Establishing Ground Truth on Pyschophysiological Models for Training Machine Learning Algorithms: Options for Ground Truth Proxies

Keith Brawner(✉) and Michael W. Boyce

Army Research Laboratory, Orlando, FL, USA
{keith.w.brawner.civ,michael.w.boycell.civ}@mail.mil

1 Introduction

One of the core aspects of human-human interaction is the ability to recognize and respond to the emotional and cognitive states of the other person, leaving human-computer interaction systems, at their core, to perform many of the same tasks. This can take the form of robotic interaction systems that respond to 'anger' [33], instructional systems that take different actions according to 'confusion' [27], and intelligent aiding systems which dynamically adjust levels of autonomy (i.e. task allocation to the human or the system) depending on continuously changing levels of 'workload' [9]. A well-designed system responds to information about the user, tailoring the experience for the purposes of enjoyment, effectiveness, or both. The emphasis of this paper focuses on understanding that emotional state to maximize human performance.

While there are many reasons why one might want to recognize emotional states within a population of individuals for the purpose of designing systems, the model creation process is fundamentally the same across much of the research. Briefly, this process can be described as the below:

1. Data about 'state' is collected, features of this data.
2. The features are distilled into markers for easier machine learning classification.
3. These markers are fed into one or more model creation algorithms.
4. Affective classification models are created.
5. Affective models are used.

While models can be built from numerous and disparate data sources, the underlying affective data is frequently suspect. Let us consider a labeled datapoint in a set which indicates 'frustration'. Is the user frustrated because they said so right now? Because they said that they were frustrated with the overall experience? Because models calibrated on another frustration event said they were? How do you know that the user is truly frustrated? Is either cognitive underload/overload [12] causing the frustration? During analysis, how can you be assured of the quality of the label in the spreadsheet?

The common aphorism is that "all models are wrong, but some are useful." The quality of labels is frequently dubious, due to the way that the labels are collected (subject to experimenter or experiment design bias), but that does not mean that it

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2017, Part I, LNAI 10284, pp. 468–477, 2017.
DOI: 10.1007/978-3-319-58628-1_35

cannot be applied to useful applications. Further, the quality can be improved through the combination of more than one labeling technique. Researchers have begun to develop adaptive multimodal recognition systems which focus on good quality samples to form training data and thus assisting to reject bad samples. The multimodal adaptive system has the added advantage of achieving better performance and a lower failure rate [11]. There are many options for collecting labeled data, each with its own advantages, disadvantages, limitations for collections, and workarounds for attempting to assure that the label is of high quality. Two significant research design problems exist, both in matching the collection technique to its intended use and the combination of labels from multiple schemes.

While a singular strong state (e.g. rage, surprise, exhaustion) is relatively easy to label, the majority of the states useful for Human Computer Interaction (HCI) research are nuanced (e.g. annoyance, confusion, underload). Distinguishing nuanced states, such as 'underload' from 'boredom', into a labeled category can be difficult. The careful selection of the method chosen for labeled data collection can help to mitigate the problem with wrong model for a specific application or use case.

Research efforts are beginning to apply hybrid techniques of multiple methods for dealing with data to improve quality [24]. Specifically, it is possible to take a qualitative analysis technique such as Grounded Theory, and apply it iteratively when performing sampling to support a machine learning algorithm. Both approaches are derived from the data, with the machine learning looking for features and the Grounded Theory looking for theories or themes that describe those features. In terms of the ability to detect affect from this data, researchers have generally tried to pull together a series of measurements via different types of physiological sensors, such as research that shows the ability to properly classify Valence and Arousal with over a 90% success rate [25], using a combination of Skin Conductance, Heart Rate, and Electromyography. Using a combination of techniques can assist in beginning to target more complex responses. As a part of their research, Noguiera et al. [25] built a regression model which runs through several iterations to assist with data scaling. Then they perform a second pass with several machine learning algorithms to merge the outputs of the regression model into an aggregated score. They use objective player modeling techniques (OPEM), which have been shown to be very consistent between administrations. Even with all these methods to collect data, a key challenge in establishing ground truth is understanding proportionally how to adjust the importance (i.e. weights) of the various data collection measures.

This paper reviews different options for obtaining "ground truth" labeled data from users. The methods examined include:

1. Using pre-existing and validated models created from a standard dataset
2. Using pre-existing and validated models created from multiple contexts of experience
3. Using manually labeled datasets
4. Through self-reported labeled ask soon after or during after the experience
5. Through self-report labeled feedback asked after a number of experiences
6. Creating sensor-based models from theory directly.

Each of these options has advantages, disadvantages, limits or restrictions, and mitigations or workarounds. This paper lightly reviews the literature for groups which have used various techniques in an effort to inform future research on the selection of the experimental labeling design best suited for the end application or use case. The expertise of the authors lies in training system applications, so the labeling techniques discussed within this paper are an extension of that lense through pragmatic application. It is an expansion on some of last years' points from the "truthiness" paper by Mark Costa and Sarah Bratt for the HCII community [8].

2 Calibrate from a Standard Dataset

One of the most attractive methods for creating emotional recognition work is to calibrate from an existing dataset. Examples of baseline datasets are the Pose Illumination and Expression (PIE) database [30], or the SEMAINE affective dataset [23]. The advantage of such a dataset is that either paid actors or in-the-wild observations of ground truth that can be used to baseline. The significant disadvantage of the approach is that the models made from such a dataset, while useful for facial detection, haven't made much, if any, progress into realtime applications. The models have been useful in the methods to develop generalized facial detection models, but have not demonstrated use in in-the-wild affect detection problems. Part of the reason for this is that the mapping between the observed face, the Facial Action Coding System (FACS) of Action Unit (AU) identification, and the actual emotion is tenuous [28].

Other work includes Conati's probabilistic assessment models and Pantic and Rothkrantz's audio/video combination methods [26]. In Conati's model, user emotions are defined through several different dimensions: student goals, variables describing student personality traits, actions to be taken by the agent, and variables describing the user's emotional state [7]. They expressed the importance of using the ability of recognizing affective state to make interactions more affective. They argue that communicative cues such as facial expressions and body movements are affected by individual's arousal. Likewise, when humans are interacting with each other humans can interpret these cues while machines have a much more difficult time. They explained that to analyze human affective feedback there needs to be an architecture which supports information coming from the visual system, information coming from the processing of audio, and information coming from touch or tactile sensors. This data in turn would undergo both feature level fusion across information types and data information interpretation to help make decisions on the appropriate feedback. Pantic and Rothkrantz classify it as data level fusion, feature level fusion, and decision level fusion [26].

While fewer emotional datasets exist for physiological signals, the reader should note the lessons from the above, such as establishing similar items from a commonly available physiological model, can be expected to encounter similar difficulties. When discussing classification for the purposes of building affective models from physiology, the authors are only generally aware of two common-access databases for the purpose: the Deap database for emotion analysis using physiological signals [21], and a dataset made available by the authors [5].

3 Calibrate from an Existing Validated Model

The general scope of this paper is to discuss evidence that all measures are a proxy of the true ground truth. While the real truth certainly exists in the brain as measured by electroencephalogram (EEG) signals, there is always the concern of the accuracy of measurement. The true brain signal is spread across the skull, subject to noise in measurement, to significant individual differences in brain topology, and varies by day and sensor placement. However, EEG signals are reliable for certain tasks and some systems have been extensively evaluated. As an example, the Advanced Brain Monitoring system can generate real-time indices of alertness, cognition, and memory [2], or measures of drowsiness/alertness [16]. There have been many (20+) studies where the patented detection algorithms have been validated over relatively stable timeframes (minutes). The studies which use functional magnetic resonance imaging (fMRI) have similarly levels of validation, with early research indicating success at tracking moment-to-moment changes in affect [20]. Each of these systems can be relied upon to give fairly accurate information in regards to labeling.

The purpose of a system may to be analyze affect during task performance, with findings useful to the system creators. An example of such a finding would be an interface which causes high levels of workload and dissatisfaction among its users. The finding can be used to redesign the system in such a way to reduce cognitive load. More frequently, however, the purpose of the system is to respond to the users' needs as they need them. An example would be an interface that re-configures based on the high workload, or a teaching system that uses knowledge of the user frustration to make changes in courseware/courseflow. The use one EEG or fMRI system for each user is fundamentally impractical.

An alternative to the use of such systems is to use the high-accuracy systems as the 'ground truth' for a series of, presumably, lower accuracy systems. In the same manner that video systems can use lipreading to distinguish words without audio, systems can be designed to use low-cost wearables and stand-off sensors in order to capture the emotion [22]. The authors have publicly shared such a dataset in the past [5]. The advantage of such an approach is that the ground truth can be considered reasonably reliable, but the disadvantage is the compounding of errors. A classifier which predicts with 80% accuracy on a signal with 80% accuracy in a system which may be barely usable with 64% accuracy. Experimenters should consider this potential compounding of inaccuracies when designing systems, but low levels of accuracy may be acceptable for systems which make slow and reliable decisions.

4 Manual Expert Label

One method of addressing the flaws of inability to attain ground truth information is to begin relying on post-hoc added labels to existing recorded data. The process of doing this relies on capturing the nuanced emotions experienced during the desired event using the classifiers in an operational setting. This is the basis for many qualitative research methods that categorize participant actions in a hope to provide more general,

overarching themes. Such qualitative approaches include thematic analysis [4] and Interpretative Phenomenological Analysis [3, 31]. On the other end of the spectrum is labeling based on physiological data such as eye fixations and saccades, as was done with an intelligent tutoring system called Metatutor [15]. As an example, consider the tasks of the classification of a fatigued driver. An experimental setup would allow for the driver to perform their normal function while being observed via a combination of bodily (e.g. EEG) and standoff sensors (e.g. webcam). The video data can then be carefully combed by expert labellers at the second-by-second resolution. These "ground truth" labels can then be used to train automatic classifiers for the bodily sensors (EEG), the standoff sensors (webcam), or use a combination of data fusion to attempt to train both.

The advantage of this approach is that, through the use of expert labellers and time-delayed recording, the ground truth information can be captured at relatively fine resolution. As an example, the first moments of affective information can be traced to their earliest FACS movements. A further advantage is that the classifiers trained are applicable in the desired application.

The first disadvantage of the approach is that the methods of classifications are not particularly guaranteed to transition beyond their initial domain. The second disadvantage is that the classifiers in this instance have the tendency to be 'jittery', rapidly classifying emotions at their earliest onset. Jittery classification can be overcome at the labeling instance, by labeling an emotion only when it is fully manifested in the desired application, or at the runtime instance, where simple rules can dampen system actions (e.g. "only act when the emotion has been present for greater than 80% of a 3 min window").

5 Self Report

All self-report data, arguably, has the same advantage that it is the ground truth, as the participant has reported it. In some manner, it is very difficult to contradict a participant which responded that they were 'bored' and 'unchallenged' (low workload) by a series of educational content presentations. Hoskin details the typical problems with self-reported data [14]. In brief, these include:

- Individual differences in introspective ability
- Individual variations in interpretation of a question
- Individual variations in rating scales, especially with large variations, such as [0–100]
- Response bias, especially in yes/no questions

Simply using a survey measure such as the NASA-Task Load Index (TLX) [13] to label a 30-second window of time can be subject to all of the above flaws. These flaws may even out over a large amount of samples, on the whole, but using them to label 1000s of datapoints from raw sensors is a gross measure, at best. This limitation is overcome if the experimenter desires a gross measure of the particular affect (e.g. 'confused' at 30 s resolution is sufficient in production).

The accuracy concerns can be mitigated through the use of the more validated instruments, such as the TLX. However, the experimenter should be aware how the frequency of polling can affect the data overall. Additionally, the experimenter should be aware that asking about an experience can change the perception of the experience. An example study where this effect is observed is in an educational study, where significant difficulties were encountered during the learning environment, but reported as interest and enjoyment after the fact [19]. It is worth noting that early results to try to build a system at the same time that it is being used have had sufficient predictive accuracies to be useful in both simulation [6] and practice [10].

5.1 Post-hoc Self-report

The general advantages and disadvantages of self-report are discussed above, being that subjects' estimate of their own emotions is arguably better than expert annotation. The notable disadvantage of post-hoc self-report is typical of most video game and learning experiences: the experience itself is somewhat challenging. When asked after an experience about the emotions experienced during the situation, the experience tends to be cast through the lens of the final moment (e.g. winning, losing, learning, etc.). The most useful workaround for this problem is to use a group-based model to create distinctive groups, each of which can be targeted for action [32].

5.2 In-Situ Self-report

The general advantages and disadvantages of self-report are discussed above. The alternative method of gathering self-report data is to ask the participant in situ to report their emotions or experience. The "think aloud protocol" allows for the experimenter to obtain a continuous feed of user affective states, resulting in a higher granularity of samples for model creation. The largest disadvantage of this approach is that the experience of "think aloud" can have a modest effect on workload and task performance [29]. This effect can be mitigated by having the "think aloud" be related to the task.

6 Physiological Sensors

The advantage of using physiological sensors, as opposed to any of the other above methods, is that the "ground truth" is objective. The raise in Galvanic Skin Response (GSR) or increase of blood flow to an area of the brain, or the frequency of brain operation, are resistant to subject recall, self-report, rater bias, or the error rate of a previously established model. These advantages are significant, but do not come without costs. The costs are that the measurement is usually not suited for the intended environment, individual responses vary significantly and change daily, and that the measurements are usually gross proxies for the things that they are measuring.

The measurement via sensors is usually not appropriate for the intended environment. As an example, much of the emotion-based research in the educational domain is eventually intended to influence the decisions of systems or teachers about the content presented to the student within the classroom. With the average classroom size in the United States around 25 students per teacher, and the cost of an fNIRS system in the range of $50,000, the educational benefits of emotional detection are simply not justified in the cost. Furthermore, many sensor-based systems require extensive set-up, which consumes time that could have been spent on the performance task.

Another downside lies simply in the quality of the data collected versus any interference that is potentially associated with it. Depending on the sensor used, what might be considered a response from a classification algorithm can be noise associated with the electronics, noise associated with the participant, or noise associated with the environment. As a very simple example, consider electrodermal activity being collected to measure arousal. Using many of the electrodermal sensors currently on the market, factors such as the ambient room temperature, skin temperature due to clothing, contact with the skin, and charge of the battery on the sensor can all lead to artifact. This does not even include gross motor movements, which can drastically impact results and the connectivity of the sensors.

Next, sensors frequently measure only a proxy of what they intend to measure. Taking the example of electrodermal activity sensor which measures the changes in skin conductance. These changes proxy are measurements for the autonomic nervous system ("fight/flight") activity, which is linked to emotional and cognitive states. A raise in GSR response can indicate stress, fear, anxiety, excitement, interest, or the anticipation of any of these things. The sensors can be calibrated over time to compensate for this weakness, but the measurement of a sensor is rarely conclusive evidence of an emotional state. Researchers, such as Picard's group, maintain a successful line of research in artifact detection and have developed screening tools to help identify responses, clear noise, and process signals against a predefined set of transformations, with tools released for others [19].

7 Conclusions

It is worth noting that early results to try to build a system at the same time that it is being used have had sufficient predictive accuracies to be useful in both simulation [6] and practice [10], which neatly avoids much of the problems of labeling.

With the development of technologies such as crowdsourcing, researchers have begun to address labeling of content in new innovative ways. Katsimerou et al. used a database of over 180 long videos which contained three different visual cues involving face and body, as well as a physical depth-based data stream from the Microsoft Kinect [18]. They used crowdsourcing to be able to make large numbers of annotations related to mood and emotion by non-expert coders. They also compared this against laboratory trained annotations to validate the non-expert inputs. As more and more information becomes available via cloud services, it is likely that labeling accomplished by larger groups of people may become the norm.

Automated detection algorithms are becoming a popular source for labeling data as well. Other researchers have used a multimodal approach (appearances collected from a camera vs context specific behaviors captured by the application) to train different classifiers to interpret affective state [1]. Kapoor and Picard used a multimodal approach in where they combined posture based data with camera based data to achieve an 86% accuracy rating of affective state, the approach was a unified Bayesian approach using Gaussian process classifiers that used expectation propagation (EP) [17].

The intended takeaway from this paper is that no *one* technique is probably sufficient for the accurate representation for the ground truth classification of affective state. However, a number of hybrid techniques can be investigated to mitigate the difficulties in any individual approach. Many experiments under various contexts using semi-reliable self-report information can be combined into reasonably reliable labels. Hybrid approaches may use active machine learning to intelligently select datapoints for labeling, with crowdsourced labeling experts providing annotations, making use of both machine learning techniques and within-task self report information [6]. Another hybrid approach may have an individual baselining period which bootstraps the machine learning classifier in batched training, updating it based on after-task self report information [10]. The authors believe that these multi-point labeling approaches tend to produce higher-quality labels overall, which result in models which are less brittle.

References

1. Alyuz, N., Okur, E., Oktay, E., Genc, U., Aslan, S., Mete, S.E., Stanhill, D., Arnrich, B., Esme, A.A.: Towards an emotional engagement model: can affective states of a learner be automatically detected in a 1:1 learning scenario. In: Proceedings of the 6th Workshop on Personalization Approaches in Learning Environments (PALE 2016), 24th Conference on User Modeling, Adaptation, and Personalization (UMAP 2016), CEUR Workshop Proceedings, this volume (2016)
2. Berka, C., Levendowski, D., Cvetinović, M., Petrović, M., Davis, G., Lumicao, M.P., Živković, V., Olmstead, R.: Real-time analysis of EEG indices of alertness, cognition and memory acquired with a wireless EEG headset. Int. J. Hum.-Comput. Interact. **17**(2), 151–170 (2004)
3. Boyce, M.W., Cruz, D., Sottilare, R.: Interpretative phenomenological analysis for military tactics instruction. In: Kantola, J.I., Barath, T., Nazir, S., Andre, T. (eds.) Advances in Human Factors, Business Management, Training and Education. AISC, vol. 498, pp. 623–634. Springer, Cham (2017). doi:10.1007/978-3-319-42070-7_58
4. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)
5. Brawner, K.: Data sharing: low-cost sensors for affect and cognition. In: Proceedings of the Educational Data Mining, London, UK (2014)
6. Brawner, K.W.: Modeling learner mood in realtime through biosensors for intelligent tutoring improvements. Department of Electrical Engineering and Computer Science University of Central Florida, p. 500 (2013)

7. Conati, C.: Probabilistic assessment of user's emotions in educational games. Appl. Artif. Intell. **16**(7–8), 555–575 (2002)
8. Costa, M., Bratt, S.: Truthiness: challenges associated with employing machine learning on neurophysiological sensor data. In: Schmorrow, D.D.D., Fidopiastis, C.M.M. (eds.) AC 2016. LNCS, vol. 9743, pp. 159–164. Springer, Cham (2016). doi:10.1007/978-3-319-39955-3_15
9. de Winter, J.C., Happee, R., Martens, M.H., Stanton, N.A.: Effects of adaptive cruise control and highly automated driving on workload and situation awareness: a review of the empirical evidence. Transp. Res. Part F Traffic Psychol. Behav. **27**, 196–217 (2014)
10. Fairclough, S.H., Karran, A.J., Gilleade, K.: Classification accuracy from the perspective of the user: real-time interaction with physiological computing. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3029–3038. ACM (2015)
11. Gupta, R., Khomami Abadi, M., Cárdenes Cabré, J.A., Morreale, F., Falk, T.H., Sebs, N.: A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 317–320. ACM (2016)
12. Hancock, P.A., Chignell, M.H.: Mental workload dynamics in adaptive interface design. IEEE Trans. Syst. Man Cybern. **18**(4), 647–658 (1988)
13. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. **52**, 139–183 (1988)
14. Hoskin, R.: The dangers of self-report. In: Science Brainwaves (2012). http://www.sciencebrainwaves.com/the-dangers-of-self-report/
15. Jaques, N., Conati, C., Harley, J.M., Azevedo, R.: Predicting affect from gaze data during interaction with an intelligent tutoring system. In: Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K. (eds.) ITS 2014. LNCS, vol. 8474, pp. 29–38. Springer, Cham (2014). doi:10.1007/978-3-319-07221-0_4
16. Johnson, R.R., Popovic, D.P., Olmstead, R.E., Stikic, M., Levendowski, D.J., Berka, C.: Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. Biol. Psychol. **87**(2), 241–250 (2011)
17. Kapoor, A., Picard, R.W.: Multimodal affect recognition in learning environments. In: ACM Multimedia 2005, pp. 677–682 (2005)
18. Katsimerou, C., Albeda, J., Huldtgren, A., Heynderickx, I., Redi, J.A.: Crowdsourcing empathetic intelligence: the case of the annotation of EMMA database for emotion and mood recognition. ACM Trans. Intell. Syst. Technol. (TIST) **7**(4), 51 (2016)
19. Kennedy, G., Lodge, J.M.: All roads lead to Rome: tracking students' affect as they overcome misconceptions (2016)
20. Knutson, B., Katovich, K., Suri, G.: Inferring affect from fMRI data. Trends Cogn. Sci. **18**(8), 422–428 (2014)
21. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2012)
22. Kokini, C., Carroll, M., Ramirez-Padron, R., Hale, K., Sottilare, R., Goldberg, B.: Quantification of trainee affective and cognitive state in real-time. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC) NTSA, pp. 2155–2166 (2012)
23. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. **3**(1), 5–17 (2012)

24. Muller, M., Guha, S., Baumer, E.P., Mimno, D., Shami, N.S.: Machine learning and grounded theory method: convergence, divergence, and combination. In: Proceedings of the 19th International Conference on Supporting Group Work, pp. 3–8. ACM (2016)
25. Nogueira, P.A., Rodrigues, R., Oliveira, E., Nacke, L.E.: A hybrid approach at emotional state detection: merging theoretical models of emotion with data-driven statistical classifiers. In: Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-vol. 02, pp. 253–260. IEEE Computer Society (2013)
26. Pantic, M., Rothkrantz, L.J.: Toward an affect-sensitive multimodal human-computer interaction. Proc. IEEE **91**(9), 1370–1390 (2003)
27. Pedro, M.O., Baker, R., Bowers, A., Heffernan, N.: Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In: Educational Data Mining 2013 (2013)
28. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**(10), 1175–1191 (2001)
29. Pike, M.F., Maior, H.A., Porcheron, M., Sharples, S.C., Wilson, M.L.: Measuring the effect of think aloud protocols on workload using fNIRS. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 3807–3816. ACM (2014)
30. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: Proceedings of the Fifth IEEE International Conference on IEEE Automatic Face and Gesture Recognition, pp. 46–51 (2002)
31. Smith, J.A.: Reflecting on the development of interpretative phenomenological analysis and its contribution to qualitative research in psychology. Qual. Res. Psychol. **1**(1), 39–54 (2004)
32. Valle, A., Núñez, J.C., Cabanach, R.G., González-Pienda, J.A., Rodríguez, S., Rosário, P., Cerezo, R., Muñoz-Cadavid, M.A.: Self-regulated profiles and academic achievement. Psicothema **20**(4), 724–731 (2008)
33. Zhang, L., Jiang, M., Farid, D., Hossain, M.A.: Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. Expert Syst. Appl. **40**(13), 5160–5168 (2013)