Accuracy of three commercial automatic emotion recognition systems across different individuals and their facial expressions

Damien Dupré
Queen's University Belfast
Belfast, United Kingdom
damien.dupre@qub.ac.uk

Nicole Andelic
Sensum Ltd.
Belfast, United Kingdom
nandelic@sensum.co

Gawain Morrison
Sensum Ltd.
Belfast, United Kingdom
gawain@sensum.co

Gary McKeown
Queen's University Belfast
Belfast, United Kingdom
g.mckeown@qub.ac.uk

***Abstract*—Automatic facial expression recognition systems can provide information about our emotions and how they change over time. However, based on different statistical methods the results of automatic systems have not yet been compared. In the current paper we evaluate the emotion detection between three different commercial systems (*i.e.* Affectiva, Kairos and Microsoft) when detecting dynamic and spontaneous facial expressions. Even if the study was performed on a limited sample of videos, the results show significant differences between the systems for the same video and per system across comparable facial expressions. Finally, we reflect on the implications according the generalization of the results provided by automatic emotion detection.**

Keywords—emotion, facial expression, automatic recognition

I. INTRODUCTION

Facial expressions are displays used to regulate social interaction. They provide information to others about inner cognitive appraisals (*e.g.* relevance or novelty of an event), about inner action tendencies (*e.g.* approach or avoidance tendencies) and social messages to others (*e.g.* willingness to make contact or to be aggressive) [1]. However, they are mainly used by others to infer the emotional state of a person. For example, a smile will be used to infer that a person is happy, on the contrary a frowning will be used to infer that a person is angry [2]. As such, facial expressions provide meaningful cues to social interaction.

One way of identifying the facial movements that are interpreted as emotions is the Emotional Facial Action Coding System (EmFACS, [3]) which associates Action Units (*i.e.* movement involving the minimal number of facial muscles) with six prototypical facial expressions of emotion. These are happiness, fear, anger, surprise, disgust and sadness. The EmFACS is not only a tool used to analyse facial expressions but it is also the starting point of the development of automatic facial expression recognition systems.

A. Current emotion detection systems

Various automatic facial expression recognition systems have been developed in order to detect people's emotions (see [4] for review). Most of them are developed for academic research only such as OpenFace [5] or IntraFace [6] but some systems are used for commercial purposes [7] (Table 1).. The commercial applications of detecting people's emotions from their facial expression are multiple. Entertainment, advertising, automotive or health care are examples of sectors in which emotion detection can be used to evaluate and predict people's emotional states.

TABLE 1. COMPANIES THAT PROVIDE AUTOMATIC FACIAL EXPRESSION RECOGNITION SYSTEMS (NON EXHAUSTIVE LIST).

Companies	Websites
Affectiva	www.affectiva.com
CrowdEmotion	www.crowdemotion.co.uk
Emo. Research Lab	www.emotionresearchlab.com
Eyeris	www.emovu.com
EyeSee	www.eyesee-research.com
GraphEQ*	www.grapheq.com
Kairos	www.kairos.com
Microsoft Azure	www.azure.microsoft.com
MoodMe	www.mood-me.com
Noldus	www.noldus.com
Nviso	www.nviso.ch
RealEyes	www.realeyesit.com
RefineAI	www.refineai.com
Seeing Machines*	www.seeingmachines.com
SightCorp	www.sightcorp.com
Visage Technology	www.visagetechnologies.com

NOTE - * Companies are not measuring emotional states but only cognitive and physical states such as attention or fatigue. The companies previously known as Emotient and Faciometrics have been removed from this list as they are now part of Apple and Facebook respectively. Companies focusing only on emotion detection in pictures were not added to the list.

Four different business models distinguish between the commercial providers of facial expression recognition systems. Some companies are Application Program Interface (API) oriented: They provide a web interface in which videos can be uploaded for analysis, and results can be downloaded. Other companies also provide a Software Development Kit (SDK) which can be embedded in a larger system and processes the emotion detection locally. Another commercial

method is the selling of local or web hosted software license. Finally, a fourth approach is service oriented as videos are manually sent to the company for internal processing.

B. *Automatic emotion detection procedure*

Most automatic facial expression recognition systems detect emotions through a three-stage analysis [8] (Fig. 1).

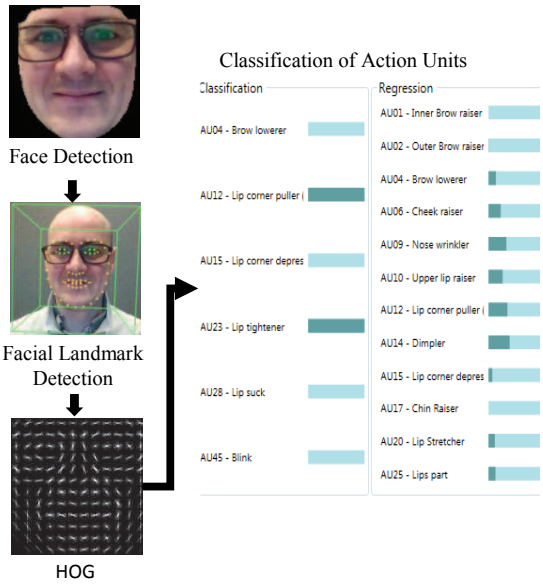

Fig. 1. Example of automated FACS coding analysis using facial landmark detection with Histograms of Oriented Gradients (HOG) features using OpenFace v0.1 [5].

The first step is to detect the face, more specifically to identify all the regions in a video that contains a face. A face is a very specific object to detect thanks to its shape, its color distinction and its characteristic organization (most of them have two eyes, a nose and mouth). Therefore, computational algorithms are trained to quickly detect faces from the background.

The second step is the facial landmark detection based on geometric features. Once the face is detected, the systems identify facial landmarks and measure their changes over time. Different facial landmark classifications are possible depending on the complexity of the model used. A classic model provided by Active Shape Models (stASM) is identifies 77 facial landmarks [9]. Another method to identify the movement of the facial landmarks is by using appearance features analysis. Statistical techniques such as Gabor wavelets filtering, Eigenface, or Active Appearance Model capture the changes in the face.

The third step is facial movement classification for emotion detection. The systems are trained on video databases made of dynamic emotional facial expressions as tagged by annotators. Facial landmark movements are classified using statistical treatments such as k-Nearest Neighbor, Neural Networks, Support Vector Machine, Bayesian Networks, AdaBoost or Hidden Markov Models to name a few. Examining the association between specific facial movements and the emotion recognized allow the systems to generalize emotion detection to new faces.

C. *Accuracy of automatic emotion detection*

Although most of the companies offering facial coding are supported by empirical research and publications, the published system evaluations are generally presented in "white papers" (*i.e.* publications that do not have the rigorous peer-review process typical of academic journals). Consequently, there is a need for independent, academic validation of these systems using objective criteria.

However, determining objective and relevant criteria to evaluate automatic emotion detection accuracy remains a challenge due to the dynamic and spontaneous nature of human emotions. Contrary to emotion detection in static stimuli (*e.g.* in pictures) that can be assessed with Confusion Matrices, emotion detection of dynamic stimuli involves changes over time that must be taken into account.

In order to overcome this problem there are methodologies that summaries the time-series by performing Confusion Matrices frame-by-frame, resulting in a Matching Score percentage [10]. However, because the level of emotion will change continuously in dynamic stimuli, criteria that takes the temporality of facial expression into account is necessary.

One way to assess emotion detection accuracy is to evaluate the Precision and Recall scores. These indicators are classic pattern recognition measures calculated to evaluate the rate of false and true positives as well as false and true negatives [11]. Precision (also called positive predicted value) is the ratio of the correctly detected events among all the events detected. Recall (also called sensitivity) is the ratio of the correctly detected event among all the events that happened. Based on statistical measures of the performance of a binary classification test, we propose to use an alternative version of the Precision and Sensitivity indicators [11] by adapting *True Positive*, *False Positive*, *True Negative* and *False Negative* rates to evaluate emotion detection accuracy of dynamic stimuli [12].

II. Methodology

In this study we compared three commercial automatic facial expression recognition systems (Affectiva Affdex SDK v3.4.1, Microsoft Project Oxford v1.0, and Kairos API v2.0). The obtained results are related to the version of the system used and the results might be different when using a newer release. These systems were chosen because they provide emotion detection of six similar labels across the three systems (i.e. *Anger*, *Disgust*, *Fear*, *Joy*, *Sadness*, and *Surprise*). Each company has granted their permission for these results to be presented.

A. *Stimuli Selection*

The emotion detection accuracy of the systems was compared using dynamic and spontaneous facial expressions

of emotion taken from the DynEmo [1] database [13]. The DynEmo database contains 358 natural facial expressions displaying affective states rated by both the expresser and naive annotators (with a 768 x 576 pixel resolution and 25 fps). From this database, we have selected eight videos (10 seconds long) labelled by annotators as displaying one of the following emotions: Disgust, Fear, Joy, and Surprise. These videos were chosen because they had the highest annotator recognition agreement in displaying the target labels. Each emotion is displayed by both a male and a female.

B. Emotion detection

The videos were processed with our three selected automatic facial expression recognition systems. As each system has a different data range, all data output was rescaled from 0 to 1.

We compared the accuracy of each of these systems by processing the videos using the following emotional labels: *Anger*, *Disgust*, *Fear*, *Joy*, *Sadness*, and *Surprise*. Even though some of the facial coding providers offer recognition of further emotions, we did not include them.

After the rescaling of the data, scores higher than 0.5 (*i.e.* 50% of detection probability) were considered to be an emotion detected (and recoded as 1) whereas scores lower than 0.5 were considered to be an emotion not detected (and consequently recoded as 0). Because the stimuli used are dynamic and spontaneous facial expressions, we expect to obtain a significant level of noise from the data recorded by the systems. Therefore, this arbitrary threshold functions as a way to not only suppress part of the false recognition noise but also as a way to provide a flexible correct recognition rate.

In order to compare the systems we first calculated their ratio of *True Positive* (only the target label is recognized), *False Positive* (the target label as well as a non-target label is recognized), *True Negative* (no label was recognized) and *False Negative* (target label was not recognized but a non-target label was).

III. Results

A. Overall emotion detection accuracy

Based on rescaled raw data (Fig. 2), emotion detection values are considered as recognized if they reach the 0.5 threshold. Each system's proportion of *True Positive*, *False Positive*, *True Negative* and *False Negative* emotion detection for all the videos is presented in Table 2.

The results of the comparison between the systems show that they have comparable detection rates in term of *True Positive* and *False Positive*. However, their detection of *False Negative* and *True Negative* differs between the facial coding systems. Affectiva and Microsoft's systems both show a significantly higher tendency to not detect non-target labels erroneously, *i.e.* to not detect non-target emotion labels when the target label is detected (*False negative*) as well as to not detect non-target emotion label when the target label is not detected (*True Negative*). In contrast, Kairos showed higher levels of false detection of emotions.

TABLE 2. MEAN PROPORTION OF TRUE POSITIVE, FALSE POSITIVE, TRUE NEGATIVE AND FALSE NEGATIVE EMOTION DETECTION FOR ALL THE VIDEOS.

System	False Negative	False Positive	True Negative	True Positive
Affectiva	0.03 *(0.02)*	0.05 *(0.04)*	0.74 *(0.11)*	0.18 *(0.10)*
Kairos	0.40 *(0.16)*	0.04 *(0.03)*	0.43 *(0.15)*	0.14 *(0.12)*
Microsoft	0.10 *(0.09)*	0.00 *(0.00)*	0.67 *(0.15)*	0.23 *(0.15)*

NOTE - Standard error in brackets

B. Emotion detection accuracy per video

When examining emotion recognition accuracy for each video/emotion, the video with a higher accuracy (*i.e. True Positive*) are the videos of a joyful facial expression. The other videos result in a proportion of target emotion detection statistically equal or lower than the detection of non-target emotion.

In order to evaluate whether the type of video and the facial coding provider have an influence on the accuracy measurements, we fitted Generalized Linear Models (GLM) with a binomial distribution [14]. The results from the GLMs for each accuracy measurement are displayed in Table 3, 4, 5 and 6.

TABLE 3. GLM INCLUDING VIDEO AND SYSTEM FOR TRUE NEGATIVE.

	Df	Dev.	AIC	LRT	Pr(>Chi)
intercept		3241	3289		
video	7	4160	4194	919	< 0.001***
system	2	3293	3337	52	< 0.001***
video:system	14	6165	6185	2924	< 0.001***

NOTE - Signif. codes: *** 0.001 ** 0.01 * 0.05

TABLE 4. GLM INCLUDING VIDEO AND SYSTEM FOR TRUE POSITIVE.

	Df	Dev.	AIC	LRT	Pr(>Chi)
intercept		1501	1549		
video	7	2463	2497	962	< 0.001***
system	2	1576	2359	75	< 0.001***
video:system	14	3630	3650	2128	< 0.001***

NOTE - Signif. codes: *** 0.001 ** 0.01 * 0.05

TABLE 5. GLM INCLUDING VIDEO AND SYSTEM FOR FALSE POSITIVE.

	Df	Dev.	AIC	LRT	Pr(>Chi)
intercept		766	814		
video	7	1107	1141	342	< 0.001***
system	2	900	944	134	< 0.001***
video:system	14	797	817	32	0.004**

NOTE - Signif. codes: *** 0.001 ** 0.01 * 0.05

TABLE 6. GLM INCLUDING VIDEO AND SYSTEM FOR FALSE NEGATIVE.

	Df	Dev.	AIC	LRT	Pr(>Chi)
intercept		1671	1719		
video	7	1799	1833	128	< 0.001***
system	2	2150	2194	480	< 0.001***
video:system	14	3360	3380	1689	< 0.001***

NOTE - Signif. codes: *** 0.001 ** 0.01 * 0.05

[1] www.dynemo.upmf-grenoble.fr

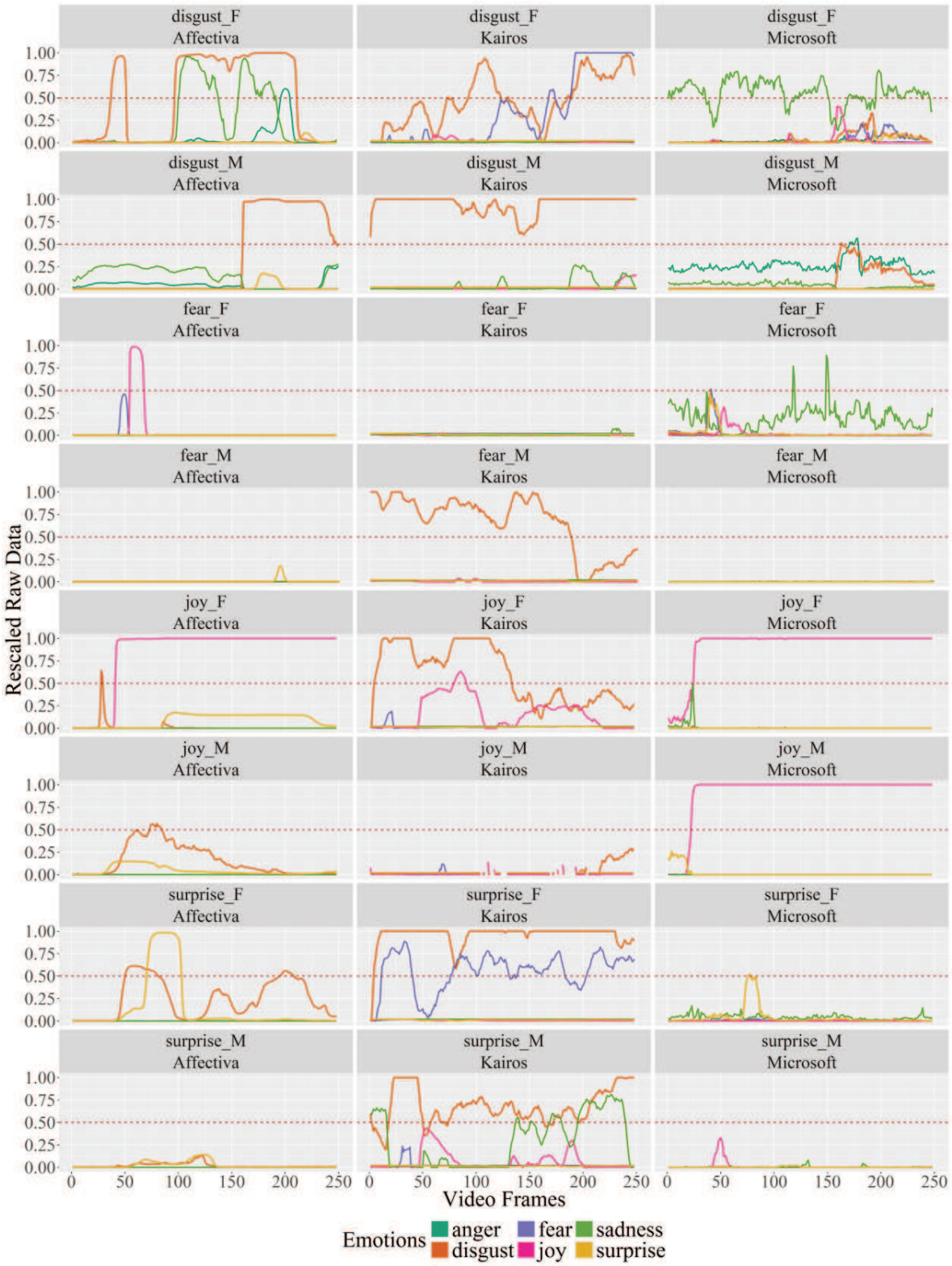

Fig. 2. Evolution of automatic emotion detection according the three different systems and the 8 emotional videos processed. Raw data rescaled from 0 to 1 for comparison. Recognition threshold set to 0.5 (dash line).

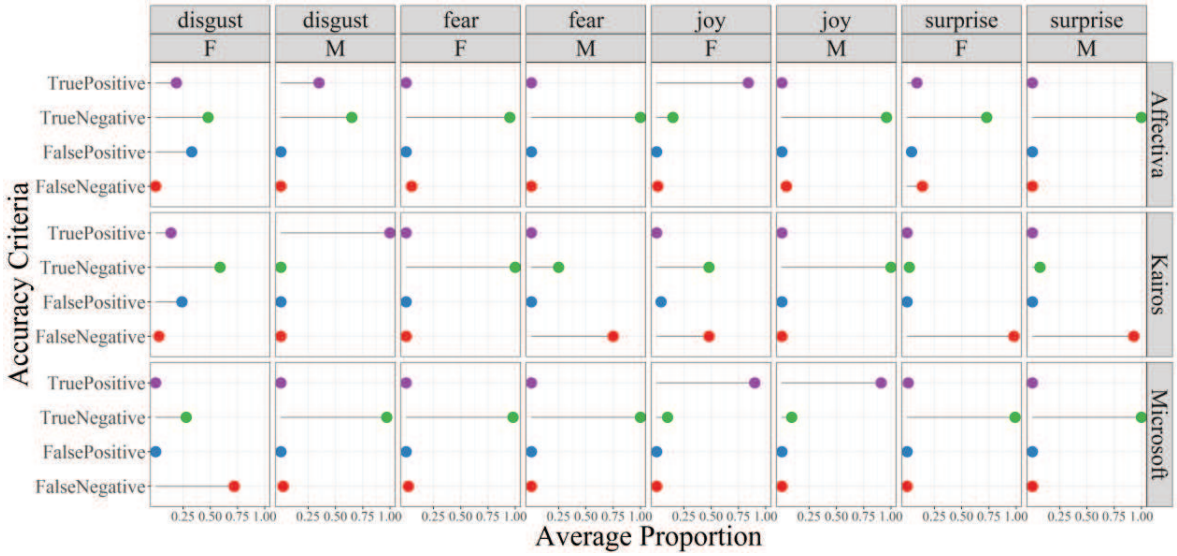

Fig. 3. Overall mean recognition proportions for the different videos tested according their ratio of *True Positive* (only the target label is recognized), *False Positive* (the target label as well as a non-target label is recognized), *True Negative* (no label was recognized) and *False Negative* (target label was not recognized but a non-target label was).

For each of the four accuracy measurements, the GLMs show that both type of video and system has a significant influence on prediction accuracy, as well as the interaction between the variables. This demonstrates that the systems are not performing equally well and that the accuracy also differs depending on the type of video that is processed.

By examining Fig. 3, it is possible to identify some patterns when comparing the videos. For example, the Joy facial expressions seem to be the most accurately recognised out of the emotions in the current study (*True Positive*). However, other facial expressions, such as Disgust and Surprise, were more likely to lead to incorrect recognitions (*False Positive* and *False Negative*).

Finally, it is interesting to see that the accuracy of the system differs depending on the individual expressing the emotion. This findings demonstrates the challenge that these systems have when faced with the idiosyncratic characteristics of spontaneous facial expressions.

IV. Discussion

The result of automatic facial expression recognition systems is a time analysis of the probability that someone is expressing an emotion X_n at the time t_i. However, because the systems are based on a range of statistical methods to detect emotions, the accuracy of different automatic systems have not yet been compared. Nevertheless, comparing the results provided by different systems is crucial. Indeed, the results provided by automatic facial expression recognition tools lead to important decisions that may involve both human and economic resources. It is therefore necessary to compare and evaluate the accuracy of these systems.

Given the dynamic and spontaneous nature of emotional facial expressions, it is difficult to assess the "true" accuracy of the emotion recognition provided by current commercial systems. To compare their accuracy we therefore proposed the use of a version of *True Positive*, *False Positive*, *True Negative* and *False Negative* detection rate.

We found that differences in accuracy were influenced both by the video being processed and the system used to process it. The differences between the systems' accuracy could be due to the way they evaluate the facial movements (facial landmark recognition) as well as how the systems classify emotions from these movements. It is also likely that there are some differences between the recognition of acted and natural expression of emotions. Whereas these systems are trained to recognize emotion in the face from posed facial expression databases (*i.e.* facial expressions displayed by actors in a controlled setting) such as JAFFEE or Cohn-Kanade databases [15], [16], we chose to evaluate them with natural facial expressions. This allowed us to test their accuracy on daily life emotions as seen 'in the wild' but it also decreased the accuracy of the systems.

However, the methodology used in the current study has some limitations. A limited number of videos were tested and it is possible that those selected are not a true representation of "everyday" emotions but are artefacts. This represents a very small sample size as there can be variation in performance based on an individual's appearance, emotion intensity, etc. To overcome this limit, future studies need to include a larger amount of stimuli in order to compare different automatic systems.

Another challenge for future research is the decision of which emotion categories that should be included. Emotion recognition providers offer detection of other emotions such as amused, persuaded, informed, sentimental or inspired [17]. The issue with the six emotions investigated in this study is

the lack of balance when comparing the valence levels (only one is positive, one is neutral and four are negative emotions). In addition, the relevance of such categorical labels is relatively low [18] and their existence 'in the wild' is still questioned.

V. Conclusion

Commercial automatic facial expression recognition systems are powerful tools used to detect people's emotions 'in the wild'. However, the accuracy of these systems remains as an open question. In this paper we compared the emotion detection accuracy of three commercial systems: Affectiva, Kairos and Microsoft. A comparison of their accuracy shows significant differences between the systems. This suggests that they are not equivalent in their ability to detect specific dynamic and spontaneous emotions. Even if automatic facial expression recognition systems are used in various contexts to detect emotions [19], their algorithms still can be improved in order to take into account the idiosyncratic characteristic of emotion expression. Therefore, users of these systems have to be aware of the strength and the potential limits of the data provided by these systems.

VI. Acknowledgments

We thank the representatives of Affectiva, Kairos and Microsoft who agreed for the communication of the data presented in this paper.

VII. References

[1] G. Horstmann, "What do facial expressions convey: Feeling states, behavioral intentions, or actions requests?," *Emotion*, vol. 3, no. 2, pp. 150–166, 2003.

[2] K. R. Scherer and D. Grandjean, "Facial expressions allow inference of both emotions and their components," *Cogn. Emot.*, vol. 22, no. 5, pp. 789–801, 2008.

[3] W. V. Friesen and P. Ekman, "EMFACS-7: Emotional facial action coding system," University of California at San Francisco, 1983.

[4] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, 2010.

[5] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," Winter Conference on Applications of Computer Vision, 2016, pp. 1–10.

[6] F. De la Torre, W.-S. Chu, X. Xiong, F. Vincente, X. Ding, and J. Cohn, "IntraFace," presented at the International Conference on Automatic Face and Gesture Recognition, 2015, pp. 1–8.

[7] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.

[8] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[9] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," European Conference on Computer Vision, 2008, pp. 504–513.

[10] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson, "Facial expression analysis with AFFDEX and FACET: A validation study," *Behav. Res. Methods*, pp. 1–15, 2017.

[11] P. Dente, D. Küster, L. Skora, and E. Krumhuber, "Measures and metrics for automatic emotion classification via FACET," Conference on the Study of Artificial Intelligence and Simulation of Behaviour, 2017, pp. 160–163.

[12] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," International Conference on Machine Learning, 2006, pp. 233–240.

[13] A. Tcherkassof, D. Dupré, B. Meillon, N. Mandran, M. Dubois, and J.-M. Adam, "DynEmo: A video database of natural facial expressions of emotions," *Int. J. Multimed. Its Appl.*, vol. 5, no. 5, pp. 61–80, 2013.

[14] W. N. Venables and B. D. Ripley, "Random and mixed effects," in *Modern Applied Statistics with S. Statistics and Computing*, W. N. Venables and Ripley, Eds. New York, NY, USA: Springer, 2002, pp. 271–300.

[15] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," International Conference on Automatic Face and Gesture Recognition, 2000, pp. 46–53.

[16] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200–205.

[17] D. McDuff, "Discovering facial expressions for states of amused, persuaded, informed, sentimental and inspired," International Conference on Multimodal Interaction, 2016, pp. 71–75.

[18] H. Gunes and H. Hung, "Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block," *Image Vis. Comput.*, vol. 55, no. 1, pp. 6–8, 2016.

[19] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in *Advances in Face Detection and Facial Image Analysis*, M. Kawulok, M. E. Celebi, and B. Smolka, Eds. New York, NY, USA: Springer, 2016, pp. 63–100.