2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Deep Emotion Change Detection for Human-robot Interaction

ByungOk Han[1], Ho Won Kim[1], and Jang-Hee Yoo[1]

Abstract— **Emotion change detection via facial expression information is an important clue to non-verbal communication that can uncover emotional context. In a human-robot interaction scenario, figuring out the timing of emotion changes using facial expression information from an user has three advantages on: 1) providing a start point to obtain time-consistent multi-modal information, 2) reducing search space in time series, and 3) producing feedback information to improve the scenario. In this regard, we introduce an initial investigation to an automatic emotion change detection framework in the field of human-robot interaction. To tackle this issue, we propose a novel method of deep emotion change detection for inferencing emotion status and detecting multiple points of emotion changes. Incorporating these ideas, we provide evaluation methods to validate the framework and the baseline performance of our approach.**

I. Introduction

In our daily life, non-verbal communication between people occurs frequently and plays an important role in the interaction process. Non-verbal communication is conveying meaning through rhyme, intonation, facial expression, behavior, gaze, gesture, context information, and so on, not language-mediated communication. Verbal communication and non-verbal communication are complementary, but in some cases, a completely different meaning can be conveyed by non-verbal elements in a face-to-face conversation [1]. For example, ”stop” with a smile face and ”stop” with a neutral face can have a entirely different meaning. In particular, facial expression information for non-verbal communication plays a crucial role in the field of human-robot interaction (HRI) as visual cues to represent and perceive an emotional status. Recently, the importance of social intelligence for human-friendly and natural interaction with service robots has emerged, and exchange of emotion information through facial expressions has become an essential element for HRI.

From this point of view, researches on automatic facial expression recognition for HRI have focused on what facial expressions are being made. In other words, the user's emotional state is deduced by figuring out what facial expressions they are making. For instance, in [3], six class facial expressions were classified based on facial muscle movements and applied to an assistant robot for HRI. With the success of deep learning technology, various facial classification algorithms have been developed as in the survey paper [4]. Meanwhile, in [5], an affective state was estimated using valence-arousal representation, and [6] proposed a multi-modal emotion recognition system that fuses facial expressions, shoulder gesture and audio cues in valence-arousal space.

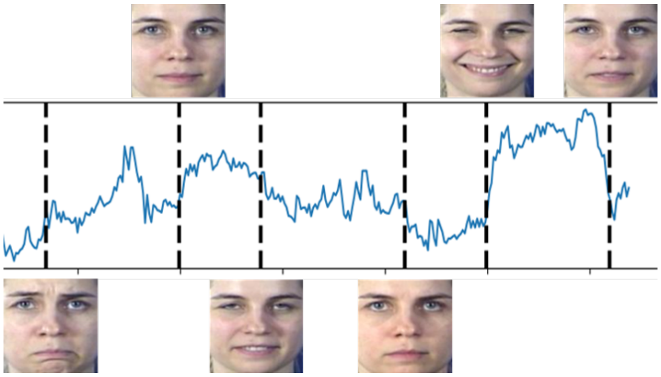

Fig. 1. Detection of emotion changes from a signal of emotion status. The dotted lines represent change points from the signal. The facial sequence is from the MMI [2] database

Emotion change detection through facial expressions, as well as emotion state recognition through facial expressions, is a vital clue to communication that can uncover social context [7]. In the field of HRI for social robots, figuring out the timing of emotion changes through facial expression information has three important meanings: 1) Robots can obtain time-consistent multi-modal information to estimate an accurate emotion status. It is difficult to deduce the accurate state of emotion with only the facial expressions [8]. To properly understand the state of emotion, non-verbal means of communication (contextual information such as intonation, behavior, gaze, gesture, atmosphere, etc.) should be interpreted at a point in time, enabling a comprehensive analysis. The timing of facial expression change can be used as the recognition point of a multi-modal system to infer emotional status, enabling fusion of time-consistent multi-modal information. 2) If multi-modal systems know when to obtain meaningful data, they can reduce search space over time series. Through emotion change detection, they can be computationally efficient. 3) A drastic change in an user's facial expression can be regarded as an abnormal event and can be used as a positive or negative feedback information to improve a HRI scenario. For example, if a robot detects a sudden negative change on an user's facial expression while communicating each other, it would change the subject of

*This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00330, Development of AI Technology for Early Screening of Infant/Child Autism Spectrum Disorders based on Cognition of the Psychological Behavior and Response)

[1]ByungOk Han, Ho Won Kim, and Jang-Hee Yoo are with the Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea byungok.han@etri.re.kr

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

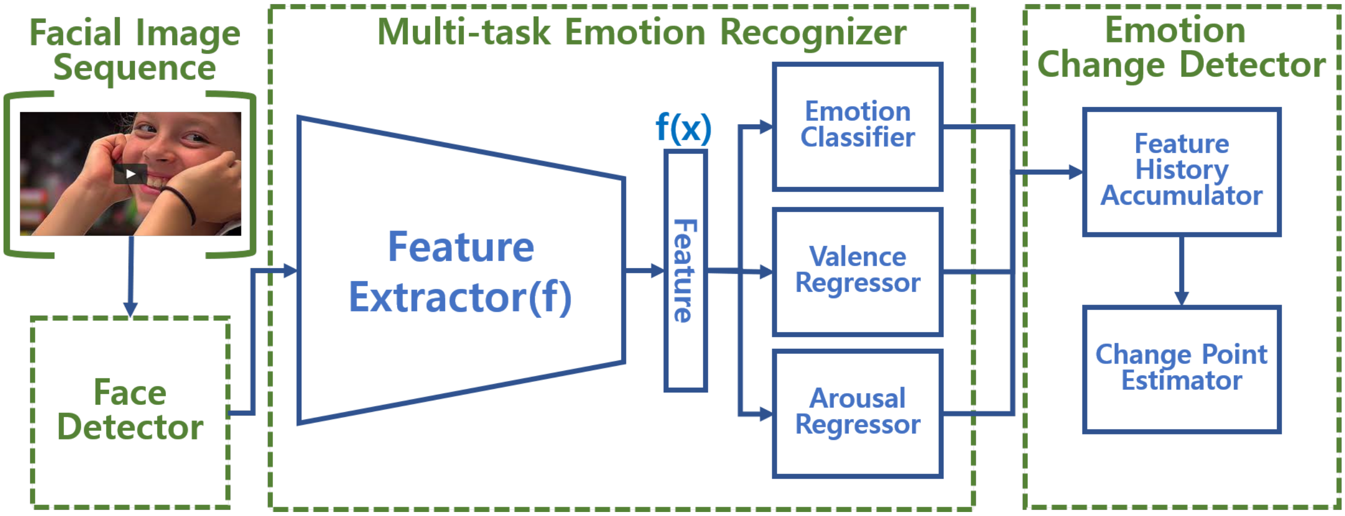

Fig. 2. The overall architecture of our framework.

the conversation in the HRI scenario to cause laughter.

Few researches have focused on automatic detection of emotion change. In [9], a detection system localizing change points is proposed based on speech signals using the Gaussian Mixture Model (GMM). However, to the best of our knowledge, our work is the first study to investigate detection of change points through facial expressions for HRI. In this workshop paper, we proposes an emotion change detector which consists of a deep multi-task learning (MTL) architecture for inferencing emotion status and a change point estimator for detecting changes of emotion status over time series.

The contributions of our work are follows: 1) We provide an initial investigation to automatic emotion change detection using facial expression information in the field of HRI. 2) We introduce a novel method for deep emotion change detection which combines a deep MTL architecture and a Pruned Exact Linear Time (PELT) algorithm. 3) We offer the baseline performance and an evaluation method for emotion change detection.

II. Proposed Method

Our method consists of three parts, a face detector, a multi-task emotion recognizer, and an emotion change detector, as shown in Fig. 2. A face detection is performed from a video input containing facial expression information. Then, based on detected regions, emotion status from the multi-task emotion recognizer is accumulated to an emotion state signal. Finally, the emotion state signal is analyzed to detect multiple change points in multivariate time series. The details are given below.

A. Multi-task Emotion Recognition

Psychologists have developed various models to explain the concept of emotional status. Among them, we infer the state of emotion through a categorical model[10] and a dimensional model[11], which are typically used in the field of facial expression recognition. The categorical model is a typical emotion model that can represent universal facial expressions based on Paul Ekman's emotion theory. The facial expressions are represented in seven discrete emotional categories, including Happiness, Sadness, Disgust, Fear, Anger, Surprise and Neutrality. On the other hand, the dimensional model was developed by Russell and represents emotional state in terms of valence and arousal dimensions rather than in discrete categories. Rather than having conflicting opinions, the two emotional models can well show some of the different aspects of complex emotions and play a complementary role. To reflect the complementary aspects of the two emotional representation tasks, we implemented them using a Deep MTL method.

Multi-task learning techniques have several advantages compared to single-task learning methods. First of all, they are memory efficient because of sharing layers inherently. In addition, since the results of several tasks can be obtained with one inference, they show the reduced time complexity. One of the most important advantages is that the associated tasks can perform complementary roles to achieve performance improvements. Taking advantage of these multi-task learning techniques, we perform emotion status prediction based on the two emotion models.

The details of our method are represented below. Given input facial images $\mathbf{x}$ and their labels $\mathbf{y}_e$, $\mathbf{y}_v$, and $\mathbf{y}_a$, we constitute a feature extractor $f(\mathbf{x};\theta_f)$ with a trainable parameters θ_f. Then, an emotion classifier is then defied as $e(f(\mathbf{x});\theta_e)$, consisting of the output of the feature extractor with a trainable parameter $\theta_\mathbf{e}$. Similarly, a valence regressor is represented as $v(f(\mathbf{x});\theta_v)$ and an arousal regressor is formulated as $a(f(\mathbf{x});\theta_a)$. The loss function of our MTL model, $MTL(\mathbf{x},\mathbf{y}_e,\mathbf{y}_v,\mathbf{y}_a;\theta_f,\theta_e,\theta_v,\theta_a)$, is defined as:

$$\begin{aligned} E(\mathbf{x},\mathbf{y}_e,\mathbf{y}_v,\mathbf{y}_a;\theta_f,\theta_e,\theta_v,\theta_a) =& CCE(e(f(\mathbf{x})),\mathbf{y}_e;\theta_f,\theta_e) \\ &+L_2(v(f(\mathbf{x})),\mathbf{y}_v;\theta_f,\theta_v) \quad (1) \\ &+L_2(a(f(\mathbf{x})),\mathbf{y}_a;\theta_f,\theta_a) \end{aligned}$$

where, CCE represents the loss function of Categorical Cross

Algorithm 1 Algorithm for PELT

```
Input: signal $(h_1, h_2, ..., h_n)$,
       cost function $E$,
       penalty value $\beta$.
Output: a set $CP$
  Initialization : $F(0) = -\beta$, $R \leftarrow \{0\}$, $CP \leftarrow \emptyset$.
  for $i = 1$ to $n$ do
      $\hat{i} = \text{argmin}_{j \in R}[E(j) + E(h_{j:i}) + \beta]$
      $F(i) \leftarrow [F(\hat{i}) + E(h_{\hat{i}:i}) + \beta]$
      $CP(i) \leftarrow CP(\hat{i}) \cup \{\hat{i}\}$
      $R \leftarrow \{j \in R : F(j) + E(h_{j:i}) \leq F(i)\} \cup \{i\}$
  end for
  return $CP$
```

Entropy for multi-class classification problem.

B. Change Detection

The results of the emotion status deduced from the multi-task emotion recognizer are represented as nine-dimensional vectors, which consist of seven-dimension for confidence values of categorical emotions, one dimension for valence values, and one dimension for arousal values. After the nine-dimensional emotion state is estimated, the emotion state is accumulated to a history signal over a time series. Based on the signal, our method detects multiple points of emotion changes using the PELT algorithm [12]. The PELT algorithm is a multiple change point search algorithm, which is known as exact and computationally efficient. Through the pruning step with in dynamic programming, it ensures that the computational complexity is linear to the length of the signal, $O(n)$ [13]. The PELT is described as in Algorithm 1.

III. Experimental Results

A. Experimental Design

We used the RetinaFace [14] algorithm for detecting facial regions from video input. Then, all detected facial images were resized to the 100×100. For the deep MTL model, we used ResNet [15] as our backbone network. To train the network, the stochastic gradient descent optimizer was used with the batch-size 512 and the number of train-epochs 120. The initial value of the learning rate was set to 0.1 and it was decayed by 10 every 30 epochs. For the PELT method, we used a kernalized cost function based on the radial basis function (rbf) [13]. The β for penalty value was set to 0.1

TABLE I

Baseline performance on AffectNet for the multi-task emotion recognizer.

Task	Backbone	Recg. Rate	MSE
Emotion Classification	ResNet-18	58.77%	-
Valence Regression	ResNet-18	-	0.093
Arousal Regression	ResNet-18	-	0.095

(Recg.: Recognition, MSE: Mean Squared Error)

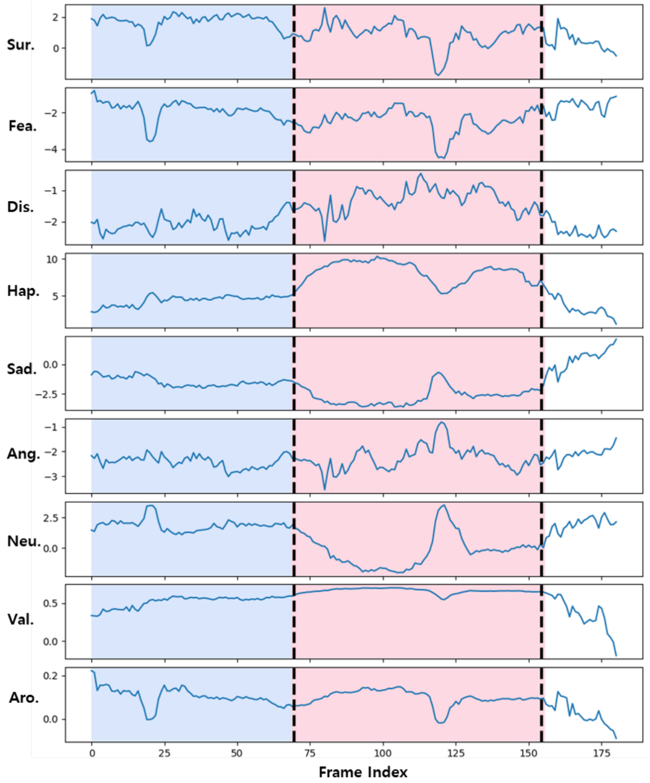

Fig. 3. A nine-dimensional emotion change signal. The vertical axis represents dimensions of the signal, which include seven-emotion categories (surprise, fear, disgust, happiness, sadness, anger, neutrality) and valence / arousal dimensions. Change points extracted our method are described as the dotted lines and the ground truth are lines that separates colors. Our method estimated two change points almost correctly.

B. Baseline Performance of Emotion State Recognition

We first report the baseline performance of our multi-task emotion recognizer, which infers emotion status through facial expression information. Our method was trained based on AffectNet [16], a representative in-the-wild dataset for facial expressions recognition. There are a total of $287,401$ facial images related to 7 emotion classes and valence/arousal values. Among them, we used $283,901$ images for a train set and $3,500$ images for a test set. As shown in Table I, we evaluated the three tasks related to emotion recognition, which include emotion classification, valence regression, and arousal regression. We achieved 58.77% for 7-class emotion classification, 0.093 mean squared error for valence regression, and 0.095 mean squared error for arousal regression. Using the results from the multi-task emotion recognition, we constructed emotion state vectors which are used as elements of an signal for emotion change detection.

C. Qualitative Evaluation of Emotion Change Detection

A signal of emotion change from facial expressions can show various forms depending on the intensity and direction of the change. Also, the variation factors in identity, pose, illumination, and occlusion of facial information can also affect the signal of emotional change. The change in emotion contains the various properties in facial information, so it is not easy to define the timing of the change. As baseline experiments for defining the timing of change, we collected video data containing the emotion change in Youtube, extracted it as a signal, estimated the timing of the emotional change, and conducted a qualitative assessment. As shown in Fig. 3, a nine-dimensional signal is extracted and two points of change were estimated based on the signal. The lines that separate colors represent the ground truth we annotated and the dotted lines are resulted points from our method. Since the emotion changes are relatively evident in the video we used, our method extracted two points almost correctly (Mean Absolute Error: 0.12 sec).

IV. CONCLUSIONS

We designed an emotion change detection framework which combines a multi-task emotion recognizer and a detector of multiple change points. A nine-dimensional emotion state vector was effectively extracted using a deep MTL model whose tasks are based on the categorical emotion model and the dimensional model in psychology. In addition, the emotion state vector is accumulated to an emotion state signal and used as a time series data for the detection of emotion changes. Multiple points of emotion changes was detected through the PELT algorithm. Our method is the initial study of emotion change detection using facial expression information and is expandable to a cognitive timing initiator of the robot system for HRI. As a future work, we plan to conduct quantitative experiments on the dataset we constructed with an online detection algorithm of multiple change points for applying to a real-time robot system.

REFERENCES

[1] A. Mehrabian, *Nonverbal communication*. Transaction Publishers, 1972.

[2] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, “Web-based database for facial expression analysis,” in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 5 pp.–.

[3] S.-C. Hsu, H.-H. Huang, and C.-L. Huang, “Facial expression recognition for human-robot interaction,” in *IEEE International Conference on Robotic Computing (IRC)*, 2017, pp. 1–7.

[4] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE Transactions on Affective Computing*, 2020.

[5] D. Kulic and E. A. Croft, “Affective state estimation for human–robot interaction,” *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[6] M. A. Nicolaou, H. Gunes, and M. Pantic, “Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space,” *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[7] Y. Yamashita, T. Fujimura, K. Katahira, M. Honda, M. Okada, and K. Okanoya, “Context sensitivity in the detection of changes in facial emotion,” *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.

[8] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, “Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements,” *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[9] Z. Huang, J. Epps, and E. Ambikairajah, “An investigation of emotion change detection from speech,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] P. Ekman and D. Keltner, “Universal facial expressions of emotion,” *Nonverbal communication: Where nature meets culture*, pp. 27–46, 1997.

[11] J. A. Russell, “A circumplex model of affect.” *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[12] R. Killick, P. Fearnhead, and I. A. Eckley, “Optimal detection of changepoints with a linear computational cost,” *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[13] C. Truong, L. Oudre, and N. Vayatis, “Selective review of offline change point detection methods,” *Signal Processing*, vol. 167, p. 107299, 2020.

[14] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 630–645.

[16] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.