Hidden Emotion Detection using Multi-modal Signals

Dae Ha Kim

kdht5022@gmail.com

Department of Electronic Engineering, Inha University
Michuhol-Gu, Incheon, Republic of Korea

Byung Cheol Song

bcsong@inha.ac.kr

Department of Electronic Engineering, Inha University
Michuhol-Gu, Incheon, Republic of Korea

Figure 1: A representative example of hidden emotion. A subject who is watching a horror movie maintains neutral facial emotion but feels fear inside.

ABSTRACT

In order to better understand human emotion, we should not recognize only superficial emotions based on facial images, but also analyze so-called inner emotions by considering biological signals such as electroencephalogram (EEG). Recently, several studies to analyze a person's inner state by using an image signal and an EEG signal together have been reported. However, there have been no studies dealing with the case where the emotions estimated from the image signal and the EEG signal are different, i.e., emotional mismatch. This paper defines a new task to detect hidden emotions, i.e., emotions in a situation where only the EEG signal is activated without the image signal being activated, and proposes a method to effectively detect the hidden emotions. First, when a subject hides the emotion intentionally, the internal and external emotional characteristics of the subject were analyzed from the viewpoint of multimodal signals. Then, based on the analysis, we designed a method of detecting hidden emotions using convolutional neural networks (CNNs) that exhibit powerful cognitive ability. As a result, this study has upgraded the technology of deeply understanding inner emotions. On the other hand, the hidden emotion dataset and source code that we have built ourselves will be officially released for future emotion recognition research.

CCS CONCEPTS

• **Human-centered computing** → **Human-computer interaction (HCI)**; • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Hidden emotion detection, inner emotion, Convolutional Neural Networks

ACM Reference Format:

Dae Ha Kim and Byung Cheol Song. 2021. Hidden Emotion Detection using Multi-modal Signals. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts)*, May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3411763.3451721>

1 INTRODUCTION

The face is a part in which psychologically and physiologically important stimuli are expressed, and it shows the inner emotion [15]. In other words, emotions through facial expressions can convey complex information such as a person's psychological state and intention. So, it is a natural approach to focus on facial expressions to analyze human emotions. However, since emotions are influenced by various environmental factors, there may be a stronger emotional change in the inner state than in the facial expression, or vice versa [11]. For example, a facial expression may be positive, but the corresponding inner emotional state may be negative. Therefore, if visual signals such as facial expressions and biological (bio) signals such as electroencephalogram (EEG) that can detect internal emotional changes are considered together, the true emotions can be better understood. In this paper, such a technology that recognizes

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8095-9/21/05...\$15.00

<https://doi.org/10.1145/3411763.3451721>

emotions based on two or more signals is defined as multimodal emotion recognition.

Multimodal emotion recognition has two main-streams: 1) Emotion recognition when heterogeneous signals, e.g., visual and physiological signals, express the same emotion [4], 2) emotion recognition when facial expressions and actual inner emotions do not match, that is, mismatch negativity (MMN) [15, 20]. However, facial expressions and inner emotional states do not always match [3]. In other words, mismatch between inner emotions and facial expressions often occurs. Therefore, detecting and analyzing mismatches such as MMN will be a way to recognize inner emotions precisely.

Existing MMN studies analyzed mismatch cases in a limited environment, and did not use advanced tools such as convolutional neural networks (CNNs) [9, 20]. Meanwhile, several techniques for recognizing emotions by analyzing images and bio-signals through CNNs have been proposed [4]. However, studies that experimentally analyze mismatch between heterogeneous signals do not exist yet, to our knowledge.

In this paper, the inner state where the facial expression maintains a neutral emotion while the bio-signal is activated is defined as a hidden emotion. Also, this paper proposes a method of analyzing feature maps of video signals and EEG signals through CNNs [13] to effectively recognize hidden emotions. Figure 1 describes the concept of the proposed hidden emotion detection (HED) task. For example, when a subject is watching a horror movie, multimodal data is acquired using a camcorder and an EEG signal measurement tool. Here, the subject was intentionally instructed to hide his/her emotion from the facial expression. In other words, the subject's external emotional state is neutral, but the internal emotional state is negative, e.g., fear. Through this HED task, we can analyze the inner and outer emotional characteristics that appear when emotions are intentionally hidden.

Specifically, the process for analyzing and detecting hidden emotions is as follows: 1) We acquire facial expressions and EEG signals of subjects concurrently who are watching various emotion-inducing video clips. Here, the image signal is pre-processed through a face region detector [14], and the raw EEG signal is also pre-processed through noise removal and normalization (Fig. 2(a)). 2) We extract feature maps from image and EEG signals using CNNs [4], and calculate the emotional level based on the extracted feature maps. Here, the emotional level is not classified into discrete categories (e.g., happy, fear, disgust, neutral, happy, sad, surprise) proposed by Ekman [5], but is regressed as the continuous valence of the Arousal-Valence (AV) domain proposed by Russell [17] (Fig. 2(b)). 3) We predict the valence value by applying adaptive fusion to feature maps obtained from multimodal signals. If at least a section where the estimate and the ground truth (GT) match exists, it is regarded as the correct case (Fig. 2(c)). Here, GT is determined by subjects' self-assessments (see Section 4.1).

The contribution points of this paper are as follows:

- For a deeper understanding of human inner emotions, so-called hidden emotions were defined and an experimental configuration to detect them was proposed. This suggests a new research direction different from the existing emotional recognition studies that assume that facial expressions and inner emotional states coincide.
- This study can be used for various applications (e.g., lie detection) to identify intentions that are not well-disclosed in facial expressions. In addition, the built-up HED dataset will be officially released soon.

2 RELATED WORKS

This section surveys the trend of emotion recognition research based on multimodal signals, and reviews existing studies on MMN analysis for detecting hidden emotions.

2.1 Multi-modal emotion recognition

In general, multimodal emotion recognition indicates a technique to fuse physiological signals such as EEG, skin resistance, and EMG/EOG through a machine learning tool, and classify emotions in the discrete domain based on the fusion information [21]. However, fusion of signals with completely different characteristics, such as image signals and bio signals, has been rarely studied. On the other hand, there are studies that have provided an affective model based on EEG and eye tracking data [22]. [22] is noteworthy in that it is a frontier emotional recognition using both visual and EEG signals. However, since [22] employed a classical way such as principal component analysis (PCA) in the feature map extraction process of multimodal signals, it did not realize sophisticated feature map analysis. In a similar manner, a study on emotion recognition based on multimodal signals [10] was reported, but the method also did not show acceptable performance. Recently, [4] has demonstrated the synergy of the video and EEG signals by fusing the two heterogeneous signals through the latest neural network and a low-rank decomposition technique. In practice, [4] achieved successful performance for well-known multimodal emotion recognition datasets, but this method is based on the assumption that multimodal signals represent the same emotional state.

2.2 Mismatch negativity (MMN) analysis for hidden emotion detection

Mismatch negativity (MMN) is based on the analysis of error signals occurring in the process of recognizing multimodal signals [20]. Specifically, the difference between sparse stimulus and frequently occurring stimulus is defined as a component of MMN [15]. For example, in the HED task, an EEG signal corresponds to a stimulus that occurs infrequently, and a visual signal corresponds to a stimulus that occurs frequently. If an EEG signal shows a negative emotion despite a neutral facial expression, this corresponds to a mismatch, i.e., a hidden emotion. A few HED studies based on facial expression analysis were reported [7], but they adopted only facial micro-expression as a clue for HED. So, for instance, if a subject hides the emotion in the facial expression, the approach of [7] may not actually work.

3 METHODS

The proposed HED framework is based on a multimodal CNN network and relies on the adaptive fusion of valences generated therefrom. Additionally, a neural network-based fusion module is proposed for the cases that the CNN network does not operate properly. Thus, the purity of bio-signal information can be improved (see Figure 3). This section describes the multimodal signal-based

Figure 2: The overall framework of hidden emotion detection.

Figure 3: The inference and training framework for HED. When the fusion module is used in the inference stage, y_e is properly scaled according to y_a 's guide.

CNN network, adaptive fusion, and fusion module, which are the basis of the proposed HED framework, one-by-one.

3.1 Visual and EEG networks

Let x_v be an input image corresponding to a face region detected by a CNN-based face detector [14]. And, let x_e be an EEG data that is a pre-processed output of the raw EEG signal acquired from a specific EEG sensor (refer to Section 4 for details on pre-processing). First, a 2D CNN f_v that receives x_v and outputs a valence domain emotion value y_v is used as an image analyzer [13]. In addition, a long-short term memory (LSTM) f_e that receives x_e and outputs a valence domain emotion value y_e is used as a biometric analyzer [18]. Note that f_e uses the 16-frame-unit EEG data $\{x_e\}_{i=1}^6$ as an input, whereas f_v receives individual x_v as an input. Here, i refers to the frame index. The next section deals with the fusion process based on $\{y_v, y_e\}$ which are the multimodal valences obtained by the image/biometric analyzers.

3.2 Adaptive fusion

The goal of HED is to detect the instances when significant y_e occurs without significant y_v . We try to realize this goal through

adaptive fusion (AF) according to the decision rule based on Eq. (1).

$$y = \begin{cases} \frac{y_v + \left(1 + \frac{|y_e - \delta|}{100}\right)y_e}{2}, & \text{if } |y_e| > \delta \\ \frac{y_v + y_e}{2}, & \text{otherwise} \end{cases} \quad (1)$$

where the threshold δ was experimentally set to 0.4. If the absolute value of y_e is greater than δ , Eq. (1) gives more weight to y_e . Otherwise, the average operation is simply performed. As a result, the fusion score y of the multimodal valence is calculated by Eq. (1). As shown in Fig. 2(c), when y reaches the GT value, the hidden emotion of the subject is found. However, if the image and EEG signals derive the valence values of opposite signs, this rule-based AF can produce an ambiguous result. In order to solve this problem, an advanced adaptive rule is required as follows.

3.3 Fusion module

The fusion module is designed to detect various hidden emotions. The proposed fusion module consists of three fully-connected (FC) layers [13], where 64, 128, and 1-dimensional linear transformation operations are performed for each layer, respectively. In all layers, the rectified linear unit (ReLU) [8] is used as the activation function.

The output value of the FC layer is defined as the emotion guidance value y_a that comprehensively reflects the valence of the multimodal signal, and y_a is trained to minimize the distance from the corresponding GT value of the HED training dataset. The objective function to minimize the distance is defined using the root mean squared error (RMSE) as follows.

$$L(y_a, y_{GT}) = \|y_a - y_{GT}\|_2^2 + \lambda \|y_a - \mu\|_2 \quad (2)$$

where $\|\cdot\|_2$ is L2 norm, y_{GT} is the GT value of the HED training dataset, and λ is the balance factor. Note that in Eq. (2), the LASSO regularization term is added for the convergence stability of y_a , which should play a role of scaling y_e through multiplication operation. In other words, a constraint term is added so that y_a reflects y_{GT} and converges to a stable point. Here, μ is set to 1.0. In sum, y_e is appropriately scaled by y_a which was obtained by the trained FC layer, and then the AF process according to Eq. (1) is performed based on the scaling. Here, scaling can convert y_e to a positive or negative value.

4 EXPERIMENTAL RESULTS

This section introduces the equipment and detailed data acquisition process for building the HED dataset, and analyzes various subjects through intensive quantitative and qualitative experiments.

4.1 HED dataset configuration

4.1.1 Details of data acquisition. A total of 23 subjects participated in the dataset acquisition experiment. Visual signals were acquired with a Sony HDR 1080p camcorder with a frame rate of 50Hz. EEG signals were collected through 12 channel electrodes of the MP150 Biopac system [16]. Continuous wavelet transform was utilized to remove noise and eyeblink artifacts of the EEG signals. The sampling frequency of EEG signals is 256Hz, and the EEG signals were synchronized based on the frequency of the visual signals during the sampling process.

The self-assessment refers to a method of directly evaluating the subject's own emotions. In this case, the AV domain-based Russell's emotion model [17] was used. Given the emotion evaluation questionnaire whenever watching each video clip, the subject chose a proper direction among eight directions on the AV domain graph, which correspond to eight emotion categories. Simultaneously, the subject selected an emotion level among a total of three levels. Also, the subject pressed the pressure sensor button whenever he/she felt a certain emotion in order to grasp the subject's actual feelings. The pressure sensor data was acquired by synchronizing with the biological signal through the MP150.

4.1.2 Experimental configurations.

Sampling and evaluation criterion: We initially obtained a total of 246 video clips for evaluation purpose, and split them into 135 and 111 respectively for training the fusion module and evaluating hidden emotions. Among them, if the self-assessment result coincides with the sign of the induced emotion of the video clip, the corresponding to video clip was selected as an evaluation data. In addition, if signals (especially bio-signals) are simply vibrating (unstable), the corresponding sections were excluded from the evaluation dataset. Finally, 80 video clips out of 111 were selected as the evaluation dataset. The criteria for correct/incorrect answers

are as follows: If there is a section where the self-assessment result and the predicted value of the multimodal CNN network coincide, the answer is evaluated as a correct answer, and other cases are evaluated as incorrect. In the example of Fig. 4, since there is a section where the multimodal predicted value (blue) intersects the GT value (black), the corresponding video clip is treated as correct. **Training details:** The encoding network of visual/EEG signals used the parameters of [13, 18] as they were without additional parameter learning. Note that only the parameters of the fusion module are learned. At this time, about 8,900 parameters for the fusion module were learned through Adam [12] optimizer with learning rate 1e-4 and the objective function of Eq. (2). In this process, the batch size was set to 32.

4.2 Quantitative results

This section analyzes the experimental results for the HED evaluation dataset. Table 1 evaluates the quantitative performance. When using only AF, hidden emotions were accurately detected for 58 out of 80 video clips. In other words, 72.5% accuracy was achieved. For example, video clips #7 and #15, which correspond to happy emotions, show very high detection accuracy. On the other hand, video clips #6 and #13, which cause negative emotions such as sadness, show relatively low accuracy. In the case of such negative video clips, visual signals caused the wrong fusion result. Note that in order to avoid this phenomenon, we additionally employed the fusion module that allows negative emotions to be well detected even in case of mismatch (see Section 3.3). As in the 3rd column of Table 1, the performance was greatly improved by additionally recognizing a total of 9 negative cases as correct answers. Although there are still cases where the performance has decreased, such as video clip #15, it is clear that overall detection performance has improved.

4.3 Qualitative results

This section analyzes some of the results in Table 1 in units of frames. As shown in Fig. 4, a strong negative emotion (valence: -0.8) that could not be detected by AF alone was detected by the proposed fusion module. Note that the pressure sensor section which the subject actually felt as a peak is almost identical with the section in which a large change in the EEG signal occurs.

In addition, Fig. 5 shows a case in which the fusion module detects hidden emotions that could not be detected with AF alone. A notable point here is that the peak section has changed. Compared with the pressure sensor results, the peak region predicted by the fusion module almost coincided with the subject's emotion-inducing region. On the other hand, we were able to observe the subject's frequent pupil movements and eye blinking in the section where the valence changes significantly. Referring to the study [1] showing that the correlation between valence and eye movement actually exists, we found that the video signals can be the indirect clue of hidden emotions.

5 DISCUSSION

Through Fig. 4 and Fig. 5, we analyzed the subject's facial expressions as well as inner emotions obtained from the 2D CNN, LSTM, and AF modules in Fig. 3, and investigated the correlation between

Figure 4: Example of HED task results (Results of the 13th video clip of Subject #6). The expression of sad emotion that could not be captured by AF alone was detected through the fusion module.

Figure 5: Example of HED task results (Results of the 5th video clip of Subject #5).

the image and EEG signals from a mismatch perspective. When a subject hides the inner emotion, we can detect it with high accuracy through the fusion module. In addition, even when irregular phenomena such as eye blinking are found in facial expressions, this could be used as a clue for detecting hidden emotions. This study

using the latest deep learning technology such as CNN for HED task will certainly help improve the performance of the conventional emotion recognition applications that could not accurately grasp the human inner emotions. For example, to recognize the inner psychological state of a person who hides emotions with

Table 1: Quantitative result based on HED evaluation dataset.

Session #	AF w/o fusion module (Correct / Total)	AF w/ fusion module (Correct / Total)	Triggered emotion
01	4 / 6	4 / 6	Angry
02	5 / 5	5 / 5	Happy
03	5 / 8	6 / 8	Fear
04	3 / 5	3 / 5	Sad
05	4 / 7	5 / 7	Happy
06	1 / 3	3 / 3	Sad
07	6 / 7	6 / 7	Happy
08	2 / 5	5 / 5	Sad
09	5 / 6	5 / 6	Fear
10	7 / 8	7 / 8	Happy
11	4 / 5	5 / 5	Disgust
12	4 / 4	4 / 4	Happy
13	1 / 2	2 / 2	Sad
14	2 / 4	3 / 4	Fear
15	5 / 5	4 / 5	Happy
Accuracy (%)	72.5 (58/80)	83.75 (67/80)	-

bad intentions (e.g., lie detection and jury witness [2], etc.), the proposed method can be used as an auxiliary tool. If it is difficult to obtain an EEG signal, a simple biological signal such as a peripheral nervous system (PNS) [19] that has similar characteristics to EEG can be an alternative. Since the proposed fusion module has only a small number of learning parameters, it is possible to properly fuse the visual signal and the PNS signal. On the other hand, in this paper, the case where the visual information maintains a neutral emotion while the bio-signal is activated is defined as the only hidden emotion. However, in a real environment, the opposite also exists. Furthermore, as a very exceptional case, all of the multimodal signals are active, but the emotions derived from multiple signals may be totally different [6]. Therefore, modeling and analyzing such various mismatch cases through CNN will be our future research.

6 CONCLUSION

This study defined and modeled the mismatch of heterogeneous signals that can occur in various situations of multimodal emotion recognition as a hidden emotion detection task. Using powerful deep learning tools like CNN, we were able to effectively detect most of the mismatch cases. In addition, various internal emotions of subjects were analyzed from the perspective of multimodality. This study, which successfully analyzed and solved the emotional signal mismatch problem in multimodal emotion recognition, which has not been tried before, using CNN, will give great motivation to future emotion recognition studies.

ACKNOWLEDGMENTS

This work was supported by Institute of Information and communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) 2020-0-01389, Artificial Intelligence

Convergence Research Center (Inha University) and Industrial Technology Innovation Program through the Ministry of Trade, Industry, and Energy (MI, Korea) [Development of Human-Friendly Human-Robot Interaction Technologies Using Human Internal Emotional States] under Grant 10073154.

REFERENCES

- [1] Silke Anders, Martin Lotze, Michael Erb, Wolfgang Grodd, and Niels Birbaumer. 2004. Brain activity underlying emotional valence and arousal: A response-related fMRI study. *Human brain mapping* 23, 4 (2004), 200–209.
- [2] Iris Blandon-Gitlin, Kathryn Sperry, and Richard Leo. 2011. Jurors believe interrogation tactics are not likely to elicit false confessions: Will expert witness testimony inform them otherwise? *Psychology, Crime & Law* 17, 3 (2011), 239–260.
- [3] Bin Chen, Pei Sun, and Shimin Fu. 2020. Consciousness modulates the automatic change detection of masked emotional faces: Evidence from visual mismatch negativity. *Neuropsychologia* 144 (2020), 107459.
- [4] Dong Yoon Choi, Deok-Hwan Kim, and Byung Cheol Song. 2020. Multimodal attention network for continuous-time emotion recognition using video and EEG signals. *IEEE Access* 8 (2020), 203814–203826.
- [5] Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion* 98, 45–60 (1999), 16.
- [6] Rinie Geenen, Linda van Ooijen-van der Linden, Mark A Lumley, Johannes WJ Bijlsma, and Henriët van Middendorp. 2012. The match-mismatch model of emotion processing styles and emotion regulation strategies in fibromyalgia. *Journal of Psychosomatic Research* 72, 1 (2012), 45–50.
- [7] Enhao Gong, Tianhe Wang, and Jing Xiong. 2013. Hidden emotion detection through analyzing facial expression. *Technical reports* 4 (2013).
- [8] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 6789 (2000), 947–951.
- [9] Craig Hedge, George Stothart, Jenna Todd Jones, Priscila Rojas Frias, Kristopher Lundy Magee, and Jonathan CW Brooks. 2015. A frontal attention mechanism in the visual mismatch negativity. *Behavioural brain research* 293 (2015), 173–181.
- [10] Yongrui Huang, Jianhao Yang, Pengkai Liao, and Jiahui Pan. 2017. Fusion of facial expressions and EEG for multimodal emotion recognition. *Computational intelligence and neuroscience* 2017 (2017).
- [11] Yongrui Huang, Jianhao Yang, Siyu Liu, and Jiahui Pan. 2019. Combining facial expressions and electroencephalography to enhance emotion recognition. *Future Internet* 11, 5 (2019), 105.
- [12] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, Yoshua Bengio

- and Yann LeCun (Eds.). ICLR, Banff, Canada, 10. <http://arxiv.org/abs/1412.6980>
- [13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.
 - [14] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. 2015. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Boston, Massachusetts, 5325–5334.
 - [15] Xiying Li, Yongli Lu, Gang Sun, Lei Gao, and Lun Zhao. 2012. Visual mismatch negativity elicited by facial expressions: new evidence from the equiprobable paradigm. *Behavioral and Brain Functions* 8, 1 (2012), 1–10.
 - [16] Zhongzhe Li and Jian Kang. 2019. Sensitivity analysis of changes in human physiological indicators observed in soundscapes. *Landscape and Urban Planning* 190 (2019), 103593.
 - [17] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
 - [18] Jürgen Schmidhuber and Sepp Hochreiter. 1997. Long short-term memory. *Neural Comput* 9, 8 (1997), 1735–1780.
 - [19] Tongshuai Song, Guanming Lu, and Jingjie Yan. 2020. Emotion recognition based on physiological signals using convolution neural networks. In *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*. Proceedings of Machine Learning Research, Shenzhen China, 161–165.
 - [20] Gábor Stefanics, Motohiro Kimura, and István Czigler. 2011. Visual mismatch negativity reveals automatic detection of sequential regularity violation. *Frontiers in Human Neuroscience* 5 (2011), 46.
 - [21] Gyanendra K Verma and Uma Shanker Tiwary. 2014. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage* 102 (2014), 162–172.
 - [22] Wei-Long Zheng, Bo-Nan Dong, and Bao-Liang Lu. 2014. Multimodal emotion recognition using EEG and eye tracking data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, Chicago, Illinois, 5040–5043.