Towards More Robust Automatic Facial Expression Recognition in Smart Environments

Arne Bernin
Department Informatik
Hamburg University of Applied Sciences
Engineering and Computing
University of the West of Scotland
ab@emotionbike.org

Larissa Müller
Department Informatik
Hamburg University of Applied Sciences
Engineering and Computing
University of the West of Scotland
lm@emotionbike.org

Sobin Ghose and Kai von Luck
Department Informatik
Hamburg University of Applied Sciences
Hamburg, Germany
sg@emotionbike.org

Christos Grecos
Computer Science Department
Central Washington University (CWU)

Qi Wang
School of Engineering and Computing
University of the West of Scotland

Florian Vogt
Innovations Kontakt Stelle (IKS) Hamburg
Hamburg University of Applied Sciences

ABSTRACT

In this paper, we provide insights towards achieving more robust automatic facial expression recognition in smart environments based on our benchmark with three labeled facial expression databases. These databases are selected to test for desktop, 3D and smart environment application scenarios. This work is meant to provide a neutral comparison and guidelines for developers and researchers interested to integrate facial emotion recognition technologies in their applications, understand its limitations and adaptation as well as enhancement strategies. We also introduce and compare three different metrics for finding the primary expression in a time window of a displayed emotion. In addition, we outline facial emotion recognition limitations and enhancements for smart environments and non-frontal setups. By providing our comparison and enhancements we hope to build a bridge from affective computing research and solution providers to application developers that like to enhance new applications by including emotion based user modeling.

CCS Concepts

•Computing methodologies → Activity recognition and understanding; •Human-centered computing → Human computer interaction (HCI);

Keywords

Affective Computing, Facial Expression Recognition, Benchmark, Application specific Clustering of Emotions

PETRA '17, June 21 - 23, 2017, Island of Rhodes, Greece

ISBN 978-1-4503-5227-7/17/06…$15.00

DOI: http://dx.doi.org/10.1145/3056540.3056546

1. INTRODUCTION

Building better and richer interactions between computing applications and users is an ongoing challenge of researchers, product designers, and solution providers alike. One promising avenue is for instance from the field of affective computing [27] to build enhanced user models to include application context and emotional response. The aim is to create user tailored and individual system dialogue to adapt level, pace and content to user needs and add responsive content.

Our research focus is to build such systems in smart environments based on facial emotion recognition and other sensors in the context of health and adaptive learning applications. For example, our EmotionBike project [25] is a system for emotion adaptive exergaming and is part of a smart home laboratory [12].

In this context we investigated the impact of provoking emotions in cycling exergames [25], which includes a stationary ergometer with a movable handlebar, that is similar to a cockpit or driver scenario. To sense the user we applied the computer expression recognition toolbox (CERT) [17]. Further we investigated the output of biophysiologial sensors for this application and we found an increase in the overall detection rates for emotion related events [24].

The problem we and many application developer and applied researcher face is that the different facial emotion recognition systems available are difficult to compare due to the lack of benchmarks in particular for non-desktop applications.

This circumstance was our motivation to write this paper in order to provide a comparison of four facial emotion recognition systems. The core of our comparison is a benchmark with three facial expression databases that cover the spectrum from desktop to smart environment situations.

In addition, we propose a new application domain specific enhancement of the recognition rate by means of emotion clustering.

2. RELATED WORK

2.1 Smart Environments and Applications

A classic laboratory setup has one stationary person in front of a desktop computer setup with optimal lighting and camera placement. More challenging are smart environments [6] where the person may be mobile indoor or outdoor and multiple cameras and people may be present. In addition, smart environments often have a wide range of possible face positions and constrains as irregular lighting, non-frontal pose or occlusions which makes it a challenging task for automatic FER [3].

Affective systems using FER have been developed for different applications in a smart environment. D'mello et al. proposed an automated learning tutoring system called AutoTutor [8].

As an additional example for a cockpit scenario similar to our setup, the NAVIEYES system architecture was developed as a lightweight system for an advanced driver assistance system using a dual camera smartphone [22].

Combining different sensors and context is often used in smart environments, Smailis et al. used a fusion of active orientation models and mid-term audio features to detect depression based on the AVEC 2014 depression dataset [28], while Abouelenien et al. detected stress via a combination of physiological signals and thermal imaging used on a self-generated dataset of a stressing task [1]. The STHENOS project [19] focuses on the development of a methodology and an affective computing system for the recognition of physiological states and biological activities in assistive environments which includes a camera based setup, an environment which is similar to our smart home lab.

Kanjo et al. [15] presented a review of different approaches and modalities for emotion recognition in pervasive environments focusing on providing "a platform of understanding for designers, computer scientists, and researchers from other related disciplines". Their article provides a good introduction to the topics.

2.2 Emotional Models and Expressions

According to Calvo et al. [5] six main perspectives exists to describe emotions: "emotions as expressions, emotions as embodiments, cognitive approaches to emotions, emotions as social constructs, neuroscience approaches as well as core affect and psychological construction of emotion."

In this work we focus on the emotions as expressions theory, which is mainly based on Ekman's theory of six basic emotions [10]. Although nowadays the number of basic emotions has been widely increased to seven (anger, fear, disgust, joy, sadness, surprise, contempt), we consider six basic emotions (without contempt) in our approach, since all the utilized algorithms and databases support just six basic emotions. Emotions are often displayed with facial expressions, although facial expressions can be seen as more abstract as they display more than only emotions, like fatigue[1].

A common approach for detecting these expressions is to generate a feature set of facial landmarks or muscle activity [33]. An approach for discrete quantification is Action Units: Action Units (AU) are part of the facial action coding system (FACS) by Ekman and Friesen [10]. They describe a set of activity based on facial muscles. Facial expressions of emotions can be coded based on the presence on these AUs and have often been used for developing FER algorithms [20, 2, 33]. As an example for low level (AU only) detection, Eleftheriadis et al. proposed a multi-conditional learning algorithm based on Bayesian learning in combination with Monte Carlo Sampling [11].

[1]In Ekman's model, fatigue is not an emotion.

2.3 Facial Expression Recognition (FER) Algorithms

2.3.1 General Approach

Many approaches exist for processing images and videos for detecting facial expressions. An overview can be found in the survey of Zeng et al. [33]. According to [33], a common approach for a FER algorithm pipeline is visualized in Figure 1. Finding the face is the first crucial step followed by a step of reducing the data size with filtering. Features are extracted from these reduced data and learning or statistic classification generates the result.

Algorithms may be trained to detect AUs [20] or Facial Landmarks [21] as intermediate step or directly trained for facial expression detection on the raw input [29].

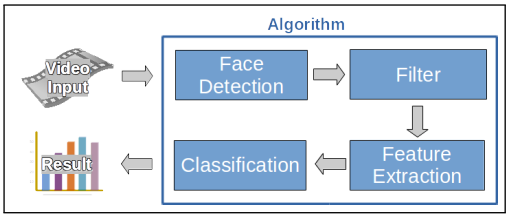

Figure 1: General structure of facial expression recognition algorithms according to [33].

While many algorithms have been proposed for research, we focus on algorithms available as commercial or non-commercial software that is ready to be used in productive applications.

2.3.2 Face Detection and Pose

The first step of FER, as shown in Figure 1, is detecting the face. The standard algorithm for this purpose is Viola-Jones [30] for frontal or near frontal images which is used in different modifications by all four FER algorithms we tested for this work. For non-frontal setups as bigger smart environments like smart homes, Brauer et al. proposed a face dection algorithm specially designed for 180 degree fisheye cameras to cover a wide area from the ceiling [4]. Detecting the silhuette as pre-stage to decting of the face, Maglogiannis et al. provided another work with this type of camera [19]. A detailed overview on recent face detection algorithms can be found in the survey of Zafeiriou [32].

2.3.3 Algorithms for Realtime Analysis

In an interactive setup, processing of the facial expressions needs to be in realtime to minimize the delay of an appropriate response. We evaluated the following four algorithm for this work:

Affectiva or Affdex SDK [20] uses a pipeline of face detection (using Viola-Jones) and localization of the key facial landmarks on each face. Extraction of texture features using histogram of oriented gradients is followed by a clas-

Table 1: Benchmarked algorithms for emotional facial expressions. The (recognized) six basic emotions are highlighted for the labeled expressions.

Software	Platforms	Interface	Labeled Expressions	Output
Emotient	Windows	IMOTIONS API/UDP	**Anger, Joy, Surpise, Sad, Disgust, Fear**, Contempt, Confusion, Frustration, Neutral	Evidence (∼ -10 - 10)
Affectiva (Affdex)	Linux, Windows, Mobile	C++, C#	**Anger, Joy, Surpise, Sad, Disgust, Fear**, Contempt, Smirk, Smile	Probability (0-100)
InSight	Linux, Windows, Mac OS, Mobile	C++	**Anger, Joy, Surpise, Sad, Disgust, Fear**, Neutral	Probability (0.0-1.0)
CERT	Mac OS	Manual csv output	**Anger, Joy, Surpise, Sad, Disgust, Fear**, Contempt, Neutral, Smile	Evidence (∼ -10 - 10)

sification of AUs using trained suppport vector machines (SVM). Modeling of prototypic emotions is achieved using EMFACS [13] based on the AUs.

InSight[2], to our knowledge, has not published documentation describing their approach. However, from the license information they credit OpenCV for the face detection which is also based on Viola-Jones. It is unpublished which feature extraction and machine learning algorithm(s) are applied.

CERT [2] implements an extended Viola-Jones algorithm approach for face detection followed by Gabor filtering for feature extraction. These features are fit into a set of linear SVMs for each AU. A second layer of SVMs is used to detect the facial expressions. In contrast to the three other algorithms, CERT does process the data near realtime but lacks an appropiate output interface to incorporate a live application.

Emotient is the successor of CERT and uses the same basic algorithms [23]. Unlike the method used in CERT, the facial expressions are not based on the output of the AUs but seperatly trained on the feature data. Emotient was aquired by another company that ended its availability as a seperate product. However, it is still commercially available as part of the IMOTIONS platform[3].

Further details on the evaluated algorithms are provided in Table 1.

2.4 Testing and Benchmarking

Performance evaluation with databases is a well established approach to benchmark FER algorithms. Many different databases for labeled facial expressions are available to the research community. The surveys of Zeng [33] and D'Mello [9] provide a comprehensive overview. The databases differ in modalities: Static images, 2D-videos [18] or even 3D-videos [31].

Benchmarking of FER algorithms is often performed either in the context of challenges on certain datasets, like [29], or for measuring the performance of a newly developed algorithm compared to alternative approaches [17]. Typically only one dataset is selected in the first case, while two or more datasets (with CK+ often used as reference set) in the second. Apart from CK+, the selection of databases often differs, which make a general comparison of algorithms difficult. We chose three databases as a core dataset to reflect important characteristica for frontal, non-frontal and smart environments.

[2]http://sightcorp.com/insight/
[3]https://imotions.com/facial-expressions

3. EVALUATION METHOD

The main principle of our approach is to conduct a "black-box testing" [26] as we can not be sure about the provided information about the algorithms and are not able to change parameters inside the algorithm.

Figure 2 illustrates our processing pipeline: After selecting the videos labeled with facial expressions from the set of databases, all are processed by the different algorithms. The output is then normalized to probability values. The values for different expressions are classified using three metrics leading to a decision identifying the primary expression. As a last step, the identified expression is compared to the label from the database.

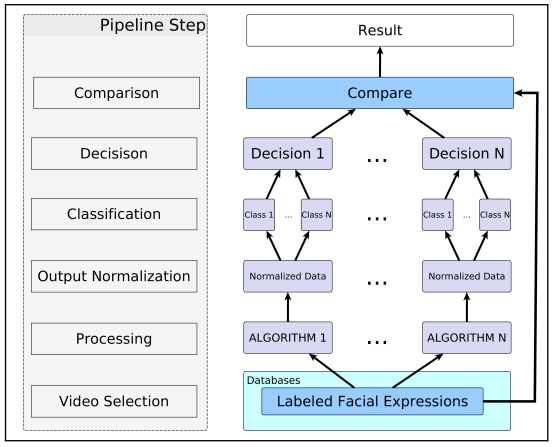

Figure 2: Processing pipeline: Labeled videos from input databases are processed, the identified primary expression is compared to the labeled emotion.

3.1 Normalize Output Probabilities

Some of the generated output needs to be preprocessed before further analysis as the output type differs for the algorithms, shown in Table 1. The data were normalized to probality values for the presence of an emotion between 0.0 and 1.0 to make it comparable. The Emotient and CERT values of evidence (The distance to the hyperplane of a SVM) were therefore transformed using Equation (1). The particular case for 0 was chosen to avoid anomalities when comparing to the other algorithms who return a value of 0 instead of 0.5 (lower part of Equation (1)).

$$p(x) = \begin{cases} 0.0 & \text{if } x = 0 \\ \frac{1}{1+10^{-1*x}} & \text{otherwise.} \end{cases} \quad (1)$$

Table 2: Videos in chosen databases for emotional facial expressions

Database	Included Data			Labeled Expressions (6BE)							Additional	
Name	Subjects	Videos	Labeled	Angry	Joy	Surprise	Sad	Disgust	Fear	All	Contempt	Neutral
AFEW	330	1645	1106	182	208	119	168	112	123	**912**	0	194
CK+	123	593	327	45	69	83	28	59	25	**309**	18	0
BU-4DFE	41	605	605	101	100	101	101	101	101	**605**	0	0

3.2 Facial Expression Recognition Algorithms

Four state-of-the-art FER algorithms were chosen for our approach. All of them are capable of realtime processing and available in commercially accessible systems. These systems were chosen based on their general availability for research emphasizing the practical approach.

3.3 Utilized Databases

Three databases (DBs) were selected due to their insight in different aspects of benchmarking for smart environments. All three databases cover different aspects of facial expressions such as temporal phases, head movement in the scene and frontal or non-frontal scenes (see Table 3).

We chose only videos labeled with a primary expression[4] for our test set (CK+ and AFEW contain additional, non-labeled videos, see Table 2) with an increasing level of difficulty in detection. CK+ [18, 14] has been often used for benchmarking - and training - and thus we included it for comparison purposes.

The BU-4DFE database [31] provides 3D-data of the head in addition to textures. This provides frontal images only containing the head on a black background that are similar to the images extracted via 3D head cropping by a depth camera with head pose tracking as in our EmotionBike setup. AFEW [7] provides close to real life videos taken from movies displaying emotions by professional actors compared to non-professional actors in CK+ and BU-4DFE. The videos from AFEW also show different view-angles, lighting, covering and settings making the analyse of these videos the most challenging task for the algorithms.

Table 3: Characteristics of videos in utilized databases.

Database	Datatype	Head movement	Performer	Origin	Phases
CK+ [18] [14]	2D frontal	no	non-professional	acted expression	onset apex
BU-4DFE [31]	2D frontal / 3D and textures	no	non-professional	acted expression	onset apex offset
Afew [7]	2D frontal / non-frontal	yes	professional actors	movies	mixed

3.4 Classification and Decision

In a database context, the frame with the greatest value of expression is often provided, see [18]. But in a natural environment and real life setup, the task is to identify the primary expression during a period of time. In our previous work, we used a window around a defined event, see [25] for details. For real life scenarios, a different procedure has to be used: We evaluated three possible metrics to identify the primary expression present in videos as defined in Equations (2), (3) and (4), where x is the expression and n is the number of processed frames.

The threshold metrics were included for boosting detection rates for expressions with lower intensity.

$$mean(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

$$meanp5(x) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 0 & \text{if } x_i <= t \\ x_i - t & \text{if } x_i > t \end{cases} \tag{3}$$

$$binaryp5(x) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 0 & \text{if } x_i <= t \\ 1 & \text{if } x_i > t \end{cases} \tag{4}$$

We applied a threshold of t=0.5 for Equation (3) and (4). A lower threshold will increase false positive detections and higher values will lead to missing instances of detections.

To decide for the primary expression, the maximum rule[5] was applied to the results of the metrics for each set of expressions and for each algorithm: The highest value classifies the detected expression and can be compared to the database label.

4. RESULTS

Before we present the detailed benchmark results we like to show and discuss the example FER output, what an application developer may be faced with, even with the fundamental CK+ database. Figure 3 shows an example for a case of false interpretation.

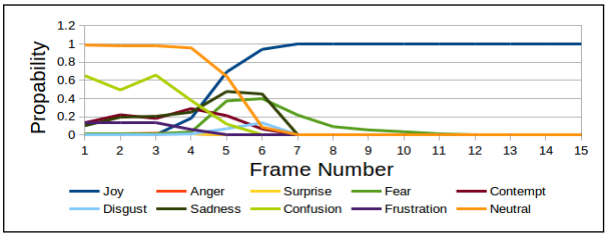

Figure 3: Example output for Emotient and subject S010 with labeled emotion joy from the CK+ database. In addition to the six basic emotions, the values of neutral, contempt, confusion and frustration are presented.

Our recommendation is to match the number of emotions to the decision relevant dimensions for the application context. First, we present the performance results for the six expressions anger, joy, disgust, fear, surprise and disgust.

Second, we provide insight into the impact of choosing the number of output expressions and into the special neutral expression.

[4]The most relevant expression

[5]Also known as the "Winner takes it all"

Table 4: Correct matches between label for the six basic emotions and recognized expressions for all three databases with mean metric and without files labeled 'neutral' or 'contempt'. The best match is highlighted.

Database	Emotion Label	# Valid DB	Number of videos (%) correctly detected by			
			InSight	Affectiva	Emotient	CERT
CK+	Fear	25	21(84%)	2(8%)	20(80%)	**23(92%)**
CK+	Anger	45	26(58%)	16(36%)	42(93%)	**43(96%)**
CK+	Disgust	59	44(75%)	**59(100%)**	56(95%)	57(97%)
CK+	Joy	69	53(77%)	51(74%)	69(100%)	68(99%)
CK+	Sadness	28	17(61%)	15(54%)	**25(89%)**	**25(89%)**
CK+	Surprise	83	**81(98%)**	66(80%)	80(96%)	74(89%)
CK+	**overall**	**309**	**242(78%)**	**209(68%)**	**292(94%)**	**290(93%)**
BU4-DFE	Fear	101	8(%)	4(%)	21(%)	**29(29%)**
BU4-DFE	Anger	101	**64(63%)**	39(%)	61(%)	54(%)
BU4-DFE	Disgust	101	9(%)	72(%)	**73(71%)**	48(%)
BU4-DFE	Joy	100	49(%)	88(%)	**97(88%)**	76(%)
BU4-DFE	Sadness	101	21(%)	25(%)	67(%)	**77(76%)**
BU4-DFE	Surprise	101	72(%)	57(%)	**86(85%)**	33(%)
BU4-DFE	**overall**	**605**	**223**	**285**	**405(70%)**	**317**
AFEW	Fear	101	**6(8%)**	0(%)	4(%)	8(%)
AFEW	Anger	157	**34(22%)**	8(%)	14(%)	26(%)
AFEW	Disgust	100	1(1%)	**46(46%)**	12(12%)	15(15%)
AFEW	Joy	178	9(%)	**97(54%)**	(%) 49	**97(54%)**
AFEW	Sadness	145	29(%)	7(%)	34(%)	**123(85%)**
AFEW	Surprise	102	33(%)	17(%)	**51(50%)**	10(%)
AFEW	**overall**	**783**	**112(%)**	**173(%)**	**164(%)**	**279(36%)**

4.1 Six Basic Emotions

4.1.1 Performance analysis

The facial expressions for the six basic emotions are the minimum of expressions shared between all algorithms and databases. As we have six classes, the minimum required detection rate is the one of guessing: ∼17%. Table 4 shows the detection rate for all databases and algorithms using mean as the best performing metric.

The detection rates have a great variety between the databases. From the output of the processing of CK+ we would assume, that all four algorithms were trained with at least parts of the database (for CERT this is known) [17].

We also noticed that some algorithms tend to prefer expressions. As an example, CERT shows a strong tendency to detect sadness ranging from 35% (in case of joy input) to 76% when analysing the AFEW videos.

4.1.2 Uncertaincy with anger and fear

Although the detection rate for fear is fairly good for the CK+ data for all algorithms except Affectiva, it constantly drops to almost zero for most of the AFEW videos. Instead of detecting fear, surprise and sadness are often detected for the AFEW videos.

A similar tendency can be observed for anger, but the results are not as distinct compared to fear. While fear drops to half (0.08) of the rate of random guessing, the rate for anger is still above (0.22). The low detection rate of fear has been reported in literature before, Valstar et al. for example reported similar results while preparing the baseline for their FERA Challenge on facial expressions [29].

4.1.3 AFEW: Videos not recognized

1106 labeled videos from AFEW where processed, for 156 no result was generated when using the mean metric (226 with meanp5, 229 with binaryp5). The average rate of no detectable expressions varies between 10% (disgust) and 17% (fear). To achieve comparibility over all three algorithms, the focus was on the six basic expressions and 'neutral' was ignored.

There are basically two possible cases for videos that could not be analysed: No face could be detected or the outcome of the FER algorithms produced very low values, so that the mean was rounded to zero. We investigated this further: InSight and Emotient failed in all 129 videos to detect a face even once. Affectiva missed the face in 104 of the files, while only 16 videos resulted in a face detection in more than one frame. In comparison with Emotient, CERT was able to detect faces in all but 17 files.

Examining the non-detectable files we found a number of conditions which made detection difficult: They often contained partly covered faces, sometimes covered by long hair or strong shadows. Other conditions contributing to low face detection rate were also an extreme head pose, people wearing glasses or reflections in the background. An unusual case we found in a number of videos are bright - but not directly blinding - lights in the background.

4.2 Different Number of Expressions

Defining the number of facial expressions used from the output of the algorithms has a significant impact on the quality of the detection rate.

4.2.1 More than Six Expressions: Confusing the Matrix

While a confusion matrix only considering the six basic emotions works well for the CK+ database, using additional facial expressions provided by the algorithm can lead to a shift in the detected primary expression. Figure 4 illustrates this case: Instead of detecting anger as primary expression (detected in 93% of the videos) when using the six basic emotions, the confusion expression is often (55% of the videos) recognized resulting in a detection rate for anger less than 10%. This has to be accounted for real life applications, especially in a context, where this additional facial expressions occur naturally and are part of the target emotion set.

4.2.2 Neutrality

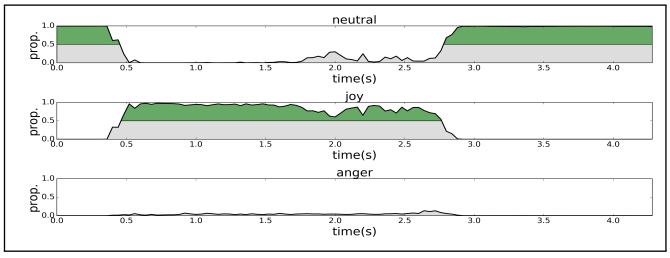

Figure 5: Example plot for BU-4DFE and InSight: Neutral is 1.0 minus all found emotions. Prob. of surprise, disgust, sadness and fear is near zero.

Lewinski defines that "a neutral face should indicate lack of emotion" [16]. There are different approaches to handle this 'neutrality' in the four algorithms. One approach is to calculate it with 1.0 minus the sum of all detected emotions (InSight uses a maximum of 1.0 overall emotions, see Figure 5).

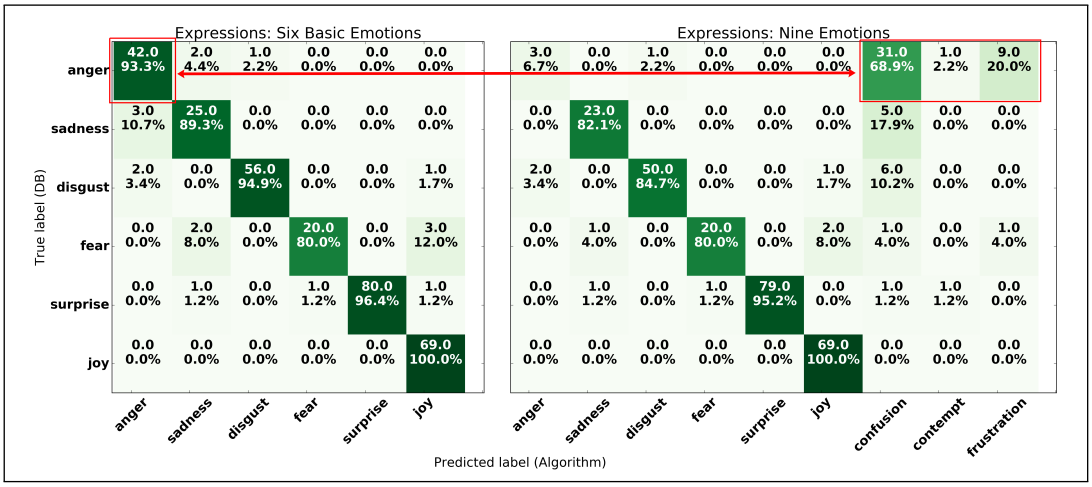

Figure 4: Example confusion matrix for Emotient using CK+ database showing only the six basic emotions and in comparison all detectable facial expressions. For Anger this leads to a shift to confusion and frustration.

Another approach is to apply separate SVMs for all different expressions (Emotient, CERT). In this case, neutral is a seperate SVM similar to all other expressions. Affectiva ignores the neutral case.

4.2.3 *Clustering of Smile and Joy*

Figure 6 visualizes the output of all algorithms for the subject F029 and the expression joy from the CK+ database. Affectiva seems to apply a constricted definition of 'joy' compared to the 'smile' facial expression resulting in a reduced detection rate. Clustering of the two emotions joy and smile could increase the detection rate.

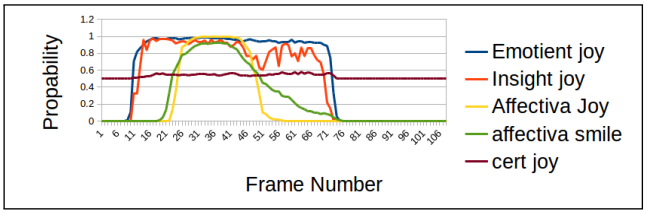

Figure 6: Example output of all algorithms for 'joy' labeled video from BU-4DFE database.

4.2.4 *Metrics*

We evaluated the three metrics mean, meanp5 and binaryp5, see section 3.4, to test their practicability in detecting the primary expression.

The mean metrics provided the best results overall datasets and algorithms, with the exception, that binaryp5 provided better results in a few cases with the CERT system. As an example for this effect, while classifying anger from the AFEW database, mean provided matches for 26 videos in contrast to the 59 with binaryp5[6], but this is only true for anger. All other expressions could not be correctly classified as primary expression. The results of meanp5 are always close to the ones provided by mean but we found no conditions where meanp5 were superior to binaryp5.

[6]meanp5 found 29 matches.

The mean metric seems the principal approach for the best results in classification of the output data of the algorithms.

5. BOOST ROBUSTNESS BY APPLICATION SPECIFIC CLUSTERING

In our database performance analysis of section 4 we have shown that false categorizations is a common problem in FER systems. Our approach to boost robustness tailors the emotion categories to the specific application context using clustering methods. For example a classic approach to reduce dimensions with a valence (positive or negative) arousal (intensity) emotion model [33] may be suitable.

In addition, we created other emotion clusters that excluded specific emotion states irrelevant for the application context.

To illustrate this, we present three use cases with clustering solution:

Usability study: For a usability test of a new smart home application, developers are interested if and when people have positive, neutral or negative emotions while testing the software. The class 'positive' consists of joy, 'negative' is defined as fear, anger, disgust or sadness. None or surprise is considered to be 'neutral'.

Cockpit scenario: A system for driver assistance should decide, at which point the driver is under negative stress resulting in more automatic support ('alarm') and when to leave control to the driver ('normal'). For this, a class 'alarm' consists of anger, fear, disgust and surprise. Joy, sadness, none are considered to be in the 'normal' class.

Learning system: For a smart learning system, a (virtual) teacher should be made aware when the students are still following (in-flow) or not. The class 'inflow' consists of surprise, none, joy and 'panic' of fear, anger, disgust. Sadness is ignored as it is considered to be an emotion not elicited by the scenario.

We applied the results of the AFEW database for all application contexts. Although the videos do not show the specific task, this database is still the most challenging and

illustrates the significant increase in application specific detection rates as shown in Table 5: Compared to the detection rates from the distribution into six emotion classes, the rates are at least doubled (19% to 42% and 48% for Affectiva) with our application based approach, making the detection and possible reactions of the smart application much more stable.

Adding 'none' to a class of expressions can be considered as setting a default (fallback) state for your application behaviour.

Table 5: Application specific clustering of six basic emotions (6BE). The detection rate is calculated related to all videos in AFEW not labeled 'neutral'.

App.	Cluster	InSight	Affectiva	Emotient	CERT
usability	positive	9	**97**	49	**97**
usability	neutral	79	60	**98**	14
usability	negative	201	278	180	**515**
usability	overall	289	435	327	**626**
usability	**detect. rate**	**32%**	**48%**	**36%**	**69%**
cockpit	alarm	262	**316**	240	199
cockpit	normal	204	221	264	**312**
cockpit	overall	466	437	504	**511**
cockpit	none	148	103	150	9
cockpit	**detect. rate**	**51%**	**48%**	**55%**	**56%**
learning	inflow	262	**302**	265	126
learning	panic	**92**	90	70	58
learning	overall	354	392	**395**	184
learning	**detect. rate**	**39%**	**42%**	**43%**	**34%**
6BE	overall	112	175	164	**279**
6BE	**detect. rate**	**12%**	**19%**	**18%**	**31%**

6. CONCLUSION

Robust automatic facial expression recognition is crucial in realizing many innovative smart environmnent applications. In this work we have proposed and explained our method of testing and benchmarking state-of-the-art FER systems using emotion labeled databases of facial expressions.

Based on extensive empirical experiments, we further presented our method of application specific clustering of expressions as a simple but practical approach to overcame current limitations of FER algorithms. Using this method with the full set of expressions provided by a FER system could change the problem with shifting of the primary expression due (described in section 4.2.1) from a limitation to a benefit.

FER systems have great potential in enhancing the interaction between human and computer, yet current limitations still leave room for improvement. We hope that this work is a step forward in improving the practicability of such systems.

7. FUTURE WORK

Expanding our datasets with a database displaying sponatious emotions like BP4D from Birmingham University [34] could increase the fundament of decision for discovering an improved combination of FER algorithms for an intelligent fusion approach.

Evaluating more algorithms would also be a logical step in this direction.

Processing our own dataset of context annotated videos containing provoked emotions from the EmotionBike project is a future task to benchmark our clustering method.

Disclaimer

This work is not company sponsored. Our aim is to provide a neutral analysis and the obervations are limited to the applied datasets. We would like our results added to further research on FER to improve the overall system performance, especially for applications in smart environments.

All product names are property of their respective owners.

Acknowledgment

We express our gratitude to the EmotionBike project team for their technical support and to the University of Applied Sciences Hamburg for funding this work.

8. REFERENCES

[1] M. Abouelenien, M. Burzo, and R. Mihalcea. Human acute stress detection via integration of physiological signals and thermal imaging. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '16, pages 32:1–32:8, New York, NY, USA, 2016. ACM.

[2] M. Bartlett, G. Littlewort, T. Wu, and J. Movellan. Computer expression recognition toolbox. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–2, Sept 2008.

[3] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 379–388, New York, NY, USA, 2015. ACM.

[4] H. Brauer, C. Grecos, and K. von Luck. *Robust False Positive Detection for Real-Time Multi-target Tracking*, pages 450–459. Springer International Publishing, Cham, 2014.

[5] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, Jan 2010.

[6] D. J. Cook and S. K. Das. How smart are our environments? an updated look at the state of the art. *Pervasive and Mobile Computing*, 3(2):53 – 73, 2007. Design and Use of Smart Environments.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, July 2012.

[8] S. D'Mello, A. Graesser, and R. W. Picard. Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, 22(undefined):53–61, 2007.

[9] S. K. D'mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.*, 47(3):43:1–43:36, Feb. 2015.

[10] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[11] S. Eleftheriadis, O. Rudovic, and M. Pantic. Joint facial action unit detection and feature fusion: A multi-conditional learning approach. *IEEE Transactions on Image Processing*, 25(12):5727–5742, Dec 2016.

[12] J. Ellenberg, B. Karstaedt, S. Voskuhl, K. von Luck, and B. Wendholt. An environment for context-aware applications in smart homes. In *International Conference on Indoor Positioning and Indoor Navigation (IPIN), Guimaraes, Portugal*, 2011.

[13] W. Friesen and P. Ekman. *EMFACS-7: Emotional Facial Action Coding System*. Unpublished manual, University of California, California.

[14] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.

[15] E. Kanjo, L. Al-Husain, and A. Chamberlain. Emotions in context: examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing*, 19(7):1197–1212, 2015.

[16] P. Lewinski. Automated facial coding software outperforms people in recognizing neutral faces as neutral from standardized datasets. *Frontiers in Psychology*, 6(1386), 2015.

[17] 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*, pages 298–305. IEEE, 2011.

[18] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, June 2010.

[19] I. Maglogiannis. Human centered computing for the development of assistive environments: The sthenos project. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '14, pages 29:1–29:7, New York, NY, USA, 2014. ACM.

[20] D. McDuff, A. N. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. E. Kaliouby. AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016, Extended Abstracts*, pages 3723–3726, 2016.

[21] P. Michel and R. El Kaliouby. Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, ICMI '03, pages 258–264, New York, NY, USA, 2003. ACM.

[22] D. Mihai, G. Florin, and M. Gheorghe. Using dual camera smartphones as advanced driver assistance systems: Navieyes system architecture. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '15, pages 23:1–23:8, New York, NY, USA, 2015. ACM.

[23] G. Mone. Sensing emotions. *Commun. ACM*, 58(9):15–16, Aug. 2015.

[24] L. Müller, A. Bernin, C. Grecos, Q. Wang, K. von Luck, and F. Vogt. Physiological data analysis for an emotional provoking exergame. In *Proceedings of the IEEE Symposium for Computational Intelligence*. IEEE, Athens, Greece, 2016.

[25] L. Müller, S. Zagaria, A. Bernin, A. Amira, N. Ramzan, C. Grecos, and F. Vogt. Emotionbike: a study of provoking emotions in cycling exergames. In *Entertainment Computing-ICEC 2015*, pages 155–168. Springer, 2015.

[26] R. Patton. *Software Testing (2Nd Edition)*. Sams, Indianapolis, IN, USA, 2005.

[27] R. W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

[28] C. Smailis, N. Sarafianos, T. Giannakopoulos, and S. Perantonis. Fusing active orientation models and mid-term audio features for automatic depression estimation. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '16, pages 39:1–39:4, New York, NY, USA, 2016. ACM.

[29] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926, March 2011.

[30] P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[31] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 211–216, April 2006.

[32] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 138:1 – 24, 2015.

[33] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, Jan 2009.

[34] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692 – 706, 2014. Best of Automatic Face and Gesture Recognition 2013.