Measures and metrics for automatic emotion classification via FACET

Pasquale Dente[1], **Dennis Küster**[1], **Lina Skora**[2], **Eva G. Krumhuber**[2]

Abstract. For dynamic emotions to be modelled in a natural and convincing way, systems must rely on accurate affective analysis of facial expressions in the first place. The present work introduces two measures for evaluating automatic emotion classification performance. It further provides a systematic comparison between 14 databases of dynamic expressions. Machine analysis was conducted using the FACET system, with an algorithm calculating recognition sensitivity and confidence. Results revealed the proportion of facial stimuli that could be recognised by the machine algorithm above threshold evidence, showing significant differences in recognition performance between the databases.

1 INTRODUCTION

The computational modelling of dynamic facial expressions is a difficult challenge [1] that must be met to understand, and ultimately to simulate natural emotions convincingly. While single images can, in principle, be coded manually on the basis of the Facial Action Coding System [2], large stimulus sets that span a range of facial behaviour require a robust automated approach. This particularly applies to naturally occurring dynamic facial expressions which are often elicited "in the wild". Rather than depicting clean exemplars of an emotion, they occur spontaneously, at varying intensities, with Action Units (AUs) that are not part of prototypical configurations [3, 4]. Automatic analysis of spontaneous as well as posed expressions therefore acts as an essential criterion from which to identify and synthesise complex emotional behaviour.

The last two decades have seen great advances in the development of stimuli for facial expression and emotion research, taking them from static to dynamic portrayals [5, 6]. In [7] we have provided a conceptual review of existing dynamic facial expression databases. The present paper describes an empirical test of 14 of the available datasets in terms of machine recognition, with a focus on the six basic emotions (happiness, sadness, fear, anger, sadness, and surprise) [8]. In doing so, we discuss different measures and metrics for automatic emotion classification and their respective role in determining detection rates.

2 CLASSIFICATION APPROACHES

14 datasets were chosen, each containing videos classified by the dataset author as portraying one of the six basic emotions (happiness, sadness, anger, fear, disgust, surprise). With the exception of DynEmo (only four emotions) and DISFA (only five emotions), we selected two portrayals of each emotion for each database, yielding 12 portrayals per dataset. Facial activity was measured through video-based analysis using the iMotions Attention Tool and its FACET module (version 5.7) [9]. FACET is a commercial facial expression recognition software based on the Computer Expression Recognition Toolbox (CERT) [10]. Recently, FACET has been used in an increasingly broad range of psychological and applied research, such as the attribution of emotions to faces of own and other races [11], the relative saliency of individual AUs [12], as well as attempts toward an automatic recognition of persuasiveness with the aid of features from facial expressions [13].

FACET outputs per-frame "evidence values" that are defined as describing how likely an "expert human coder" would be to categorise an expression in a given frame as reflecting the intended emotion [14]. FACET evidence values are recommended for any in-depth analysis as per the manual, and are described as "very similar" to a Z-score centred around zero, i.e., the set value is assumed to reflect an even chance that an expression is to be categorized as neutral [14]. FACET outputs these per-frame values in a range from -4 to +4. Unfortunately, no recommendations are made by FACET concerning the aggregation of evidence values for interpretation beyond the level of individual frames. In the present research, we therefore decided to further aggregate the output evidence values, and to test the results empirically against the database emotion labels used as the ground truth. We specified the threshold to indicate a positive per-frame recognition for a given expression as evidence > 0. In order to evaluate machine recognition performance at the per-video level across the databases, we computed two additional metrics: recognition sensitivity and recognition confidence.

Recognition sensitivity

The sensitivity metric is a simple measure of the percent of frames containing the target evidence > 0. It can be used to assess which databases show the largest percentage of frames with the target expression (e.g., happiness) above the detection threshold. As such, it provides guidance for the evaluation of databases that show target expressions for a substantial amount of time. In the present context, the databases vary substantially in average stimulus (i.e. expression) duration, as well as the proportion of emotional frames as opposed to neutral or low-intensity frames. To account for this variability, recognition sensitivity was computed for each expression as the percentage

[1] Dept. of Psychology and Methods, Jacobs Univ. Bremen, 28759 Bremen, Germany. Email: {p.dente, d.kuester}@jacobs-university.de.

[2] Dept. of Experimental Psychology, University College London, WC1H 0AP, UK. Email: {p.skora, e.krumhuber}@ucl.ac.uk.

of frames with target evidence > 0, divided by the sum of all frames, multiplied with 100. The result was then aggregated across the whole database to yield an average percentage score indicating the overall proportion of frames that were correctly identified as containing evidence for the target expression. This approach is thus broadly in agreement with the statistical definition of "sensitivity" in so far as it reflects the extent to which a positive item was correctly classified. However, we use this term only loosely due to the lack of precision in the definition of evidence values generated by FACET.

$$\text{Sensitivity} = \frac{N\ (\text{frames with target evidence} > 0)}{N\ (\text{frames})} * 100$$

Based on the guidelines provided by FACET [14], evidence > 0 can be interpreted as the least conservative threshold for positive classification. As such, there is more evidence for the presence of a given expression than evidence for its absence. However, evidence values can be substantially higher than 0, up to the point where a near perfect certainty (> 2) can be assumed that an expression is present. More stringent thresholds place higher demands on classification rates, which results in lower expression recognition as the evidence threshold increases.

From inspection of Figure 1, above-threshold recognition across all 14 databases did not decay equally for all emotional expressions. While 54.71% (*SD* = 27.67) of frames in happiness were classified with near-perfect evidence (a decrease of 17.4%), only 3.84% (*SD* = 7.79) of frames in sadness were classified with the same evidence threshold > 2 (a decrease of 36.46%). This suggests that for sadness only a small number of stimuli could be classified with high certainty.

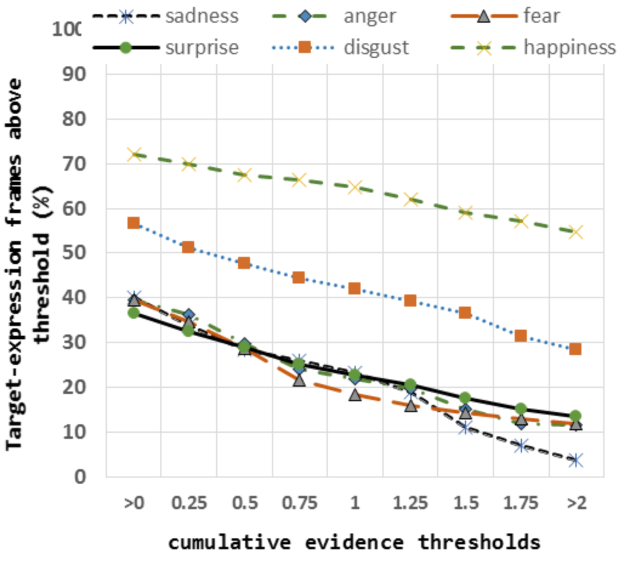

Figure 1. Percentage of Target-Expression Frames as a Function of Evidence Threshold.

Recognition confidence

Recognition confidence reflects the proportion of above-threshold target evidence (x) relative to the total above-threshold evidence, consisting of the target evidence (x), plus non-target evidence (y). Both (x) and (y) were computed as a sum of all frames (i) of a given clip, for FACET evidence values above the rejection threshold > 0. We excluded any evidence below this threshold because evidence < 0 reflects an assessment of the system that a given expression was not present. The ground truth to distinguish "target" vs. "non-target" evidence was provided by the expert labels provided for the validated databases. By multiplication with 100, the score yields a "percentage" value that is comparable to human confidence measures. For example, if a clip was labelled as "happy" in the validated database, and FACET only reported above rejection threshold evidence for happiness but no above-threshold evidence for any other expression, recognition confidence would be 100%. Recognition confidence thus provides a more robust metric that takes into account false-positive classifications, as well as the summative confidence reflected by the per-frame FACET evidence values.

$$\text{Confidence} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} * 100$$

While a more conservative threshold implies stringent sensitivity in the classification of a target expression, non-target expressions may tend to be detected more frequently even at lower thresholds. As a result, recognition confidence should be more robust (compared to plain sensitivity scores) because conservative thresholds allow for more cases of non-target evidence to be filtered out. Furthermore, recognition confidence is weighted by the respective evidence values.

As can be seen in Figure 2, recognition confidence scores were more robust, with slight increases for happiness and disgust up to thresholds of 0.75 to 1.00. For these two emotions, recognition confidence remained stable even at very conservative thresholds, suggesting that target expressions could be easily identified. The other four emotions showed some decline in recognition confidence with higher thresholds. Yet, variation in confidence scores as a function of the evidence threshold was still modest compared to the results obtained for recognition sensitivity.

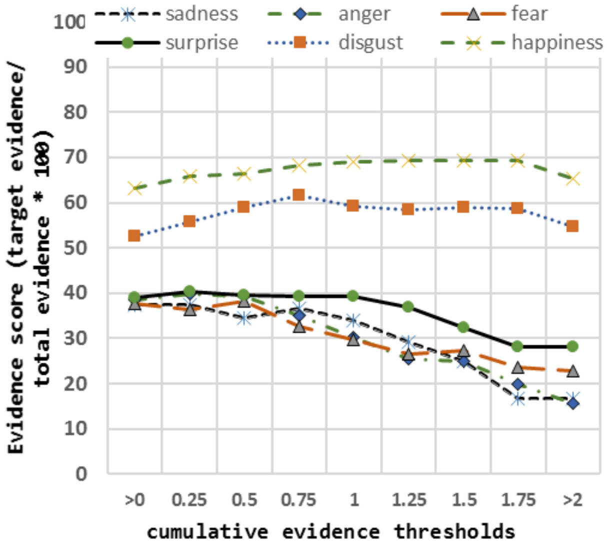

Figure 2. Evidence Scores as a Function of Evidence Threshold.

3 DETECTION RESULTS

Based on the sensitivity and confidence scores, it was possible to evaluate the extent to which each of the 14 databases yielded > 0 threshold evidence for emotion detection. For this, scores for the two portrayals per emotion were averaged for each database. Non-parametric bootstrap ANOVAs (*N* boots = 5000) were performed on the machine classification data. Significant differences occurred for the type of emotion, *F*(5, 124) = 5.86, *p* < .001, as well as the type of database (i.e. spontaneous vs. posed; *F*(1, 124) = 12.00, *p* < .01, with target expressions being detected with higher confidence in posed than spontaneous databases. Happiness was overall recognised with the highest confidence (90.86, *SD* = 23.24), followed by disgust (76.45, *SD* = 32.32), while fear was recognised with the lowest degree of confidence (47.70, *SD* = 39.39), and sadness performing slightly better (51.62, *SD* = 44.39).

As can be seen in Table 1, there was a spread of machine recognition performance between databases, ranging from complete failures to detect any evidence (STOIC) to near perfect performance (ADFES). Additionally, a few databases (e.g., BU-4DFE) appeared to perform substantially better when recognition confidence as opposed to sensitivity was assessed. Such databases may provide relatively clear expression data for machine analysis, albeit likely with a somewhat larger proportion of below-threshold frames. Overall, sensitivity and recognition confidence scores for individual clips were highly correlated (Pearson's *r* = 0.81), suggesting a clear linear relationship between both metrics. This was the case in particular for the subset of stimuli drawn from spontaneous databases (*r* = 0.91).

Database	**Sensitivity** ***Mean (SD)***	**Confidence** ***Mean (SD)***
ADFES	**80.04** (3.15)	**95.70** (9.52)
BINED (spon)	**33.73** (47.15)	**39.01** (47.44)
BU-4DFE	**62.99** (28.74)	**90.46** (27.00)
CK	**63.78** (24.30)	**86.01** (28.55)
D3DFACS	**61.33** (34.36)	**56.05** (36.40)
DaFEx	**36.32** (31.75)	**47.91** (42.96)
DISFA (spon)	**46.05** (40.44)	**56.82** (47.15)
DynEmo (spon)	**19.81** (28.47)	**21.14** (26.84)
FG-NET (spon)	**22.29** (23.21)	**37.67** (44.86)
GEMEP	**29.55** (30.71)	**29.80** (31.96)
MMI	**46.48** (30.17)	**74.86** (38.22)
MPI	**52.31** (22.98)	**68.97** (36.92)
STOIC	**failed**	**failed**
UT Dallas	**42.62** (41.23)	**61.54** (47.71)
Mean	**47.95** (33.86)	**62.73** (41.38)

Table 1. Machine Sensitivity and Confidence Mean Scores for 14 Databases. Spon = Databases with Spontaneous Portrayals

By ranking the databases on both metrics, Figure 3 demonstrates the relative advantage of taking non-target evidence into account in the assessment of a dataset's recognition confidence. For example, D3DFACS is the only database showing lower confidence than sensitivity. This is likely to be due to low per-frame evidence found by the system for semi-profile views of the 2D video clips in this database. While there was an overall significant effect of database type, performance of FACET at the level of each individual database suggests that factors related to the construction of a database may be more important than their posed/spontaneous nature per se.

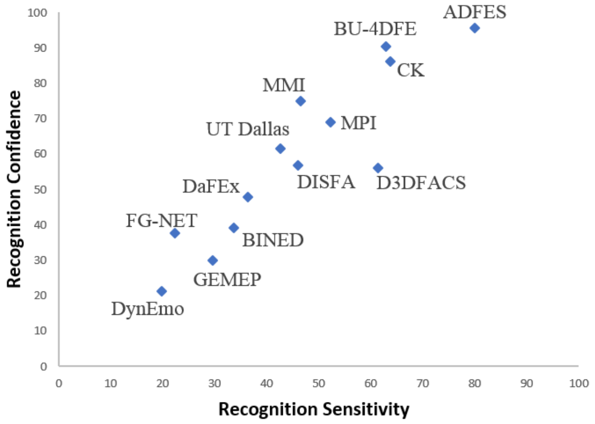

Figure 3. Correlation between Sensitivity and Confidence Scores across the Databases.

4 CONCLUSIONS & FUTURE WORK

The present research introduced and assessed two measures for machine recognition using FACET. Both sensitivity and confidence scores provided a robust method for evaluating emotion classification performance. They also allowed for a systematic comparison between 14 databases with dynamic facial expressions. This is the first empirical challenge of dynamic datasets in terms of automatic emotion detection. Detection rates were above 50% for the majority of the databases. Nonetheless, there was also a substantial number of sets with relatively weak performance (DynEmo, FG-Net, BINED), especially when portrayals were spontaneous rather than posed. In view of the limitations of machine analysis to deal with changes in viewing angle and overall visibility of the face [9, 10], the relatively uncontrolled nature of natural/spontaneous expressions, e.g., in online interaction [16], appears to pose additional challenges for automatic classification. Future work could aim for more high-quality data samples [11] on par with the technical recording setup used for some of the best performing posed databases such as ADFES or BU-4DFE. We suggest that an approach that combines confidence and sensitivity metrics can shed light on potential issues and limitations of dynamic facial expression databases. Full results including a larger set of stimuli will indicate how machine classification performs across a number of mediating factors such as the number and type of emotions, gender, and age [12]. The measures and metrics presented in this paper are comparable to human recognition performance for comprehensive database examination in the future.

REFERENCES

[1] S.C. Marsella, J. Gratch, and P. Petta, P. Computational Models of Emotion. In: *A Blueprint for Affective Computing: A Sourcebook and Manual.* K. R. Scherer, T. Bänziger, E. Roesch (Eds.). Oxford University Press (2010).

[2] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System: The manual on CD ROM.* Research Nexus, Salt Lake City, UT (2002).

[3] A. Kappas, E.G. Krumhuber, and D. Küster. Facial Behavior. In: *Nonverbal Communication (Handbooks of Communication Science, HOCS 2).* J. A. Hall, M. L. Knapp (Eds.). Mouton de Gruyter (2013).

[4] J. Cohn, Z. Ambadar, and P. Ekman. Observer-Based Measurement of Facial Expression with the Facial Action Coding System. In: *Handbook of Emotion Elicitation and Assessment*. J. A. Coan, J. J. B. Allen (Eds.). Oxford University Press (2007).

[5] E.G. Krumhuber, A. Kappas, and A.S.R. Manstead. Effects of Dynamic Aspects of Facial Expressions: A Review. *Emotion Review,* 5: 41-46, (2013).

[6] E.G. Krumhuber and P. Skora. Perceptual Study on Facial Expressions. In: *Handbook of Human Motion*. B. Müller, S. Wolf (Eds.). Springer-Verlag (2016)

[7] E.G. Krumhuber, P. Skora, D. Küster, and L. Fou. A Review of Dynamic Datasets for Facial Expression Research. *Emotion Review* (in press).

[8] P. Ekman. An Argument for Basic Emotions. *Cognition and Emotion,* 6: 169-200 (1996).

[9] iMotions Biometric Research Platform 5.7, *Emotient FACET*, iMotions A/S, Copenhagen, Denmark (2016).

[10] G. Littlewort, J. Whitehill, T. Wu, I. Fasel., M. Frank, J., Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on IEEE: 298-305* (2011)

[11] C.S. Hu, Q. Wang, T. Han, E. Weare, and G. Fu. Differential emotion attribution to neutral faces of own and other races. *Cognition and Emotion*, 31, 360-368 (2017).

[12] M.G. Calvo, A. Gutiérrez-García, and M. Del Líbano. What makes a smiling face look happy? Visual saliency, distinctiveness, and affect. *Psychological Research*, 1-14 (2016).

[13] B. Nojavanasghari, D., Gopinath, J., Koushik, T., Baltrušaitis, and L.P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction,* 284-288 (2016). ACM.

[14] iMotions A/S. *Attention Toll FACET Module Guide 130806* (2016). Retrieved from https://imotions.com/guides/ on [03/07/2017]

[15] P. Dente, D. Küster, and E.G. Krumhuber. Boxing the Face: A Comparison of Dynamic Facial Databases Used in Facial Analysis and Animation. In: *Procs. 1st Joint Conf. on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP),* Vienna, Austria, ACM Press, p. 5 (2015).

[16] D. Küster, E.G. Krumhuber, and A. Kappas. Nonverbal Behavior Online: A Focus on Interactions with and via Artificial Agents and Avatars. In: *Social Psychology of Nonverbal Communications*. A. Kostic, D. Chadee (Eds.). Palgrave Macmillan (2014).

[17] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, A Survey of Facial Affect Recognition Methods: Audio, Visual and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 31: 39-58 (2009).

[18] G. Sandbach, S. Zafeiriou, M. Pantic, and J. Yin. Static and Dynamic 3D Facial Expression Recognition: A Comprehensive Survey. *Image and Vision Computing,* 30: 683-697 (2012).