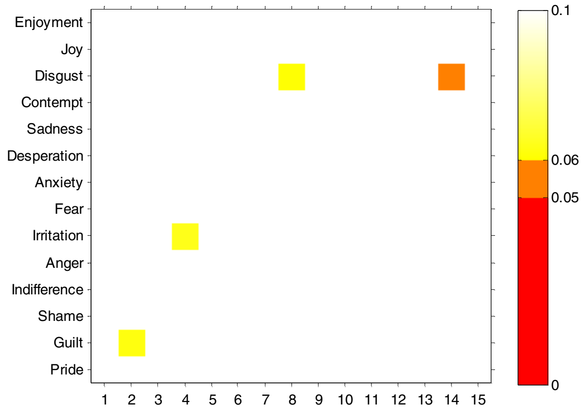


Figure 2: Significance matrix, showing the p-values of independent *t*-tests, for all combinations of emotions (vertically) and appraisal dimensions (horizontally). Red squares indicate significant *t*-tests at $p < 0.05$ and orange squares indicate marginally significant effects.

Despite this one, marginally significant difference ($p = 0.055$), we conclude that the data from our experiment are very similar to those gathered by Scherer [8]. We explain the small difference by noting that, in retrospect, some of our pictures intended to evoke contempt could have actually (also) have evoked disgust. For instance, we intended to use a photo of a teenager saluting in front of a swastika flag to evoke contempt, but many people found this photo to evoke disgust. External normative standards (dimension 14) would therefore be an appraisal dimension where this difference, between our data and that of Scherer [8] would show up.

4.2. PCA analyses and reduction of dimensions

We carried out a Principal Component Analysis (PCA; including several rotational variants) to investigate whether we needed all fifteen appraisal dimensions to explain most of the variance in the data from the questionnaires. The PCA revealed that the first principal component (i.e., the direction in which the variance of the data is largest) described 27% of the data and that the first three components sum up to 51%. In order to get above 80% of variance explained, the first 9 principal components were needed (see Figure 3).

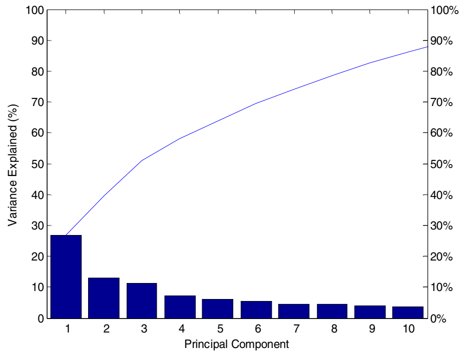


Figure 3: The percentage of variance explained per principal component (bars) and total percentage of variance with each additional component (blue line).

Figure 4 shows the data projected onto the first and second principal components found by PCA. The first observation to be made is that there seems to be little sub-clustering of data, i.e., the global structure in the data seems to be a single cluster without sub-groupings (e.g., per emotion). This is confirmed by the application of Agglomerative Hierarchical Clustering that only succeeds in recognizing the positive versus negative emotions.

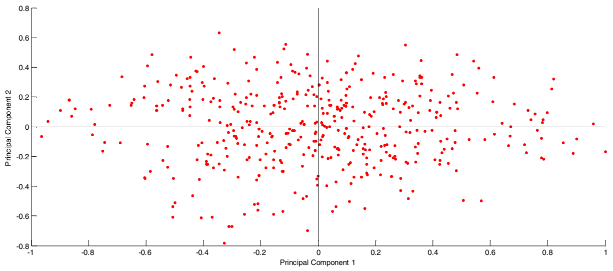


Figure 4: Projection of data samples onto the first and second principal component.

Projecting the original (appraisal) dimensions onto the dimensions found by PCA, as depicted in Figure 5, provides insight in the interrelations of appraisal dimensions, reflected in the data. The length and orientation of the blue line segments indicate how much the appraisal dimension contributed to the first and second PCA components. For instance, appraisal dimensions Conduciveness and Expectedness were closely aligned along the first component. This can be expected because, often, unexpected events are not conducive regarding one's goals. Another interesting observation was that Novelty and Causality by Chance were rather opposite to Control. Also this could be expected because often people do not feel (fully) in control when new and/or seemingly random events happen.

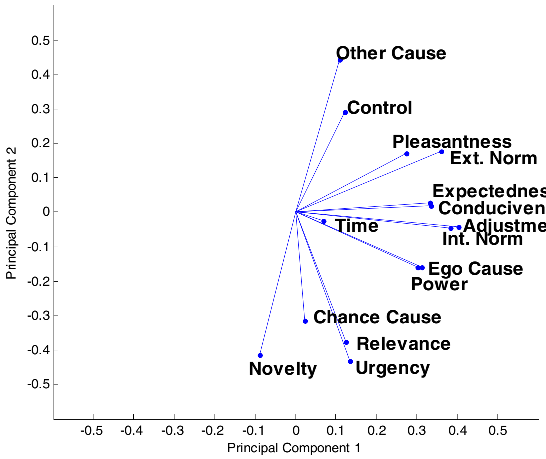


Figure 5: Projection of the appraisal dimensions into the space spanned by principal components 1 and 2.

Because components 1-3 collectively explained a little over half of the variance, we tried to label these components to come to a better understanding of how the data was structured. An initial visual inspection hinted that components 1-3 could perhaps be related to the valence, arousal, and dominance dimensions of Bradley and Lang [1].

We asked a new group of participants (N=29) to rate each of the 14 emotion words on a valence, arousal, and dominance scale and used these data to assess whether any of the three components would correlate to one of these affective dimensions. We exhaustively calculated correlations between the three components and the three affective dimensions (see Table 1) and observed that in general the correlations were not exceptionally strong.

Comparing our work to that of Scherer [10], we noted that although component 1 seemed to distinguish between positive emotions (particularly, Enjoyment, Joy, and Pride) from negative emotions, the correlations did not reflect this (see Table 1), because it turned out that within the positive and negative emotions the first principal component does not reflect the valence values properly. The strongest correlation was between the first principal component and dominance.

Table 1. Correlations between the first three principal components and affective dimensions.

Affective dimensions Principal components	Valence	Arousal	Dominance
1	-0.34	0.06	-0.56
2	0.25	-0.07	0.24
3	-0.07	0.44	-0.36

This finding could be explained with reference to Figure 5. This figure shows that Expectedness, Conduciveness, and Adjustment were contributing most to the first principal component. We suspect that experiencing a highly unexpected event having low conduciveness to achieving one's goals and low adjustment ability to change the outcome of the event could result in emotions having high dominance.

4.3. Classification results

In the end, the ultimate goal is to use the appraisal data as predictor for the corresponding emotions. Let us therefore investigate the classifying performance, using several classifying methods. We chose four techniques of different nature, including two classical classification techniques: k-Nearest Neighbors (kNN), which uses majority voting over the closest k neighboring training samples; and Artificial Neural Network (ANN; three layer feed forward, back propagation), which trains weights, connecting nodes on succeeding layers (including input and output layer), thereby composing a network that relates output (class label) to input (data). These techniques can be considered as black box techniques, that is, it is hard (if not impossible) to infer from the classifier what data properties it has learned to use to discriminate between classes.

The other two techniques are more transparent. They are prototype-based learning methods, which use prototypes in the same (mathematical) space as the data to represent the classes. These prototypes are interpretable as they show differences and similarities between classes. The simplest uses the means per class as prototypes; more advanced is Robust Soft Learning Vector Quantization (RSLVQ) [11], which (implicitly) takes into account variances and prior probabilities when learning the prototypes.

As a baseline reference we used a baseline classifier that returned the label of the class with highest prior probability, as specified by equation (1). It describes the best guess one can do without inferring any knowledge about the structure of the data.

$$\gamma = \arg\max_{c} p_c \text{, where } p_c = \frac{n_c}{\sum_{c'} n_{c'}} \text{,} \qquad (1)$$

where p_c is the class prior, inferred from the number of samples n_c representing class c.

Table 2. Classifier performance averaged over 10x10-fold cross validations for the complete dataset and specified subsets.

	Complete dataset: 14 classes		Rated 20+ times: 10 classes		Rated 35+ times: 5 classes	
Method	Hyper-parameter	Performance	Hyper-parameter	Performance	Hyper-parameter	Performance
Baseline		14.9 %		16.7 %		26.4 %
kNN	k = 7	24.5 ± 0.8 %	k = 13	27.9 ± 1.4 %	k = 13	48.9 ± 2.0 %
ANN	$N_{hidden} = 3$	25.2 ± 0.7 %	$N_{hidden} = 5$	28.1 ± 1.5 %	$N_{hidden} = 3$	46.2 ± 2.7 %
RSLVQ	$v_{soft} = 1$	24.3 ± 1.2 %	$v_{soft} = 10$	29.1 ± 0.8 %	$v_{soft} = 10$	47.3 ± 1.8 %
Means		21.2 ± 1.4 %		27.4 ± 1.0 %		48.5 ± 1.5 %

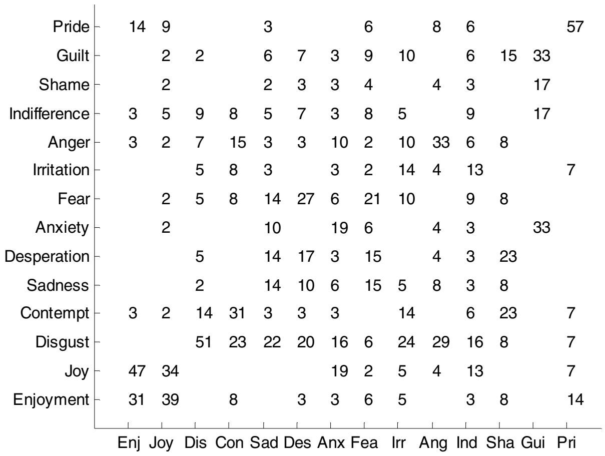
The classifiers were applied with 10-fold cross validation and the results reported are means and standard deviations over 10x10-fold cross validations.

As Table 2 shows, the performance of the classifiers ranged from 21.2% to 25.4% on the complete dataset which was less than 2 times the performance of the baseline classifier (14.9%). Interestingly, the Artificial Neural Network performed best when the hidden layer was small (N_{hidden}=3), thereby only returning 5 to 6 (out of 14) class labels, effectively disregarding the other emotions. Apparently its capacity was too small to fit to all 14 classes.

The dataset showed a large diversity of priors, indicating that some classes were underrepresented. Therefore, the classifiers might be unable to learn the properties of these classes to a sufficient degree. To test this, we filtered the data to only include emotions that had been rated at least 20 times. The filtering did not increase performance much as indicated in the fifth column of Table 2. Reducing the dataset even further to only the top 5 most rated classes (rated at least 35 times) showed a larger increase of performance (rightmost column of Table 2), up to 49.5%. However, the baseline performance also increased significantly (to 26.4%) because of the reduced number of classes. Finally we tried using only the 'basic' emotions (anger, disgust, fear, joy, and sadness) [4], resulting in similar findings. In summary, over all classification assessments, classifier performance was roughly twice the baseline performance.

Scherer also worked on a classifier of appraisal vectors onto emotion labels in the 'Geneva Expert System on Emotion' [8]. When comparing our results, in general, to the performance of Scherer's expert system (78% correct), they highlighted a big gap. However, Scherer's system considered both the first and second hit in the performance. The second best option was given in case the first option was perceived incorrect by a user. Therefore, we recalculated Scherer's performance by only including the first hit and found it to be 60.0% where the baseline performance is 22.9%. Moreover, this performance was not a true classification performance because the class labels were not identified separately from the classifier predictions, i.e., users were asked whether the prediction could be plausible.

We analyzed the confusion matrix (Figure 6) of RSLVQ to study whether particular emotions being classified incorrectly were responsible for the classifier performance, or whether for all emotions the performance failed. The analysis revealed that for most emotions at least some instances are classified correctly. However, no single instance of Shame was classified as such, indicating that Shame was easily mistaken for other emotions. As could be expected, Joy and Enjoyment were confused often, as well as Disgust and Contempt. Fear turned out to be confused often with Desperation and Sadness. Surprisingly such confusions did not hold for Anxiety and Fear or Anger and Irritation, which could be expected to be close as well. The confusion matrices for the other classifiers did not reveal (other) interesting confusion patterns.



	Enj	Joy	Dis	Con	Sad	Des	Anx	Fea	Irr	Ang	Ind	Sha	Gui	Pri
Pride	14	9			3			6		8	6			57
Guilt		2	2		6	7	3	9	10		6	15	33	
Shame		2			2	3	3	4		4	3		17	
Indifference	3	5	9	8	5	7	3	8	5		9		17	
Anger	3	2	7	15	3	3	10	2	10	33	6	8		
Irritation			5	8	3		3	2	14	4	13			7
Fear		2	5	8	14	27	6	21	10		9	8		
Anxiety		2			10		19	6		4	3		33	
Desperation			5		14	17	3	15		4	3	23		
Sadness			2		14	10	6	15	5	8	3	8		
Contempt	3	2	14	31	3	3	3		14		6	23		7
Disgust			51	23	22	20	16	6	24	29	16	8		7
Joy	47	34					19	2	5	4	13			7
Enjoyment	31	39		8		3	3	6	5		3	8		14

Figure 6: Confusion matrix showing the (user indicated) labeled emotions on the horizontal axis and the (RSLVQ) classified instances vertically. The numbers represent the percentages of classified emotions per labeled emotion.

Finally, we used the inherent transparency of prototype-based learning by assessing the (relative) positions of the prototypes to get a better understanding of how the data was structured. We took the learned prototypes of RSLVQ and applied a distance preserving projection on two dimensions using Curvilinear Component Analysis (CCA) [3]. This method reduces the number of dimensions by optimally preserving the relative distances between the data points (in this case the prototypes). Figure 7 shows that the prototypes settled at non-coinciding locations and showed a large (relative) distance between the positive emotions (Enjoyment, Joy, and Pride) and the negative emotions. Repeated tries (using different subsets of data for training the prototypes) resulted in similar graphs. We did find that the negative emotions seemed to vary quite a lot amongst themselves.

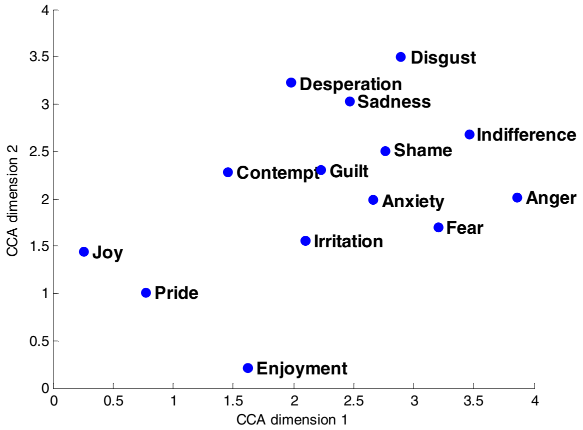


Figure 7: RSLVQ prototypes projected in a reduced (2D) space, using CCA.

5. Discussion

To summarize, we investigated whether empathizing with an emotional event resulted in experiencing an emotion in a similar way as when recollecting personal emotional events; the latter being the cornerstone of Scherer's component-process (appraisal) model [8]. We gathered data to assess this question and to determine whether machine-learning techniques would enhance earlier classification attempts of appraisal data to emotion.

We show that, although using a slightly different methodology, the data, collected from 240 participants, is highly similar to the data reported in [8]. This finding indicates that the Scherer's CPM generalizes to other contexts than recollecting personal emotional events to evoke an emotion.

A Principal Components Analysis and correlation analysis do not show that the affective dimensions valence, arousal, and dominance map onto components 1-3 as we expected from [10]. The strongest connection that we observe is one between the first principal component and dominance. This seems to contrast with findings in [10] that describe a far better fit of the principal components and affective dimensions. With the current data available, we cannot explain these differences yet.

A classification task shows that classifying performance is rather low and does not differ much between classifier techniques. The use of prototype based classifiers enables us to further explore the differences and similarities between emotions. It reveals that, although the collected dataset is noisy, a quite clear distinction between the positive and negative emotions is present.

To conclude, we find that the CPM generalizes from contexts of recollection of personal emotional events to empathizing with (generic) emotional events. Our attempt to discriminate between discrete emotions on the level of appraisal, in a pure classification task, does not lead to very promising results. Machine-learning techniques have a hard time finding the class boundaries because of relatively high levels of intra-class variation. Future studies could look into exploring whether the high intra class variation could be explained by the notion of mixed emotions [2]. That is, it could well be that, from a computational perspective, emotions could better be represented as a vector of contributions (between 0 and 1) of all emotions instead of a single discrete emotion (which could be represented as a contribution vector in which a single contribution is 1, the others having value 0).

Acknowledgements

We thank three internal and three anonymous external reviewers for their helpful comments that have improved this paper.

References

[1] M. M. Bradley, and P. J. Lang. Measuring emotion: The Self-Assessment Manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25: 49-59, 1994.

[2] P. Carrera, and L. Oceja. Drawing mixed emotions: Sequential or simultaneous experiences? Cognition and emotion 21(2): 422-441, 2007.

[3] P. Demartines, and J. Hérault. Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets. IEEE Transactions on Neural Networks 8(1): 1197-1206, 1997.

[4] P. Ekman, Universals and cultural differences in facial expressions of emotion. In J. Cole (ed.), Nebraska Symposium on Motivation 19: 207-283. Lincoln, NE: University of Nebraska Press, 1972.

[5] A. Ortony, G. L. Clore, and A. Collins. The cognitive structure of emotions. Cambridge, Cambridge University Press, 1994.

[6] T. J. M. Overbeek, A. van Boxtel, J. H. D. M. Westerink. Development of an emotion-eliciting stimulus set: Results of emotional pictures and film fragments ratings. Unpublished technical report (PR-TN2007/00574), 2007.

[7] J. A. Russell. Core Affect and the Psychological Construction of Emotion. Psychological Review 110, 145-172, 2003.

[8] K. R. Scherer. Studying the emotion-antecedent appraisal process: An expert system approach. Cognition and Emotion 7: 325–355, 1993.

[9] K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, and J. Johnstone (eds.). Appraisal processes in emotion: Theory, methods, research: 92–120, New York: Oxford University Press, 2001.

[10] K. R. Scherer, E. S. Dan, and A. Flykt. What determines a feeling's position in affective space? A case for appraisal. Cognition and emotion 20(1): 92–113, 2006.

[11] S. Seo, and K. Obermayer. Soft learning vector quantization. Neural computation 15: 1589–1603, 2003.