Multimedia Tools and Applications (2019) 78: 3267–3276
https://doi.org/10.1007/s11042-018-6733-7

Personalized smart home audio system with automatic music selection based on emotion

Dongwann Kang[1] · Sanghyun Seo[2]

Received: 19 March 2018 / Revised: 23 August 2018 / Accepted: 27 September 2018 /
Published online: 29 October 2018

Abstract
In this paper, we introduce a personalized home audio system that uses IoT technologies to recommend and play music remotely based on a user's estimated emotion. This system estimates a user's emotion based on texts on their smartphone collected during outdoor activities. Based on this emotion, our system then searches for music that matches it from a music database. The system automatically detects the user when they return home, and plays the recommended music via a connected audio system. Consequently, personalized emotion-based music recommendation is provided transparently without the user's awareness.

Keywords Internet of things · Emotion estimation · Music recommendation

1 Introduction

Recently, Internet of Things (IoT) technologies have been rapidly developing due to a convergence of ubiquitous wireless communication, embedded systems techniques, and the popularization of using smartphones and wearable devices [29]. Modern smartphones are equipped with up to 10 built-in sensors, such as Global Positioning System (GPS) sensors, accelerometers, and gyroscopes, to enable them to capture different types of information for users, from their location to ambient light conditions. Moreover, as the use of wearable devices, such as smart watches and activity trackers, has been gradually increasing, sensors equipped in these devices broaden the range of applications for smartphones. These sensors construct a network, known as a body sensor network [27], and communicate with smart devices so that various statuses about a user and their environment can be collected and analyzed to provide a better user experience.

The development of IoT technologies enables personalized services in various fields. In the healthcare field, applications which monitor heart rate using a heart rate sensor in a smart

✉ Sanghyun Seo
shseo75@gmail.com

[1] Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, Korea

[2] Division of Media Software, Sungkyul University, Anyang, Republic of Korea

watch provide personalized healthcare information [9, 16]. In the fitness field, to suggest a personalized exercise plan, smartphone applications are commonly used in which an activity tracker communicates with running and cycling machines to obtain precise fitness data and track these on a smartphone. Also, for smart homes, various smartphone and wearable device applications have appeared, such as a smart door lock [3, 7] which enables the user to lock and unlock doors and a smart kitchen which allows the user to monitor gas leaks or the carbon monoxide state and control thermal and lighting conditions [13, 24].

In this study, we propose a home audio system, which remotely recommends and plays music based on a user's estimated emotion, as a type of personalized multimedia service using IoT technologies. In this proposed system, a smartphone application continually collects texts on the user's smartphone during outdoor activities. This application communicates with a server at the user's home by transmitting the texts. The server then estimates the user's emotion based on the texts and searches for music that matches the emotion from a music database. When the user returns home, the server automatically detects it and plays the recommended music via a connected audio system. As a result, personalized emotion-based music recommendation is provided.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of studies related to personalized music recommendation and emotion estimation. We then present the details of our personalized IoT technology-based multimedia system for recommending music based on a user's emotion in Section 3. The implementation and the results are given in Section 4. Finally, we conclude with a summary of our ideas and discuss future work in Section 5.

2 Related work

In the cognitive psychology field, many studies have conducted work to quantitatively measure human emotion. Russell [20] suggested a circumplex model for representing emotions as a two-dimensional space consisting of two axes: arousal and valence. Similarly, Thayer [25] also proposed a two-dimensional model in which the main factors are energy and tension. These models have been widely used in various emotion-related works, which have used facial expression and keywords to assess emotions. By expanding these approaches, Bigand et al. [4] showed that emotions, especially for music, can be represented by using a three-dimensional model through multi-dimensional scaling. Schimmack and Rainer [22], as well as Schimmack and Grob [21], also showed that a three-dimensional model, which consists of arousal, energy, and tension, cannot be reduced to a two-dimensional model.

Studies for recognizing emotion from music started with computational media aesthetics (CMA) [15], which was an approach to understand content based on its low-level features to be addressed for content management. Based on CMA, Feng et al. [6] proposed a method for retrieving music by emotion. They classified the emotion of a musical track by using relative tempo and silence ratio. Many studies have focused on predicting emotion from music by using parametric regression. Yang et al. [28] extracted 114-dimensional musical features by using spectral contrast and Daubechies Wavelet Coefficient Histograms. Then they employed principal component analysis to reduce the data space and adopted support vector regression with various boosting algorithms to estimate arousal and valence values. Similarly, Han et al. [8] proposed a method for music emotion classification. They collected high-dimensional temporal and spectral features. Then they applied nonnegative matrix factorization to reduce the dimensionality of the features and employed a support vector machine classifier to classify the music. By utilizing music emotion classification, they also proposed a music recommendation method based on a user's current and desired emotions. However, contrary to our study, they obtained a users' emotion by directly querying them. Lee et al. [14] extracted 376 acoustic features from music, and predicted the emotion of music by using linear regression. In their study, the emotion of a musical track was used to find a painting which matched it. To achieve this, the emotion of paintings was also estimated [11]. This concept, which matches different content using similar emotions, is employed in our study as well. Rho and Yeo [19] briefly summarized emotion models and acoustic features of multimedia which were used in previous multimedia emotion recognition studies.

Several studies have focused on the emotion of words. Bradley and Lang [5] provided a set of emotional ratings for 1,034 words. The words were rated in terms of pleasure, arousal, and dominance. Warriner et al. [26] extended this database to 13,915 words. Their database is directly employed in our study to predict a user's emotion based on texts on the user's smartphone. Seo and Kang [23] utilized emotional keyword-based image retrieval to acquire a training dataset of images. In their study, the word database produced by Warriner et al. [26] was also employed to determine the emotion of retrieved images.

Many studies have conducted work for recommending multimedia content including music. In these studies, a general predilection for certain music genres analyzed by social network [10], the user's situation [18] and user's sentiments extracted from social networks [12] were used to recommend multimedia content. Unlike these approaches, we estimate the emotion of the user and music, and recommend music by matching them.

3 Personalized music recommendation system based on emotion estimation

3.1 System overview

A system overview of this study is shown in Fig. 1. During outdoor activities, the emotion predictor, which is installed on a smartphone as an application, predicts a user's emotion based on texts on the smartphone and transmits them to the server at the user's home. The server collects the transmitted emotion data, and selects a user's representative emotion for

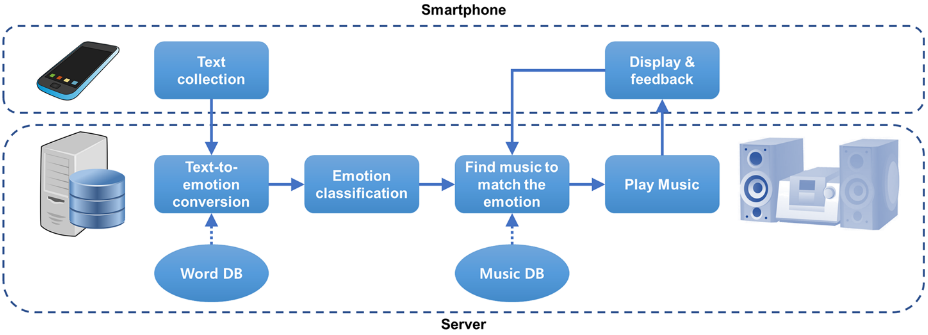


Fig. 1 System overview. The system consists of two parts: smartphone and server. The smartphone application collects texts on a user's smartphone and transmits them to the server. The server converts texts to emotions by selecting representative emotions using emotion classification and plays music that matches the user's emotion

the day by classifying this data. Once the user's return has been detected, the server selects music which matches the representative emotion from the user's music pool on the server and plays them through a connected audio system. The user can submit feedback on the results of the smartphone application while the music is played. The music recommendation performance is enhanced by assessing this feedback.

3.2 Text-based emotion estimation

This study aims to develop a system that works in the background to recommend music by estimating a user's emotion during everyday life. Therefore, an electroencephalogram-based emotion recognition approach, which is a traditional and reliable method for predicting human emotion, is not suitable for our system because it requires cumbersome equipment. Questionnaire-based emotion estimation is also inappropriate due to the hassle of frequently asking the user about their emotions. Instead of using cumbersome equipment or asking the user about their emotions, we propose a method for estimating a user's emotion based on texts with other people on the user's smartphone.

We first collected texts via push notifications from text message services on the user's smartphone, and picked only texts with words that existed in our word database. This database, which is employed in [26], consists of pairs of English words whose emotion has been estimated through a crowdsourced user study. To represent emotion quantitatively, the emotion of a word is defined as a two-dimensional coordinate, where x and y coordinates correspond to arousal and valence, respectively, using Russell's model [20], as shown in Fig. 2. Consequently, the database consists of words and their two-dimensional coordinates.

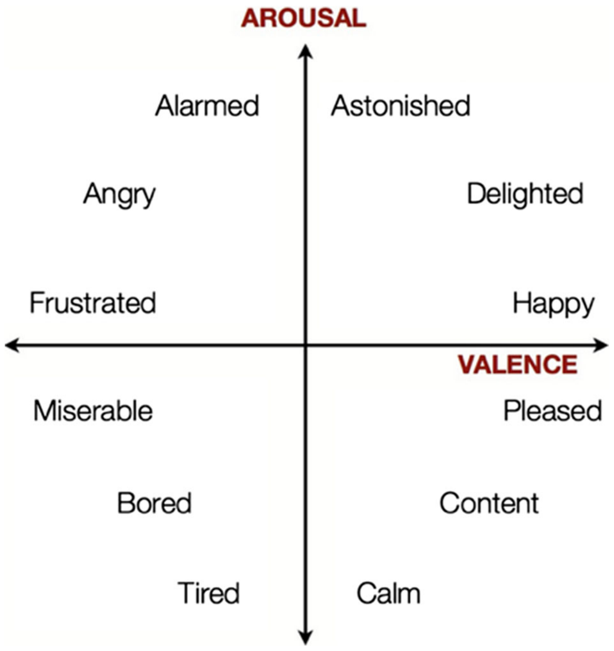


Fig. 2 Emotional words represented by a two-dimensional model consisting of arousal and valence axes

3.3 Representative emotion estimation

We collected texts from push notifications on a user's smartphone, and mapped them on to Russell's two-dimensional emotion space. Then, a number of texts are placed on this space, as shown in Fig. 3. Among them, representative emotions should be selected to be used for recommending music. To achieve this we employed a mean shift, which is a density-based clustering algorithm. On the two-dimensional emotion space where the texts are located, we calculated the mean shift vector of a text i, V_i, which indicates the direction towards the local maxima of density, by using (1).

$$V_i = 1/C(j) \sum p_j - p_i \quad j \in N(i) \tag{1}$$

where $N(i)$, $C(j)$, and p_i denote the neighbors of i, the number of j, and the coordinate of i on the emotion space, respectively. Once we had calculated the mean shift vector of a text, we shifted the coordinate of this text along the vector. We then iteratively obtained and shifted mean shift vectors until the updated coordinates converged. For all texts we found their converged coordinates. We regarded these coordinates as the emotional modes of the texts, and defined them as the representative emotions of the texts. Due to the non-parametric characteristics of the mean shift algorithm, this method works appropriately on multimodal distributions. However, some modes have relatively low densities compared to other modes even though they correspond to local maxima. To avoid selecting these modes as representative emotions, we considered the influence of mode m_i by using (2).

$$f(m_i) = C(j) * C(n)/T^2 \quad j \in N(m_i), n \in S(m_i) \tag{2}$$

where $C(x)$ denotes the number of x again, and T and $S(m)$ are the number of all texts and a set consisting of the elements which converge into mode m, respectively. We excluded the modes of which the $f()$ value is lower than a pre-assigned threshold for representative emotions.

3.4 Emotion-based music recommendation

Once a user's emotion has been estimated based on texts, our system recommends music that matches this emotion. To match music to a given emotion, we found the emotion of musical tracks and selected music which has a similar emotion. To find the emotion of musical tracks, we employed two approaches, as follows.

3.4.1 Expert-based approach

We employed a music database, which was generated by classifying music according to emotions by experts. In this database, emotions are represented by emotional words so that they can have a two-dimensional emotion coordinate and be utilized in our system. We found the coordinates of words from the word database used in Section 3.2 and tagged music with these coordinates.

3.4.2 Feature-based approach

We provided music recommendations for music that was not included in the database created by experts. To achieve this, we predicted the emotion of a musical track by analyzing the relationship between the features of music and emotion. We extracted acoustic features, such as dynamics, timbre, harmony, register, rhythm, and articulation, from music by using

Lee et al.'s method [14] and established a prediction model by using linear regression. To find the emotion of a musical track, we obtained the emotional responses of volunteers by asking them for their arousal and valence in a range from 1 to 9. Then we assigned the music its average values and employed this as emotion coordinates.

Our system regards that the music matches a certain emotion as long as it is within a pre-defined distance from the emotion, represented by emotion coordinates. In the case where there is no music within this distance, we recommend the music which is nearest to the emotion. For user convenience, our system provides several options for recommendations: to recommend 1) music which matches all representative emotions, 2) music which matches most representative emotions for which the value of Equation 2 is high, and 3) music which is within a pre-defined distance from texts that belong to representative emotions. In our system users can select an option according to their preferences.

Sometimes, users might want to listen to their favorite music regardless of their current emotion. To reflect a user's taste for certain music, our system receives user feedback by using two options: play more or play less. Initially, our system assigns each musical track a priority value of zero. If the user selects "play more" as feedback the priority is increased in increments of one up to 9. For "play less" the priority is decreased in the same manner down to -9. We utilized the priority value when calculating the distance between a musical track and a representative emotion by using (3).

$$D(p_m, p_e) = k^r \times |p_m - p_e| \quad (k < 1) \tag{3}$$

where r is the priority value (integer), and p_m and p_e denote coordinates of music and a user's representative emotion, respectively. We used 0.9 as a value k which controls the influence of the priority value. If a user prefers to play a musical track more or less, the priority value affects the distance function so that the music is selected more or less frequently.

4 Results

Our smartphone application (Fig. 3) was developed in the Android environment [1]. The application stored the GPS coordinates of the user's home and collected texts from the notifications of the user's text message services once their current coordinates were more than a pre-defined distance from their home. Then it transmitted the texts to the server at the user's home. This procedure was continuously performed until the user returned home.

The server (Fig. 4) located at home was set up using a Raspberry Pi 3 [17] and Apache Tomcat [2]. The server received texts from the smartphone application through HTTP-based communication. The server application was developed using Python, and it converted texts into emotion coordinates, which matched the words in the database described in Section 3.2, and these were stored accumulatively during the user's outdoor activities. When the user's home-coming was detected, by using GPS coordinates, the smartphone application notified the server. Finally, the server automatically played recommended music via an external audio system which was connected through a 3.5 mm audio jack from the Raspberry Pi and an AUX input jack to the audio system. The smartphone application displayed information for the music which was played and received the user's feedback. This feedback was transmitted to the server to update the priority value of this music and utilized in further music recommendations.

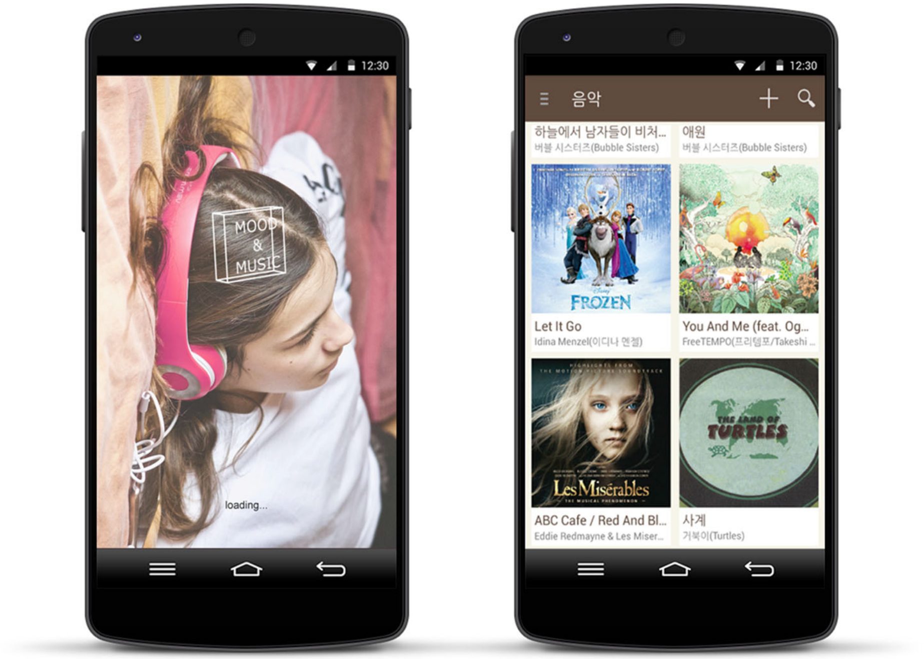

12:30
MOOD & MUSIC
loading...
12:30
음악
하늘에서 남자들이 비처...
버블 시스터즈(Bubble Sisters)
애원
버블 시스터즈(Bubble Sisters)
FROZEN
Let It Go
Idina Menzel(이디나 멘젤)
You And Me (feat. Og...
FreeTEMPO(프리템포/Takeshi ...
Les Misérables
ABC Cafe / Red And Bl...
Eddie Redmayne & Les Miser...
THE LAND OF TURTLES
사계
거북이(Turtles)

Fig. 3 The smartphone application which transmits texts collected from notifications and displays information on played music

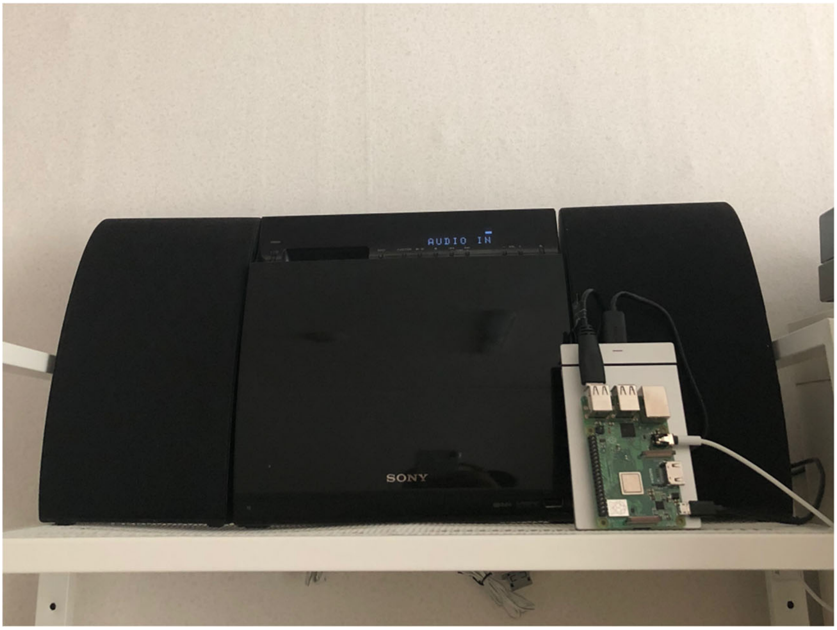


Fig. 4 Server set up using a Raspberry Pi 3 and an external audio system

5 Conclusions

In this paper, we introduced a system for recommending and automatically playing music which matched a user's emotion by using emotion estimation based on texts from the user's smartphone. To achieve this, we first obtained texts from the notifications of the text message services on the user's smartphone and converted them into two-dimensional emotion coordinates, which consisted of arousal and valence values, by using a pre-studied word database. Then, to find a user's representative emotions from these words, we performed mean shift clustering on the emotion coordinates, and selected the modes that had high influence factors as the representative emotions. We determined the emotions of the music by using expert-based and feature-based approaches. This allowed our system to select music whose emotion coordinates were close to the representative emotion and recommend them. In this system, users could give feedback as to whether they wanted to play the recommended music more or less according to their preferences, and this feedback was reflected in further music recommendations.

In future work we plan to develop more accurate emotion estimation methods. Photos taken on smartphones might enhance accuracy through facial expression and the mood of the color. Also, smart watches equipped with built-in heart rate sensors might be able to give information on when users are more aroused. We believe that these additional elements will enhance emotion estimation accuracy.

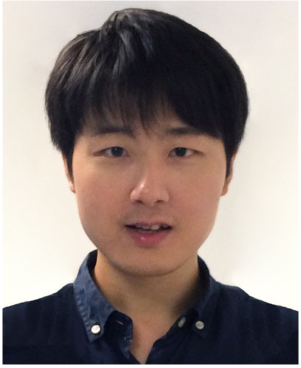
Acknowledgements This work has supported by Seoul National University of Science & Technology and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2016R1D1A1B03935378).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Android api guide. http://developer.android.com/guide/index.html
2. Apache tomcat ®. http://tomcat.apache.org
3. August home. http://august.com
4. Bigand E, Vieillard S, Madurell F, Marozeau J, Dacquet A (2005) Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cogn Emotion 19(8):1113–1139. https://doi.org/10.1080/02699930500204250
5. Bradley MM, Lang PJ (1999) Affective norms for english words (anew): instruction manual and affective ratings
6. Feng Y, Zhuang Y, Pan Y (2003) Music information retrieval by detecting mood via computational media aesthetics. In: Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003), pp 235–241. https://doi.org/10.1109/WI.2003.1241199
7. Fernandes E, Jung J, Prakash A (2016) Security analysis of emerging smart home applications. In: 2016 IEEE Symposium on security and privacy (SP), pp 636–654. https://doi.org/10.1109/SP.2016.44
8. Han BJ, Rho S, Jun S, Hwang E (2010) Music emotion classification and context-based music recommendation. Multimed Tool Appl 47(3):433–460. https://doi.org/10.1007/s11042-009-0332-6
9. Hello heart. http://helloheart.com
10. Jun S, Kim D, Jeon M, Rho S, Hwang E (2015) Social mix: automatic music recommendation and mixing scheme based on social network analysis. J Supercomput 71(6):1933–1954. https://doi.org/10.1007/s11227-014-1182-1
11. Kang D, Shim H, Yoon K (2018) A method for extracting emotion using colors comprise the painting image. Multimed Tool Appl 77(4):4985–5002. https://doi.org/10.1007/s11042-017-4667-0
12. Kim M, Park SO (2013) Group affinity based social trust model for an intelligent movie recommender system. Multimed Tool Appl 64(2):505–516. https://doi.org/10.1007/s11042-011-0897-8

13. Kummer M (2017) Review of the ecobee4 smart thermostat with homekit
14. Lee T, Lim H, Kim DW, Hwang S, Yoon K (2016) System for matching paintings with music based on emotions. In: SIGGRAPH ASIA 2016 Technical briefs, SA '16. ACM, New York, pp 31:1–31:4. https://doi.org/10.1145/3005358.3005366
15. Nack F, Dorai C, Venkatesh S (2001) Computational media aesthetics: finding meaning beautiful. IEEE MultiMedia 8(4):10–12. https://doi.org/10.1109/93.959093
16. Phan D, Siong LY, Pathirana PN, Seneviratne A (2015) Smartwatch: Performance evaluation for long-term heart rate monitoring. In: 2015 International symposium on bioelectronics and bioinformatics (ISBB), pp 144–147. https://doi.org/10.1109/ISBB.2015.7344944
17. Raspberry pi 3 model b. https://www.raspberrypi.org/products/raspberry-pi-3-model-b/
18. Rho S, Song S, Nam Y, Hwang E, Kim M (2013) Implementing situation-aware and user-adaptive music recommendation service in semantic web and real-time multimedia computing environment. Multimed Tool Appl 65(2):259–282. https://doi.org/10.1007/s11042-011-0803-4
19. Rho S, Yeo SS (2013) Bridging the semantic gap in multimedia emotion/mood recognition for ubiquitous computing environment. J Supercomputing 65(1):274–286. https://doi.org/10.1007/s11227-010-0447-6
20. Russell JA (1980) A circumplex model of affect. J Person Soc Psychol 39(6):1161
21. Schimmack U, Grob A. (2000) Dimensional models of core affect: a quantitative comparison by means of structural equation modeling. European J Personal 14(4):325–345. https://doi.org/10.1002/1099-0984(200007/08)14:4<325::AID-PER380>3.0.CO;2-I
22. Schimmack U, Rainer R (2002) Experiencing activation: energetic arousal and tense arousal are not mixtures of valence and activation. Emotion 2(4):412
23. Seo S, Kang D (2016) Study on predicting sentiment from images using categorical and sentimental keyword-based image retrieval. J Supercomput 72(9):3478–3488. https://doi.org/10.1007/s11227-015-1510-0
24. Stojkoska BLR, Trivodaliev KV (2017) A review of internet of things for smart home: Challenges and solutions. J Cleaner Product 140:1454–1464. https://doi.org/10.1016/j.jclepro.2016.10.006
25. Thayer RE (1990) The biopsychology of mood and arousal. Oxford University Press, Oxford
26. Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 english lemmas. Behav Res Methods 45(4):1191–1207. https://doi.org/10.3758/s13428-012-0314-x
27. Yang GZ, Yang G (2006) Body sensor networks, vol 1. Springer, Berlin
28. Yang YH, Lin YC, Su YF, Chen HH (2008) A regression approach to music emotion recognition. IEEE Trans Audio, Speech, Language Process 16(2):448–457. https://doi.org/10.1109/TASL.2007.911513
29. Zhang D, Ning H, Xu KS, Lin F, Yang LT (2012) Internet of things. J UCS 18:1069–1071



Dongwann Kang is an Assistant Professor in the Department of Computer Science and Engineering at Seoul National University of Science and Technology. He received his PhD from Chung-Ang University in Korea in 2013, where he has been a research fellow until Jun 2015. He was a lecturer of Undergraduate Interdisciplinary Program in Computational Sciences, Seoul National University, Korea (from Mar 2014 to Jun 2015); a lecturer at the Department of Multimedia, Sookmyung Women¡¯s University, Korea (from Mar 2014 to Dec 2014); a visiting researcher (from Jul 2015 to Jan 2018) and a Marie Sk©©odowska-Curie fellow (from Feb 2018 to Aug 2018) at the Faculty of Science and Technology, Bournemouth University, UK. His research interests include non-photorealistic rendering and animation, emotional computing, image manipulation and GPU processing.

Sanghyun Seo received his B.S. degrees in Computer Science and Engineering from Chung-Ang University, Seoul, Korea, in 1998 and M.S. and Ph.D. degrees in GSAIM Dep at Chung-Ang University, Seoul, Korea, in 2000 and 2010. He was senior researcher at G-Inno System from 2002 to 2005. He was the postdoctoral researcher at Chung-Ang University, in 2010, and the postdoctoral researcher at LIRIS Lab, Lyon 1 University from February 2011 to February 2013. He had worked at the ETRI (Electronics and Telecommunications Research Institute), DaeJeon, Korea, May 2013 to February 2016. Now, he is currently a faculty of Department of Media Software at Sungkyul University He has been a reviewer in Multimedia Tools and Applications (Springer), Computer and Graphics UK (Elsevier), Journal of Supercomputing (Springer), Visual Computer (Springer), and Program Committee member in many international conferences and workshops and has edited a number of international journal special issues as a guest editor, such as Journal of Real-Time Image Processing, Journal of Internet Technology, and Multimedia Tools and Applications and so on. He has been appointed as an Associated-Editor Journal of Real-Time Image Processing since 2017. His research interests are in the area of computer graphics, non-photorealistic rendering and animation, 3D GIS system, real-time rendering using GPU, VR/AR, and game technology.