OVERVIEW OF THE ADVANCEMENTS IN AUTOMATIC EMOTION RECOGNITION: COMPARATIVE PERFORMANCE OF COMMERCIAL ALGORITHMS

A PREPRINT

Mariya Malygina
Department of Psychology
National Research University
Higher School of Economics;
Neurodata Lab
Moscow, Russia
mmalygina@hse.ru

Mikhail Artemyev
Neurodata Lab
Moscow, Russia
m.artemyev@neurodatalab.com

Andrey Belyaev
Neurodata Lab
Moscow, Russia
a.belyaev@neurodatalab.com

Olga Perepelkina
Neurodata Lab
Moscow, Russia
o.perepelkina@neurodatalab.com *

December 24, 2019

ABSTRACT

In the recent years facial emotion recognition algorithms have evolved and in some cases top commercial algorithms detect emotions like happiness better than humans do. To evaluate the performance of these algorithms, the common practice is to compare them with human-labeled ground truth. This article covers monitoring of the advancements in automatic emotion recognition solutions, and here we suggest an additional criteria for their evaluation, that is the agreement between algorithms' predictions. In this work, we compare the performance of four commercial algorithms: Affectiva Affdex, Microsoft Cognitive Services Face module Emotion Recognition, Amazon Rekognition Face Analysis, and Neurodata Lab Emotion Recognition on three datasets AFEW, RAVDESS, and SAVEE, that differ in terms of control over conditions of data acquisition. We assume that the consistency among algorithms' predictions indicates the reliability of the predicted emotion. Overall results show that the algorithms with higher accuracy and f1-scores that were obtained for human-labeled ground truth (Microsoft's, Neurodata Lab's, and Amazon's), showed higher agreement between their predictions. Agreement among algorithms' predictions is a promising criteria in terms of further exploring the option to replace human data labeling with automatic annotation.

***K*eywords** Emotion Recognition · Affective computing · Facial expression

1 Introduction

Various automatic facial expression recognition systems have been developed to detect human emotions. Affective computing research becomes more reliable, valid, and accessible as computer science develops [1]. In general, we see the increasing support for the idea that automated facial expression analysis is technically achievable. Advanced emotion classifiers are close to reaching the level of emotion recognition that people show. The commercial applications of emotion recognition using facial expressions are multiple: banking, market research, human resources, advertising,

*We express our gratitude to Francesca Del Giudice for editing the text of the article.

and health care are the examples of industries where emotion prediction is valuable [2]. Specific industries require rich annotations for algorithms to perform with high accuracy. In this work, we are exploring the possibility of annotating using emotional labels assigned by automatic classifiers instead of human annotations. We assume that automatic labeling could be more efficient than human-based annotation in certain conditions. We tested the performance of four commercial algorithms on the datasets with different level of control over conditions of data acquisition. As different algorithms use distinct statistical methods and datasets to train the machine learning procedures, they differently classify emotions [3]. If algorithms are capable to recognize what people actually express (in other words, technologies work), then different algorithms would give consistent predictions per video recording, using particular emotion labels. Predictions should be similar despite the different learning history of the algorithms. The present research is to reveal whether agreement between algorithms' predictions may be a relevant criteria to evaluate the performance of automatic emotion recognition. If predictions obtained from different algorithms show high agreement, then annotating new unlabeled data using them could be promising instead of manual human annotations.

2 Related issues

2.1 Human-related challenges of automatic emotion recognition

Performance of automatic emotion recognition is usually compared with performance of human annotators[4]; however we note that the latter is not infrequently flawed by various perceptual and conceptual errors. Humans appear to be biased observers when it comes to emotional expressions of others. Properties of emotional stimuli, as well as human cognitive and emotional state, influence the processing of emotional information. One of the factors is the effect of mood congruence on process of information encoding [5], which is well-evidenced especially when it comes to perception of emotional faces. According to the mood congruence effect, stimuli are processed easier if they correspond to the emotional state of an annotator. Cognitive biases related to facial expressions are present even on the level of neural response to emotional faces. For instance, it was shown that the effect of selective adaptation to facial expression has event-related potential correlates, when the same neutral face is perceived as sad following the exposure of a happy face and perceived as happy following the exposure of a sad face [6]. The aftereffects of this exposure are decreased sensitivity to emotional faces that were exposed and increased sensitivity to emotions of other classes. Such visual aftereffects, as well as preference for stimuli congruent with the current mood of a person, may become a serious issue when considering annotation procedure where humans rate large amounts of emotional data. In addition, high cognitive load and tiredness may contribute to shifts in evaluations of emotional expressions [7]. Keeping this in mind, we should consider that automatic emotion recognition could be more efficient than human-based annotation when it comes to tasks or challenges in certain industries. Moreover, algorithms may predict some of the emotion classes even better than humans. As Dupré, Krumhuber, Kuster, and McKeow [8] have shown, some algorithms outperformed human-level accuracy of recognition for happiness and reached human-level accuracy for sadness when classifying emotions in BU-4DFE and UT-Dallas datasets. Thus, considering that observer-based ground truth incorporates certain amount of errors, human rates should not necessarily be considered the only criterion for estimating the algorithms' performance. To evaluate the recognition performance, we are going to compare the predictions of algorithms under review with each other. If several algorithms show similar results (e.g. estimated weights/probabilities of each emotion label per frame), then we will assume that the consistency among predictions indicates the reliability of the predicted emotion. In this paper, statistical procedures are applied: 1) to compare ground truth data to algorithms' predictions; 2) to evaluate the agreement between the predictions of four algorithms. We believe that the agreement between algorithms' predictions could be a relevant criteria to evaluate the performance of automatic emotion recognition. Such comparison could become a means of monitoring of advancements in automatic emotion recognition solutions.

2.2 Challenges related to training and testing in automatic emotion recognition

In emotion recognition research, labeled data used for training and testing are limited due to difficulties in acquiring, particularly when it comes to 'in the wild' data [9]. Labelling takes considerable time and requires human effort. Since we have plenty of unlabelled data available, a lot of effort is aimed at labelling said data while simultaneously reducing the costs of this operation. The use of crowdsourcing platforms is a good way of reducing costs while obtaining enough annotations. Alternatively, we could reduce the very amount of human effort. We can distribute human efforts more efficiently if we concentrate on emotion classes most relevant for algorithm training. This approach is used for semi-autonomous learning and the model of active learning [10]. Another method is dynamic labeling that needs less annotations per case if the annotations are in high agreement [11]. However, in this paper, we are exploring the possibility of annotating using emotional labels assigned by automatic classifiers with high agreement among their predictions.

Algorithms mostly use acted facial expressions for training which affects their performance on mixed or hidden emotions that are more common 'in the wild'. Both quality and nature of training datasets have a significant impact on accuracy of recognition, and training data should resemble the one the algorithm is supposed to deal with. Another possible issue is the problem of overfitting [12]: they classify correctly only the emotions similar to those in acted training datasets, and lose their flexibility when applied to complex facial expressions. The specifics of acted datasets is that the offset phase of emotion is frequently missing [13]. To summarize, the annotation of spontaneous expressive data remains a major topic in emotion research due to the limited labelling and existing level of disagreement among annotators.

3 Materials and Methods of Data Analysis

The present analysis aims to compare the performance of four commercial algorithms of automatic facial expression recognition: Affectiva Affdex SDK, Microsoft Cognitive Services Face API, Amazon Rekognition Face Analysis[1], and Neurodata Lab Emotion Recognition. All four were tested on three publicly available datasets: AFEW, RAVDESS, and SAVEE that contain a total of 4358 videos. We note that the term "performance of algorithms" traditionally means the accuracy of their predictions compared to human-labelled ground truth. However in this study, we use this term both as the accuracy of their predictions and f1-scores. Firstly, we compared accuracy to human-labelled ground truth. Secondly, we compared accuracy of four algorithms to each other by measuring agreement among them. To do so, we used correlations between predictions for every emotion class. Thirdly, we calculated f1-score for a) predictions and human-labelled ground truth, and b) between predictions of all four algorithms. These steps allowed us to identify possible patterns or dependencies in how the algorithms performed with ground truth and how they were related to each other.

AFEW, RAVDESS, and SAVEE datasets description. In this study, we used three dynamic datasets containing short audiovisual fragments with one specific emotion expressed in every fragment: Acted Facial Expressions in the Wild (AFEW) [14], The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[15], and Surrey Audio-Visual Expressed Emotion (SAVEE)[16]. These datasets are observer-based and assessed with categorical annotation (see Table 1 for details). The datasets differ in terms of data acquisition, while all of them have resolutions sufficient for performing facial expression analysis. RAVDESS is a set of records of professional actors shot in laboratory settings, SAVEE is a set of records of non-professional actors in laboratory settings, while AFEW consists of movie fragments where actors were recorded in natural settings with environmental noise.

Table 1: Description of data used for analysis. Note that not all information that datasets contain was included (e.g. video labeled as "Calm" or audio-only recordings were excluded).

	Algorithms					
Dataset	Year of release	Subjects	Videos	Modalities	Type	Emotion labels
AFEW	2011	330	1426	audio-visual	spontaneous	Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise
RAVDESS	2018	24	2452	audio-visual	posed	Angry, Disgust, Fearful, Happy, Neutral, Sad, Surprised
SAVEE	2011	4	480	audio-visual	posed	Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise

SAVEE and RAVDESS captured face and neck regions on plain background. Actors pronounce phrases (in SAVEE) or pronounce phrases and sing (in RAVDESS) while simultaneously expressing different emotions, which reflects within-subject design of these datasets. AFEW contains close to natural facial expressions, head poses, and movements of actors up to full body capture in a between-subjects format. The choice of datasets was conditioned by several factors. Firstly, they are publicly available for research purposes and together they represent a large amount of various data. In addition, RAVDESS is a relatively new dataset (released 2018) and has not been used for algorithm training yet. Therefore the selection of RAVDESS enabled us to test algorithms on unfamiliar data. While SAVEE and RAVDESS were recorded in lab environment, AFEW presents 'in the wild' data. The comparison of 'in the wild' data to lab-controlled data shows how stable advanced algorithms react to changing conditions.

Algorithms description. Microsoft's, Affectiva's, Amazon's, and Neurodata Lab's emotion recognition solutions classify emotions by analyzing facial expressions and return the confidence level for each emotion[17, 18, 8]. Given algorithms apply the principle of single-frame analysis. Affectiva Affdex SDK, Microsoft Cognitive Services Face API, and Amazon Rekognition Face Analysis share the Facial Action Coding System standard for measuring emotional facial expressions using the facial landmark points [17]. Neurodata Lab Emotion Recognition classifies emotions by analyzing the holistic image of a face without specifying any action units based on end-to-end deep learning approach. These particular algorithms were chosen as they ascribe similar 'basic' emotion labels to predictions, though the labels

[1]We compute the analysis for Affectiva, Amazon, and Microsoft using the versions of their cloud API available in May 2019.

themselves are slightly different from each other. Microsoft Cognitive Services Face API gives predictions for eight emotions: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise [19]. Affectiva Affdex SDK predicts seven emotional classes: anger, contempt, disgust, fear, joy, sadness, and surprise [17]. Amazon Rekognition works with seven emotion labels: angry, calm, confused, disgusted, happy, sad, and surprised [18]. Neurodata Lab Emotion Recognition gives predictions for seven emotion classes: anger, anxiety, disgust, happy, neutral, sad, and surprise.

4 Performance of the algorithms

4.1 Recognition accuracy compared to ground truth

Three-way ANOVA (Algorithm x Dataset x Emotion label) was conducted to examine the differences in the proportion of correct classifications. We compared the recognition of target emotion using five emotion classes: Anger, Disgust, Happiness, Sadness, and Surprise. These emotion classes are analyzed by the aforementioned algorithms. We note that due to the specifics of the algorithms, Affectiva does not predict such emotion class as Neutral, and Amazon does not recognize such class as Fear. Therefore, the video recordings that were labelled as Neutral and Fear were excluded from ground truth data for ANOVA analysis. However, the rest two algorithms do recognize Neutral and Fear. So when the rest two algorithms predicted the emotion classes of Neutral and Fear, we discarded this result and instead took for analysis the emotion with second highest recognition confidence score per video. This way the comparison of the algorithm performance was made legit. For ANOVA, accuracy was considered as a dependent variable. Accuracy refers to a proportion of correct predictions among all cases, when predicted emotion label matches the ground truth label.

Table 2: Results obtained with ANOVA for recognition accuracy

	Sum of Squares	df	Mean Square	F	p-value
Dataset	39.169	2	19.584	113.92	< .001
Algorithm	33.444	3	11.148	64.85	< .001
Emotion Label	135.937	4	33.984	197.69	< .001
Dataset * Algorithm	7.686	6	1.281	7.45	< .001
Dataset * Emotion Label	103.351	8	12.919	75.15	< .001
Algorithm * Emotion Label	153.616	12	12.801	74.47	< .001
Dataset * Algorithm * Emotion Label	48.785	24	2.033	11.82	< .001
Residual	1247.707	725	0.172		

Significant main effect was found for the relation between algorithm type and recognition accuracy $F(3, 7317) = 64.85$, $p < .001$, $\eta_p^2 = .039$. Post hoc comparisons using Tukey HSD correction identified that Microsoft Face Emotion Recognition outperformed other algorithms ($p < .001$ for Microsoft M = 0.62, 95% CI [0.59, 0.64]; Neurodata Lab M = 0.53, 95% CI [0.51, 0.55]; Amazon M = 0.49, 95% CI [0.47, 0.51]; Affectiva M = 0.39, 95% CI [0.37, 0.41]).

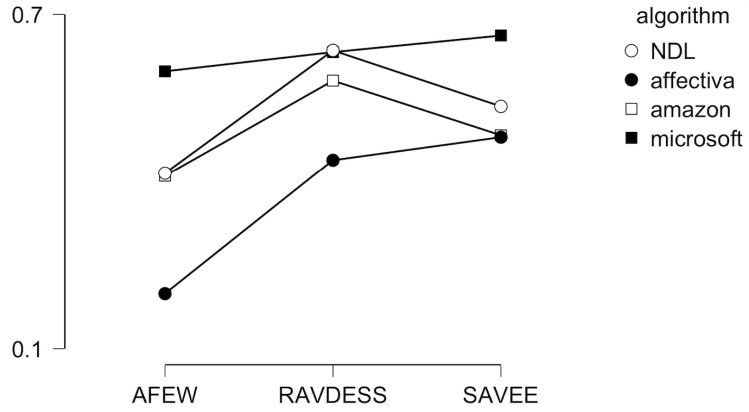


Figure 1: A descriptive plot of significant interaction between the datasets (AFEW, RAVDESS, SAVEE) and the algorithms' accuracy scores (Affectiva Affdex, Microsoft Face Emotion Recognition, Amazon Rekognition Facial Analysis, and Neurodata Lab Emotion Recognition). The ordinate scale corresponds to accuracy scores.

We established the main effect of dataset type on recognition accuracy F(2, 7317) = 113.92, p < .001, η_p^2 = .033, with AFEW containing the least detectable emotions (M = 0.40, 95% CI [0.38, 0.42], p < .001), and RAVDESS containing the most detectable emotions M = 0.57, 95% CI [0.56, 0.59], p < .001. Consistent with the results of other studies [3], accuracy scores for emotion labels were higher for acted facial expressions (RAVDESS M=0.57, 95% CI [0.56, 0.59]; SAVEE M = 0.54, 95% CI [0.52, 0.56])in comparison with more challenging 'in the wild' expressions (AFEW M =0.40, 95% CI [0.38, 0.42], p < .001). Significant interaction was revealed between algorithm type and dataset type F(6, 7317) = 7.45, p < .001, η_p^2 = .007. Compared to other algorithms, Microsoft Face Emotion Recognition showed better performance when analyzing 'in the wild' AFEW dataset (Fig. 1). For AFEW dataset in general, the spread of values in recognition accuracy among algorithms is large (Microsoft M=0.56, 95% CI [0.53, 0.59]; Neurodata Lab M=0.41, 95% CI [0.38, 0.44]; Amazon M=0.39, 95% CI [0.36, 0.42]; Affectiva M=0.25, 95% CI [0.20, 0.29]), while the spread of values in recognition accuracy for RAVDESS and SAVEE is smaller.

Significant main effect of emotion label on recognition accuracy (F(4, 7317) = 197.69, p < .001, η_p^2 = .093) was found. It shows inconsistency in recognition accuracy of target emotions. Tukey pairwise comparisons confirmed that predictions for Surprise (M=0.72, 95% CI [0.69, 0.74]), Happy (M=0.60, 95% CI [0.57, 0.62]), and Sad (M=0.57, 95% CI [0.55, 0.59]) were the most accurate (p < .001) across all considered datasets and algorithms.

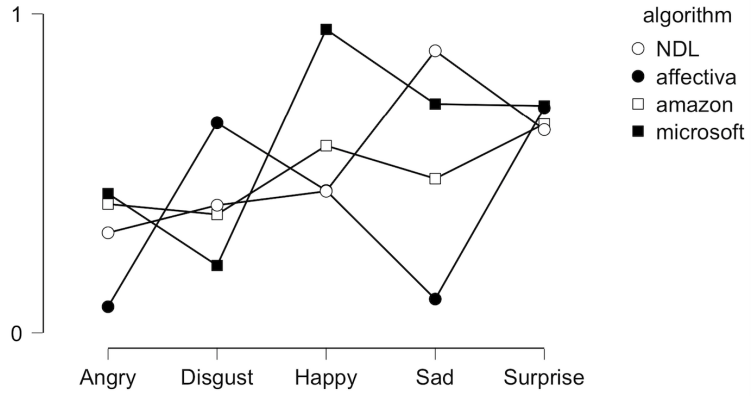


Figure 2: Descriptive plot on significant interaction between ground truth emotion label and accuracy of the algorithms, based on three datasets. The ordinate scale corresponds to accuracy rates.

We found significant interaction between algorithm type and emotion label (F(12, 7317) = 74.47, p < .001, η_p^2 = .125). Further comparisons with Tukey HSD correction showed that the algorithms' performance on specific emotion labels depended on the algorithm considered (Fig.2). Microsoft predicted Happy most accurately (M=0.97, 95% CI [0.92, 1.00), Neurodata Lab showed the highest accuracy with Sad (M=0.88, 95% CI [0.83, 0.92]), while Amazon and Affectiva classified Surprise with the highest accuracy scores (Amazon M=0.71, 95% CI [0.67, 0.76]; Affectiva M=0.71, 95% CI [0.65, 0.76]). In general, algorithms have shown consistency in recognition of Surprise (no significant statistical difference for pairwise comparisons between algorithms was found).

4.2 Generic performance levels for emotion recognition accuracy compared to ground truth

In order to compare the predictions of algorithms to the ground truth data (as provided in datasets), emotion recognition performance per video was analysed with the use of confusion matrices. All emotion labels retrieved from three datasets, as described in Table 1, were used for building the confusion matrices. Macro-average F1-score metric was chosen, because the sample of videos is unbalanced (regarding the number of videos representing each class), and macro-average F1-score allows to counterbalance the contribution of each class. F1-score combines recall and precision and distributes equal weights to each of them. Table 3 presents macro-average F1-scores achieved by every algorithm. Among the datasets, RAVDESS acquired the highest scores for recognition and was followed by SAVEE. Among the algorithms, Microsoft reached the highest scores for AFEW and SAVEE, while Neurodata Lab showed best performance with RAVDESS.

Table 3: F1-scores obtained by comparing every algorithm's predictions and ground truth

F1 score	Neurodata Lab	Affectiva	Amazon	Microsoft
AFEW	0.26	0.09	0.28	0.36
RAVDESS	0.49	0.26	0.39	0.33
SAVEE	0.32	0.24	0.25	0.42

4.3 Agreement among algorithms

4.3.1 Spearman Rank Correlation Coefficient as a metric for the agreement between the algorithms' predictions

Since algorithms were trained on different data and every algorithm uses its own scale to make predictions, we cannot compare the results of their predictions directly. In turn, we expect every algorithm to highlight certain emotion labels on the same videos. The whole analysis is based on the assumption that the algorithms would predict certain emotion classes in correspondence with the datasets' ground truth. We rank the datasets' video fragments in accordance with predictions for every emotion class. For instance, videos with Happy as ground truth achieve high predicted confidence level for Happy and are placed at the top of the Happy ranking. Videos with Sad as ground truth achieve low confidence levels for Happy and are placed at the bottom of the Happy ranking. First we created such a ranking for each emotion class respectively, and then for each algorithm respectively. Thus, the ranking of videos per each emotion label was considered. Then the rankings built for each algorithm were compared with each other. We applied Spearman Rank Correlation with Holm-Bonferroni sequential correction to the rankings of each emotion label. Spearman Rank Correlation Coefficient (SRCC) was calculated for Anger, Disgust, Happiness, Sadness, and Surprise, as these emotion labels are present among four considered algorithms.

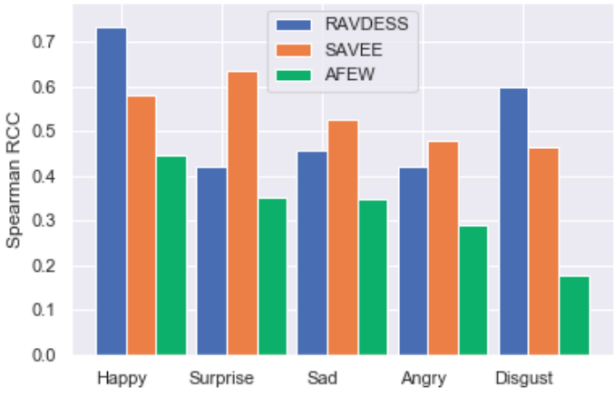
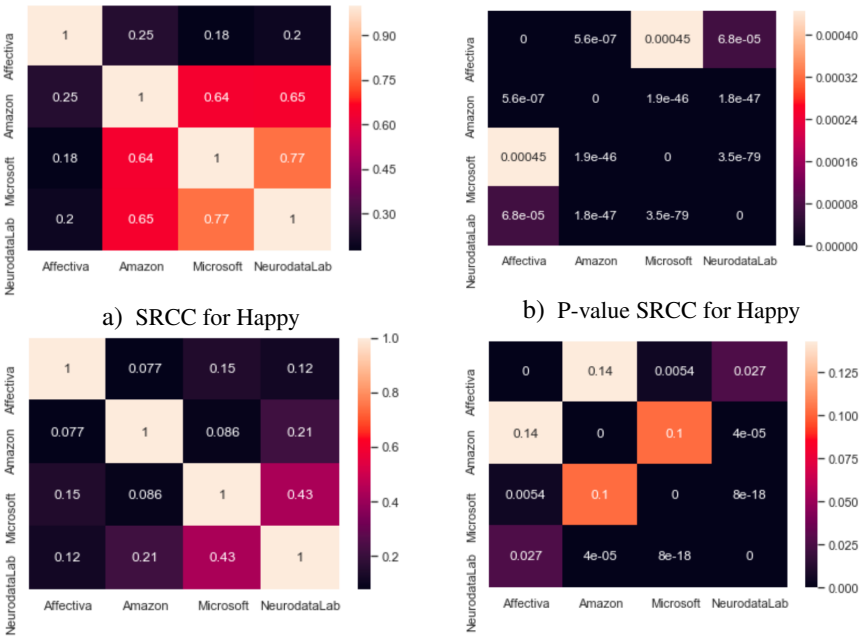


Figure 3: Barplot for emotion labels on mean SRCC. Each bar represents mean Spearman Rank Correlation Coefficient achieved among Neurodata Lab, Affectiva, Amazon, and Microsoft algorithms in 6 pairwise comparisons.

If emotion labels are predicted correctly, then the rankings of videos for each emotion would be highly correlated for different algorithms. In general, all pairwise SRCCs were statistically significant and positively related. There was a considerable overlap between predicted confidence levels for emotions from RAVDESS and SAVEE datasets correspondingly, suggesting that recognition patterns were similar for datasets with high control over data acquisition (RAVDESS r_{mean} = .53; SAVEE r_{mean} = .54). Correlations between videos ranked by confidence level were weaker for 'in the wild' AFEW dataset (r_{mean} = .32). If more complex data is used for analysis, then each algorithm demonstrates more specific pattern of recognition.

If we consider distinct emotion labels, the agreement patterns are dataset-depending (See Figure 3). Happy showed the highest agreement for RAVDESS (r = .73, p <.001) and AFEW (r = .45, p <.001), while Surprise showed the highest agreement in SAVEE (r = .64, p <.001). Happy and Surprise have obtained highest accuracy of recognition when compared with ground truth, and SRCC scores for these two labels have shown the highest agreement between all four algorithms. A large variance of agreement scores was revealed for Disgust (RAVDESS r = .60, p <.001; SAVEE r = .46, p <.001; AFEW r = .18, p <.001). This indicates that the algorithms have performed inconsistently with Disgust depending on dataset. In total, moderate or low consistency of SRCC scores is observed for all emotion labels depending on the dataset. Thus, performance significantly depends on the type of the data.



a) SRCC for Happy

b) P-value SRCC for Happy

c) SRCC for Disgust

d) P-value SRCC for Disgust

Figure 4: Spearman Rank Correlation is depicted here as a measure of agreement using the example of AFEW dataset. See the examples of strong (Pic. a) and weak (Pic. c) correlations between predicted confidence levels for emotional labels. Significance level is set at .01.

We have explored predictions for AFEW dataset in detail, as AFEW is challenging for recognition (M =.40 for accuracy). Figure 4 shows two emotion labels, Happy and Disgust, that are different in relation to agreement. Correlation scores achieved for rankings of confidence levels are presented with corresponding significance values. Happy has reached stronger correlations (r_{mean} = .45) between algorithms, while Disgust has reached weaker correlations between algorithms (r_{mean} = .18). At the same time, Happy has achieved higher accuracy (M=.42, 95% CI [0.39, 0.46]), in contrast Disgust has achieved lower accuracy (M=.18, 95% CI [0.14, 0.23]). Therefore we may assume, if emotion is accurately recognized, its rankings of confidence levels are correlated stronger.

SRCC scores were statistically significant for all five emotion labels (Happy, Disgust, Angry, Sad, Surprise) only between Microsoft and Neurodata Lab as shown in Figure 4. Other pairwise comparisons of algorithms contained at least one statistically insignificant result for emotion labels. Mean SRCC as a measure of ranking agreement (based on distribution of videos by target emotion ranks) was found for Microsoft Face Emotion Recognition and Neurodata Lab Emotion Recognition ($r_{mean} = .54$). The relative performance of algorithms contains a certain level of disagreement if we assess datasets with low control over data acquisition.

4.3.2 F1 Score as a metric of algorithms' predictions agreement

To evaluate the agreement between predictions of selected algorithms, emotion recognition performance per video was analysed using the metric of macro-average F1-score (see Fig. 5, 6, 7). Predictions of one algorithm (we consider prediction here as one emotion label per video which has got highest probability level) were considered as ground truth, while predictions of another algorithm were considered as predicted emotion labels, pairwise. Predictions for five emotion labels were used per algorithm: Angry, Disgust, Happy, Sad, Surprise.

Higher agreement was achieved pairwise for Microsoft - Neurodata Lab (f1=.56 for AFEW, f1=.62 for RAVDESS, f1=.39 for SAVEE), Amazon - Neurodata Lab (f1=.47 for AFEW, f1=.5 for RAVDESS, f1=.45 for SAVEE), and Microsoft - Amazon (f1=.53 for AFEW, f1=.56 for RAVDESS, f1=.28 for SAVEE). In contrast, f1-scores were low for Affectiva. Agreement drops noticeably for Affectiva with more complex data to analyze. Agreement was more sustainable for Microsoft, Amazon, and Neurodata algorithms depending on the complexity of data.

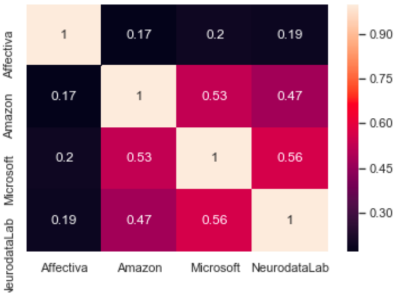


Figure 5: AFEW F1-scores

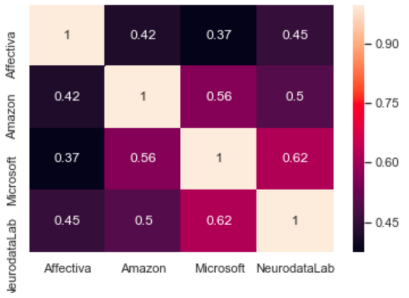


Figure 6: RAVDESS F1-scores

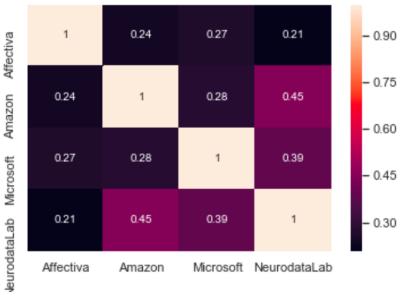


Figure 7: SAVEE F1-scores

5 Discussion

This work is dedicated to monitoring of the advancements in automatic emotion recognition solutions. Initially, we investigated how different commercial algorithms analyzed the publicly available datasets that had different control over data acquisition. In present research, we have compared the predictions of Affectiva's, Amazon's, Microsoft's and Neurodata Lab's emotion classification algorithms with ground truth data (datasets AFEW, SAVEE, and RAVDESS) and evaluated the degree of agreement between the predictions of these four algorithms. We also tested whether the agreement between the algorithms' predictions could be a relevant criteria to evaluate the performance of automatic emotion recognition.

Among the four algorithms, we observed onsiderable variance in recognition accuracy (ranging from .39 to .62), where Microsoft Face Emotion Recognition and Neurodata Lab Emotion Recognition have outperformed Affectiva Affdex and Amazon Rekognition Facial Analysis algorithms. The performance of these four algorithms in our research is consistent with previous studies [20, 8]. Thus, the performance of advanced algorithms depends on the algorithm considered.

We revealed the effect of dataset type on algorithm performance. Prediction scores (accuracy and f1-score) were higher for acted facial expressions compared to the more challenging 'in the wild' expressions and environmental noise. RAVDESS played by actors has achieved highest recognition accuracy and f1-scores, even while being a newly released dataset. SAVEE that contains the recordings of non-professional actors was less detectable. Algorithms showed similar recognition patterns for acted data with high control over conditions of data acquisition. If more complex data was used for analysis, than every algorithm demonstrated more specific pattern of recognition. When algorithms performed on 'in the wild' AFEW dataset, they gave a wide variation of recognition accuracy (from .25 for Affectiva to .56 for Microsoft). Yet algorithms have shown more homogeneous scores for accuracy performing with RAVDESS and SAVEE, which contained laboratory controlled data.

The recognition of Surprise and Happy emotion classes were the most accurate for all algorithms and datasets, whereas additional pairwise comparisons suggested algorithm-specific recognition of certain emotion labels over others. The moderate classification performance was achieved for Sad, and weaker performance was achieved for Anger and Disgust. The recognition of particular emotions was dataset-specific and algorithm-specific.

The similarity of performance patterns was examined using agreement metrics. If an emotion label was predicted accurately (compared to human-labeled ground truth), its predictions were highly correlated for considered algorithms. Correlations for Angry, Disgust, Happy, Sad, and Surprise were strong for RAVDESS and SAVEE, suggesting that recognition patterns were similar for datasets with high control over data acquisition. F1-scores as metrics of agreement were strong or moderate between Amazon's, Microsoft's, and Neurodata Lab's algorithms, and tended to be lower for Arrectiva's pairwise. The highest agreement was achieved for Microsoft - Neurodata Lab pairwise when videos from all datasets were taken for analysis. The relative performance of algorithms contained a certain level of disagreement when we assessed the AFEW dataset with low control over data acquisition.

Agreement scores between algorithms might be an informative criterion. Firstly, the algorithms with higher accuracy and f1-scores that were obtained for human-labeled ground truth (Amazon's, Microsoft's, and Neurodata Lab's), showed higher agreement between their predictions. Affectiva's algorithm was the least accurate, and was also in low agreement with other algorithms (as shown at Figure 1, which represents the accuracy of algorithms per dataset, and Figures 5, 6, 7, which describe the agreement between algorithms per datasets measured as f1-scores). Secondly, accuracy that algorithms reached for ground truth drops when we use more complex data, the same as agreement between algorithms decreases when we test algorithms on more complex data. For more accurate algorithms (Microsoft's, Neurodata Lab's, and Amazon's), agreement between their predictions remained similar for laboratory controlled data and for 'in-the-wild' data.

6 Conclusions

Current state-of-the-art machine learning technologies technically allow us to achieve a very high recognition accuracy for different objects and classes, including the patterns of facial muscle movements. However, considering emotional facial expressions, the achievement of high accuracy meets a certain methodological challenge: that is the task of annotating affective data. The general level of inter-observer agreement reported in affective computing studies is 0.39 kappa [21]. Annotation agreement for 'in the wild' data tends to be lower [22]. Observer-based annotations are expensive and time consuming, and sometimes inefficient. For human-labeled data, the solution could be found in searching for a balance between approximated ground truth and accuracy of recognition. The use of accurate algorithms instead of manual human labeling is interesting since the agreement between several algorithms represents a cumulative agreement of many annotators per each algorithm. We have compared the predictions of four commercial algorithms with each other to see if the agreement between these classifiers helps to solve the problem of obtaining emotional labeling. The achieved agreement level ranged from .42 to .73 for emotion classes in acted datasets, and from .18 to .45 for 'in the wild' dataset, showing that the predictions were similar to a certain extent despite the different learning history of the algorithms. This suggests that particular amount of data may be annotated automatically. Since more accurate algorithms have shown similar predictions, then accuracy and agreement might be potentially related, and agreement between the algorithms' predictions could be a relevant criteria to evaluate the performance of automatic emotion recognition. In order to make automatic annotation effective in practie, algorithms could be chosen considering two conditions: a) algorithms have high accuracy that was obtained for human-labeled ground truth, and b) algorithms' predictions have high agreement with each other.

References

[1] Peter Lewinski, Tim M den Uyl, and Crystal Butler. Automated facial coding: Validation of basic emotions and facs aus in facereader. *Journal of Neuroscience, Psychology, and Economics*, 7(4):227–236, 2014.

[2] Damien Dupré, Nicole Andelic, Gawain Morrison, and Gary McKeown. Accuracy of three commercial automatic emotion recognition systems across different individuals and their facial expressions. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 627–632. IEEE, 2018.

[3] Sabrina Stöckli, Michael Schulte-Mecklenbeck, Stefan Borer, and Andrea C Samson. Facial expression analysis with affdex and facet: A validation study. *Behavior research methods*, 50(4):1446–1460, 2018.

[4] Alec Burmania and Carlos Busso. A stepwise analysis of aggregated crowdsourced labels describing multimodal emotional behaviors. In *INTERSPEECH*, pages 152–156, 2017.

[5] Brendan D Murray, Alisha C Holland, and Elizabeth A Kensinger. Episodic memory and emotion. 2013.

[6] Jenna L Cheal, Jennifer J Heisz, Jennifer A Walsh, Judith M Shedden, and MD Rutherford. Afterimage induced neural activity during emotional face perception. *Brain research*, 1549:11–21, 2014.

[7] Alison M Mattek, Paul J Whalen, Julia L Berkowitz, and Jonathan B Freeman. Differential effects of cognitive load on subjective versus motor responses to ambiguously valenced facial expressions. *Emotion*, 16(6):929–936, 2016.

[8] Damien Dupré, Eva Krumhuber, Dennis Küster, and Gary McKeown. Emotion recognition in humans and machine using posed and spontaneous facial expression. 2019.

[9] Björn W. Schuller. *Acquisition of Affect*, pages 57–80. Springer International Publishing, Cham, 2016.

[10] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008.

[11] Yue Zhang, Eduardo Coutinho, Björn Schuller, Zixing Zhang, and Michael Adam. On rater reliability and agreement based dynamic active learning. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 70–76. IEEE, 2015.

[12] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 2017.

[13] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34–41, 2012.

[14] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Acted facial expressions in the wild database. *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, 2, 2011.

[15] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5), 2018.

[16] S. Haq and P.J.B. Jackson. *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pages 398–423. IGI Global, Hershey PA, 2010.

[17] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 3723–3726. ACM, 2016.

[18] Amazon rekognition documentation. https://docs.aws.amazon.com/rekognition/index.html.

[19] Dhwani Mehta, Mohammad Siddiqui, and Ahmad Javaid. Facial emotion recognition: A survey and real-world user experiences in mixed reality. *Sensors*, 18(2):416, 2018.

[20] De'Aira Bryant and Ayanna Howard. A comparative analysis of emotion-detecting ai systems with respect to algorithm performance and dataset diversity. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 377–382. ACM, 2019.

[21] Sidney K D'Mello. On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing*, 7(2):136–149, 2015.

[22] Hongying Meng, Andrea Kleinsmith, and Nadia Bianchi-Berthouze. Multi-score learning for affect recognition: the case of body postures. In *International Conference on Affective Computing and Intelligent Interaction*, pages 225–234. Springer, 2011.