Artificial Neural Networks Can Distinguish Genuine and Acted Anger by Synthesizing Pupillary Dilation Signals from Different Participants

Zhenyue Qin, Tom Gedeon(✉), Lu Chen, Xuanying Zhu, and Md. Zakir Hossain

Research School of Computer Science,
Australian National University, Canberra, Australia
{zhenyue.qin,lu.chen,xuanying.zhu,zakir.hossain}@anu.edu.au,
tom@cs.anu.edu.au

Abstract. Previous research has revealed that people are generally poor at distinguishing genuine and acted anger facial expressions, with a mere 65% accuracy of verbal answers. We aim to investigate whether a group of feedforward neural networks can perform better using raw pupillary dilation signals from individuals. Our results show that a single neural network cannot accurately discern the veracity of an emotion based on raw physiological signals, with an accuracy of 50.5%. Nonetheless, distinct neural networks using pupillary dilation signals from different individuals display a variety of genuineness for discerning the anger emotion, from 27.8% to 83.3%. By leveraging these differences, our novel Misaka neural networks can compose predictions using different individuals' pupillary dilation signals to give a more accurate overall prediction than even from the highest performing single individual, reaching an accuracy of 88.9%. Further research will involve the investigation of the correlation between two groups of high-performing predictors using verbal answers and pupillary dilation signals.

Keywords: Emotion veracity · Neural networks · Pupillary dilation

1 Introduction

Dilation of the pupil reflects a range of cognitive processes including interest [10], motivation [23], and emotionality [28]. Research has revealed that pupil dilation reflects the mechanisms of creating and retrieving memories [11]. Specifically, pupil dilation is positively correlated with people's confidence in their memories [21]. Hence, pupillary reflex has persisted as an index of cognitive demand [13]. Later research suggests pupillary reactions *summed index* of brain processes during cognitive activities [16]. In particular, pupillary dilation reflects the inner activity of the autonomic nervous system, which is vital for maintaining the equilibrium of the body [17], and is hence not under conscious control.

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11305, pp. 299–310, 2018.
https://doi.org/10.1007/978-3-030-04221-9_27

The autonomic nervous system consists of two sub-systems, namely the sympathetic nervous system and the parasympathetic nervous system [28]. The former encourages whereas the latter inhibits pupil dilation. That is, these two sub-components operate in an antagonistic fashion to support the processes of the autonomic nervous system [17], the effects of which are visible via pupil dilation. Studies on the relationship between cognitive activities and pupillary dilation often utilize discrete stimuli, since emotional stimuli can cause significant effects on the autonomic nervous system [1]. These stimuli can also result in distinctive waveforms corresponding to different mental activities [28]. For example, the pupil dilates when people observe beneficial as well as adverse pictures [27].

Emotions can be characterized as a combination of two dimensions, arousal and valence [18]. Arousal corresponds to how strong the emotion is, and valence shows how positive the emotion is, as Fig. 1 indicates. Despite the simplicity of this model for representing emotions, researchers have successfully utilized it in a range of tasks, such as emotion recognition and memory studies [15,22,29].

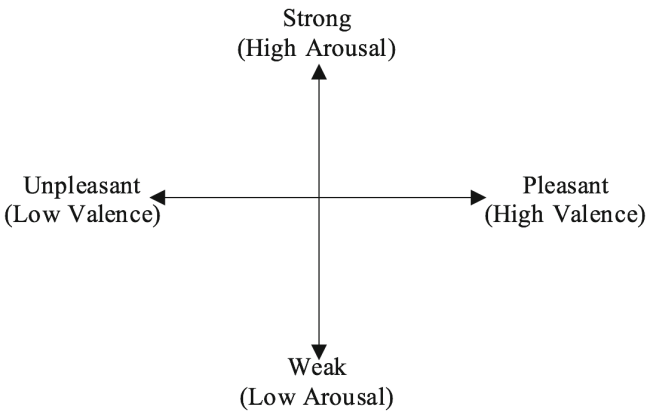

Fig. 1. The model of emotions [4].

Facial expressions are effective means of communication that can convey emotional information more promptly than languages, with which humans can rapidly recognize affective states of others [3]. Unlike the physiological signals mentioned above that are directly governed by the autonomic nervous system, people may perform acted facial expressions that contradict the expresser's true affective state [5]. For example, a salesperson may present acted smiles to pretend friendly attitudes. Research has indicated six basic facial expressions that are readily recognizable across dissimilar cultural backgrounds, namely anger, happiness, fear, surprise, disgust, and sadness [3]. In this paper, we consider anger due to its fundamental importance as a basic emotion.

Due to the divergent emotional strengths of genuine and acted emotions on the autonomic nervous system, there may exist differences between the physiological signals corresponding with genuine and acted emotions. Previous work done by Qin [25] revealed that peoples' capabilities for discerning the veracity of emotions vary. Neural networks can leverage this variety by assigning positive and negative weights to high and low accuracy discerners in order to aggregate peoples' responses and give a final higher precision prediction [25]. Preliminary statistical tests also indicate that people's physiological signals differ on distinguishing genuine and acted anger videos. That is, some participants' physiological signals corresponding to the two different kinds of videos vary more than others. Thus, we hypothesize that we can also aggregate physiological signals from different participants in order to give higher accuracy prediction of the source video label. That is, whether the stimulus is genuine or acted anger.

In this paper, we propose novel Misaka networks, which can predict the veracity of a person's expressed emotions by aggregating various observers' pupillary dilation signals. While previous related work focuses on using psychological signals from single participant, we novelly utilized physical signals from multiple individuals and showed better results. This research can be potentially applied to identify the true emotion of a person from his or her observers. Compared with collecting verbal answers from participants, utilizing physiological signals does not require a further process of interviewing and can possibly predict others' emotions ad-hoc and in-time. That is, if one can obtain observers' dynamic pupillary dilation signals in a timely fashion, one may predict the observed person's current emotion in real time.

This paper is organized as follows: We will introduce the structure of our Misaka neural networks using crowdsourcing techniques. Then, we will report our results and present discussion on the results obtained. We conclude this paper with a discussion of the limitations in our work and future work to tackle those limitations.

2 Method

2.1 Stimuli

This paper utilizes the pupillary dilation signals used by Chen [5]. The elicitation stimuli are videos sourced from YouTube, with 10 each corresponding with genuine and acted anger, respectively. We choose anger as the emotion to study because anger is one of the six basic emotions that are identifiable independent from cultural backgrounds [7], so we hypothesize that the results of this paper on anger can be generalized to other primary emotions.

Genuine emotion expressions were collected from live news reports and documentaries and acted ones were sourced from movies containing similar scenes. These videos were picked to balance ethnicity, gender and background context as far as possible. Further, they have been processed with greyscale normalization to reduce the differences between videos other than the veracity of emotions.

Avezier indicated that the contextual backgrounds for demonstrating emotional expressions are vital in order for humans to effectively discern different emotions [2]. Therefore, the stimuli adopted in this paper retain some contextual backgrounds to better simulate scenes from daily life.

Due to the different number of frames of different stimuli, we remove the two shortest videos due to their significantly lower number of frames, namely 60 and 89. Then, we truncate all the stimuli to only include the beginning 105 frames, which is the number of frames for the shortest remaining stimulus videos. Moreover, although there were 20 participants who took part in our experiment in total, due to the fact that some people were absent from some experiments, we only have 12 participants' complete pupillary dilation signals with all the videos. Thus, we only utilize data from these 12 participants in this study. This degree of loss of data is within the normal range with the mobile sensors use, trading non-intrusiveness and hence more natural behavior for occasional data loss.

Figure 2 demonstrated how the experiments were conducted. Participants were provided with oral instructions by the experimenters prior to the experiments. As Fig. 2 indicates, pupillary dilation was tracked with a remote Eye Tribe tracker at 60 Hz [6]. Furthermore, we also collected the subjects' skin conductance, blood volume pressure, and heart rate during the same experiments. They have not been utilized in the analysis of this paper and may be used for future work.

Fig. 2. The experimental setup [5].

2.2 Neural Networks

Artificial neural networks are simulations of animal brain information processing. Theoretically, a multiple layer feedforward neural network can approximate any measurable functions [14]. Due to the recent increase of computational power and the vast amount of data, neural networks now have dramatically improved in applications, such as object detection and speech recognition [26].

2.3 Crowd Prediction

Social science researchers have demonstrated promising results of utilizing the crowd to predict future outcomes [30]. For example, the prediction results of five US presidential elections using crowd prediction were more accurate than the traditional polls, where the latter was basically random guessing [19]. Crowd prediction has also been extended to a range of industrial applications, including in healthcare companies and technology corporations [24].

Further research on crowd prediction has also acknowledged that people vary in capabilities. That is, instead of assembling every predictor's opinion into equal consideration, top-performing predictions will be extracted first and their answers synthesized as the representation responses of the whole crowd [20]. This elite-based method showed a 50% greater accuracy than composing crowd fore-casting teams [8].

2.4 Misaka Networks

In this paper, we combine both neural network and crowd prediction techniques in order to predict a person's true emotion from observer reactions to their video performance. A Misaka network includes a collection of feedforward neural networks to predict whether a pupillary dilation physiological signal corresponds to a genuine or acted anger. The name Misaka network? was inspired by a Japanese anime where clones of Misaka can demonstrate much more powerful capabilities when working as a cohesive group than as individuals [12]. Each of these neural networks is trained to predict whether a pupillary dilation signal belongs to genuine or acted anger, using one participant's data, trained on that participant's reactions to 18 videos. We call these neural networks discerners. In our case, we have trained 12 discerners corresponding to the 12 participants. Specifically, a discerner has an input layer with 105 nodes (corresponding to 105 frames), a hidden layer with 100 nodes and an output layer with 2 nodes. Each discerner is trained using leave-one-out cross-validation.

A Misaka network also contains another feedforward neural network to combine all the discerners' responses for one video, based on the discerners' previous accuracies. We call this neural network an aggregator. For example, if the accuracy of discerner i was higher previously than discerner j, then the aggregator should assign more weight to a future prediction from discerner i. Moreover, the aggregator should also be able to reverse answers from the poor-performing discerners, and in general, learn to best aggregate the information from the discerners. An aggregator has an input layer with 12 nodes, a hidden layer with 120 nodes and an output layer with 2 nodes.

As Fig. 3 indicates, we collected 18 pupillary dilation time-series signals from each participant by letting him or her watch 18 videos, in an order balanced fashion to eliminate the effects of presentation order. Among these 18 videos, 10 correspond to genuine and the other 8 correspond to acted anger, respectively. Then, we train a discerner to predict the source of a signal. That is, whether a signal comes from watching a genuine anger video, or an acted one. We conduct

leave-one-video-out cross-validation for each discerner to predict the label of every video. Thus, we utilize 17 signals to train and let the discerner predict the source of the remaining one signal. As a result, for each of the 12 discerners, we will have 18 predicted results for each video.

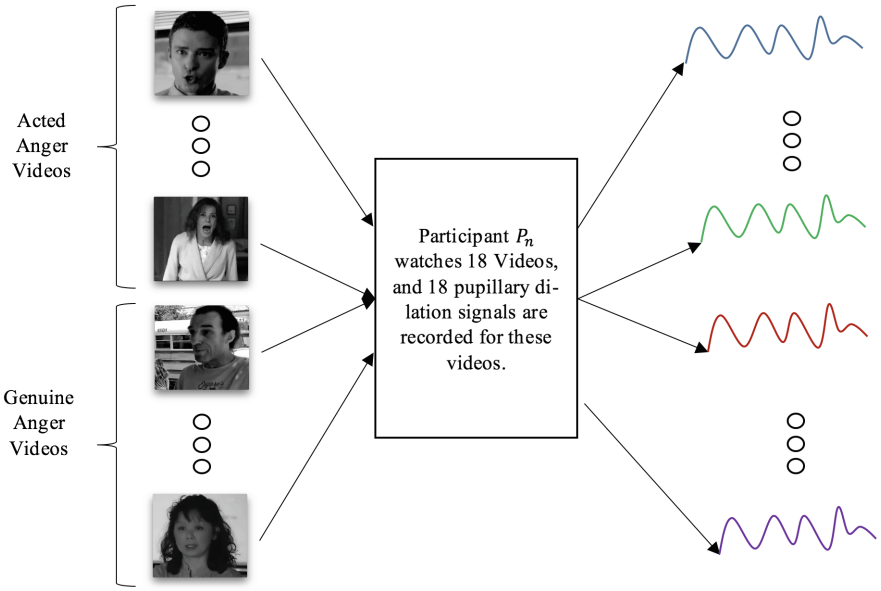

Fig. 3. An illustration of recording pupillary dilation signals from a participant.

Subsequently, we utilize the aggregator to learn the reliability of discerners in order to best combine their answers to give more accurate predictions. In our case, the input data is an 18×12 matrix, where each feature row represents a discerner's prediction for a given anger video. For example, a vector $<1, 0, ..., 1, 1>$ means that the predictions from the first two, ..., and the last two regarding an anger video are genuine, acted, ..., genuine and genuine, respectively. That is, the aggregator will learn from the historical answers from discerners in order to determine their reliabilities. Based on these reliabilities, the aggregator will eventually combine all the discerners' opinions and ideally, give a more precise prediction. We also conducted leave-one-video-out cross-validation for the aggregator. Nevertheless, due to our limited amount of participants, we test on *the same* videos. However, the aggregator will not know the correct labels, it only learns the credibilities from the discerners. Therefore, using the same data again will not result in unreliable issues.

Our previous work on crowdsourcing verbal responses has indicated that a minimal number of participants being 20 may be necessary for an accurate aggregator to work properly [25]. Here, we only have 12 participants. In order to reduce the unexpected effects caused by insufficient participants, we only utilize

discerners that are strongly accurate or inaccurate (if significantly worse than chance) by only considering those outside $50\% \pm 10\%$ accuracy. This is because we wish to minimize the learning pressure on the aggregator by removing inputs that may confuse it, as discerners close to 50% (the chance value) provide noisy signals.

3 Results and Discussion

3.1 Discerning Reliabilities of Detecting Anger Differ with People's Varying Pu0pillary Dilation Signals

When observing genuine and acted anger facial expressions, people's pupillary dilation signals corresponding to those two kinds of emotional expressions can vary. Our preliminary analysis indicated this from Student's t-tests. For each participant, we averaged each signal so that we would collect $10 + 8$ means, which correspond to genuine and acted stimuli, separately. The calculated statistical significances among the first 8 pairs of signal means differ among different participants, as Table 1 indicates.

Table 1. Statistical t-test significance calculated from the means of every individual's genuine and acted pupillary dilation signals.

Participant	P1	P2	P3	P4	P5	P6
T-test significance	0.8856	0.3494	0.6720	0.5765	0.8530	0.3363
Participant	P7	P8	P9	P10	P11	P12
T-test significance	0.0036	0.7592	0.2997	0.4486	0.5623	0.5959

In more detail, we may interpret the t-test significances as a reflection of the deviance between signals corresponding to genuine and acted anger emotions. These different values imply that some people's pupillary dilation signals may be more distinguishable for the veracity of the source anger videos than others.

Table 2 shows the accuracy of discerners, each corresponding to a participant's pupillary dilation signals. For instance, discerner D1 is trained with participant P1's physiological signals. The accuracy is calculated by averaging the cross-validation results of predicting the signal source with a discerner. Specifically, taking discerner D1 as an example, we train it with 17 signals from participant P1. Then we let D1 predict whether the remaining signal from P1 is sourced from a genuine anger video, or an acted one. The accuracy of D1 is defined as the proportion of correctly predicted signal sources over the total number of signals. In our data, discerner D1 demonstrated the highest accuracy, whilst discerner D5 showed a poor accuracy with merely 27.8%, which is noticeably worse than chance. We could speculate that this participant has had an unusual emotional background, such as only encountering anger in videos, and so judges genuine anger incorrectly, consistently.

Table 2. Different discerners have distinct prediction accuracies, with 6 close to chance.

Discerner	D1	D2	D3	D4	D5	D6
Accuracy	83.3%	44.4%	50.0%	50.0%	27.8%	50.0%
Discerner	D7	D8	D9	D10	D11	D12
Accuracy	50.0%	38.9%	44.4%	61.1%	66.7%	38.9%

3.2 Aggregating the Prediction Results from Various People's Pupillary Dilation Signals Can Increase the Accuracy of Prediction

As discussed previously, we used **the same** data for aggregation due to the limited amount of collected signals. The aggregator result is 88.9% accurate, showing that an aggregator, by combining multiple discerners' predictions based on their prior response accuracies, can outcompete the highest-accuracy discerner. The aggregator, as illustrated in Fig. 4, successfully learned the pattern of their reliabilities and that it should assign more accurate discerners like D1 mostly positive weights. Conversely, it gave poor-performing ones like D5 mostly negative weights. Eventually, this aggregator demonstrated 88.9% accuracy with cross-validation.

3.3 A Mathematical Explanation for the Feasibility of Misaka Networks

The problem faced by Misaka networks can be formally abstracted as given a collection of N discerners $D_1, D_2, ..., D_N$, each with an accuracy $A_1, A_2, ..., A_N$. If these N discerners have reached a consensus on predicting a binary result as 1, what is the probability of the result actually being 1?

We apply Bayes' theorem. The probability of the result R being 1 can be expressed as

$$\begin{aligned} &P(R=1|D_1=1, D_2=1, ..., D_N=1) \\ &= \frac{P(D_1=1, D_2=1, ..., D_N=1|R=1) \times P(R=1)}{P(D_1=1, D_2=1, ..., D_N=1)} \end{aligned} \tag{1}$$

With a further assumption of $P(R=1) = P(R=0) = 0.5$ and $D_1, D_2, ..., D_N$ are conditionally independent given R, Eq. 1 can be further expressed as

$$\frac{\prod_{i=1}^{N} P(D_i=1|R=1) \times P(R=1)}{\prod_{i=1}^{N} P(D_i=1|R=1) \times P(R=1) + \prod_{i=1}^{N} P(D_i=1|R=1) \times P(R=0)} \tag{2}$$

For example, given we have two discerners D_1 and D_2 with accuracies A_1 being 0.8 and A_2 being 0.7 and they have reached a consensus predicting the result being 1. Formula 2 tells us that the overall probability of the result being 1 is approximately 0.9032, which is higher than 0.8.

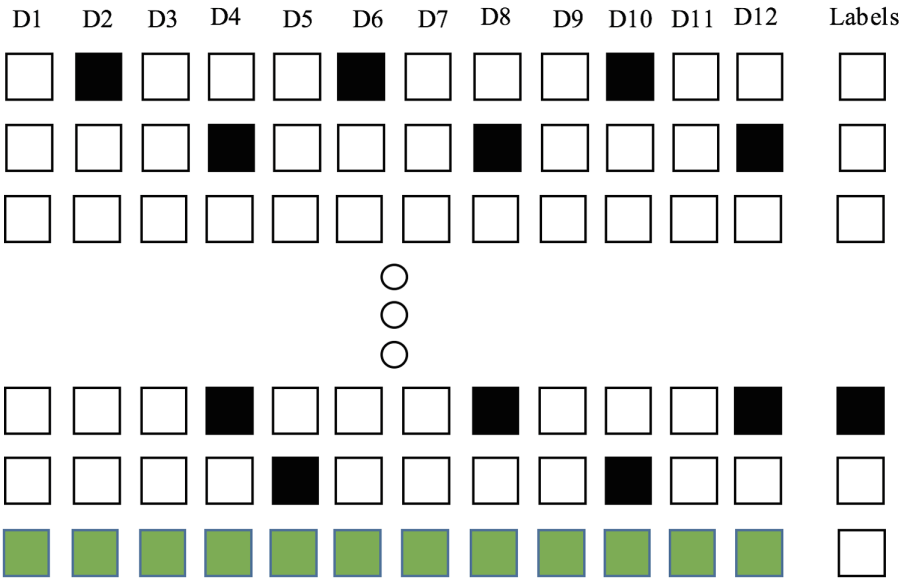

Fig. 4. An illustration of an aggregator. White and black represent genuine and acted anger respectively. Each row represents predictions from different discerners regarding the same video, whose correct label is given in the last "labels" column. The aggregator learns from the first 17 to predict the label for the last green row (genuine in the example shown). (Color figure online)

4 Limitations and Future Work

4.1 Limitations

The number of stimuli is limited, specifically, only 18 videos, also there are a limited number of participants. Thus, it is infeasible to split these videos into two groups, where the first half is to train discerners including cross-validation within that group, and then use the second half to train and test the synthesizer by predicting the genuineness of each video, based on the previous learnt accuracies of each discerner, again using cross-validation on that group. This is because it requires about 20 videos to consistently infer the reliability of each discerner. Instead, cross-validation was undertaken with the same videos for both discerners and aggregators. Nonetheless, our reported results are still relatively reliable since discerners will not provide correct labels to aggregators. In the future, we will collect more videos containing anger and other expressions as stimuli in order to more thoroughly test a Misaka network by splitting videos into two groups, one for training all the discerners and the other for training and testing the aggregator.

4.2 Future Work

Further, people's capabilities on discerning the veracity of emotions also differ when they verbalize their thoughts. Therefore, we will investigate whether there exists a correlation between the two groups of high-performing individuals, being verbal high-performers and physiological signal high-performers. That is, we would like to discover whether the people who are capable of giving quality emotional discerning results from verbalizing produce even more discriminating physiological signals, or whether they are just in better touch with their bodies/emotions, or to discover whether it is possible to be more correct verbally than from physiological signals. We suspect that the degree of emotional valence in the stimuli may have a substantial effect on these alternatives. That is, more exaggerated acted expressions may make people put more belief on their genuineness. We may also investigate whether fuzzy logic can help in assembling predictions from individual psychological signals [9].

5 Conclusion

We introduced our Misaka networks, which use reliability signals to aggregate outputs of discerner networks trained on participants' raw physiological signal data. We achieved state-of-the-art results on the (small) sizes of datasets common in close-to-real-world recording of emotional and physiological data.

In summary, we discovered that people's pupillary dilation signals vary in their ability to discern the veracity of anger facial expressions. This variety ranges from 27.8% to 83.3%. We can leverage these differences by training another neural network to learn these patterns of these different reliabilities in order to assign appropriate weights for each participant. After aggregating answers from different participants' physiological signals, the prediction accuracy can be boosted to 88.9%. This combination of discerning networks trained on individuals and an aggregator trained on their reliability compose our Misaka network.

Acknowledgments. The authors acknowledge Dongyang Li, Liang Zhang and Zihan Wang for the suggestion of applying Bayes' theorem in the probability calculation.

References

1. Andreassi, J.L.: Psychophysiology: Human Behavior & Physiological Response, 5th edn. Lawrence Erlbaum Associates Publishers, Mahwah (2007)
2. Aviezer, H., Hassin, R., Bentin, S., Trope, Y.: Putting facial expressions back in context. In: Ambady, N., Skowronsky, J.J. (eds.) First Impressions, chap. 11, pp. 255–286. Guilford Press, New York (2008)
3. Batty, M., Taylor, M.J.: Early processing of the six basic facial emotional expressions. Cogn. Brain Res. **17**(3), 613–620 (2003)
4. Chanel, G., Ansari-Asl, K., Pun, T.: Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: 2007 IEEE International Conference on Systems, Man and Cybernetics, pp. 2662–2667 (2007)

5. Chen, L., Gedeon, T., Hossain, M.Z., Caldwell, S.: Are you really angry?: detecting emotion veracity as a proposed tool for interaction. In: Proceedings of the 29th Australian Conference on Computer-Human Interaction, Brisbane, Queensland, Australia, pp. 412–416. ACM (2017)
6. Dalmaijer, E.: Is the low-cost eyetribe eye tracker any good for research? Technical report, PeerJ PrePrints (2014)
7. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)
8. Frood, A.: Work the crowd. New Sci. **237**(3166), 32–35 (2018)
9. Gao, Y., Xiao, F., Liu, J., Wang, R.: Distributed soft fault detection for interval type-2 fuzzy-model-based stochastic systems with wireless sensor networks. IEEE Trans. Ind. Inf. (2018, early access version)
10. de Gee, J.W., Knapen, T., Donner, T.H.: Decision-related pupil dilation reflects upcoming choice and individual bias. Proc. Nat. Acad. Sci. **111**(5), E618–E625 (2014)
11. Goldinger, S.D., Papesh, M.H.: Pupil dilation reflects the creation and retrieval of memories. Curr. Dir. Psychol. Sci. **21**(2), 90–95 (2012)
12. Haimura, M.: A Certain Magical Index. ASCII Media Works, Tokyo (2013)
13. Hess, E.H., Polt, J.M.: Pupil size in relation to mental activity during simple problem-solving. Science **143**(3611), 1190–1192 (1964)
14. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
15. Jirayucharoensak, S., Pan-Ngum, S., Israsena, P.: EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. Sci. World J. (2014)
16. Kahneman, D., Beatty, J.: Pupil diameter and load on memory. Science **154**(3756), 1583–1585 (1966)
17. Kim, K.H., Bang, S.W., Kim, S.R.: Emotion recognition system using short-term monitoring of physiological signals. Med. Biol. Eng. Comput. **42**(3), 419–427 (2004)
18. Lang, P.J.: The emotion probe: studies of motivation and attention. Am. Psychol. **50**(5), 372 (1995)
19. Manski, C.F.: Interpreting the predictions of prediction markets. Econ. Lett. **91**(3), 425–429 (2006)
20. Mellers, B., et al.: Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspect. Psychol. Sci. **10**(3), 267–281 (2015)
21. Papesh, M.H., Goldinger, S.D., Hout, M.C.: Memory strength and specificity revealed by pupillometry. Int. J. Psychophysiol. **83**(1), 56–64 (2012)
22. Partala, T., Jokiniemi, M., Surakka, V.: Pupillary responses to emotionally provocative stimuli. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, pp. 123–129. ACM (2000)
23. Pletti, C., Scheel, A., Paulus, M.: Intrinsic altruism or social motivationwhat does pupil dilation tell us about children's helping behavior? Front. Psychol. **8**, 2089 (2017)
24. Polgreen, P.M., Nelson, F.D., Neumann, G.R., Weinstein, R.A.: Use of prediction markets to forecast infectious disease activity. Clin. Infect. Dis. **44**(2), 272–279 (2007)
25. Qin, Z., Gedeon, T., Caldwell, S.: Neural networks assist crowd predictions in discerning the veracity of emotional expressions. arXiv Preprint arXiv:1808.05359 (2018)
26. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)

27. Steinhauer, S.: Pupillary dilation to emotional visual stimuli revisited. Psychophysiology **20**, S472 (1983)
28. Steinhauer, S.R., Siegle, G.J., Condray, R., Pless, M.: Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. Int. J. Psychophysiol. **52**(1), 77–86 (2004)
29. Wagner, J., Kim, J., André, E.: From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: IEEE International Conference on Multimedia and Expo, ICME 2005, Amsterdam, Netherlands, pp. 940–943. IEEE (2005)
30. Wolfers, J., Zitzewitz, E.: Prediction markets. J. Econ. Perspect. **18**(2), 107–126 (2004)