Multimedia Recommendation System Using Adaptive Resonance Theory Neural Model for Digital Storytelling

Ju-Youn Park, Woo-Ri Ko and Jong-Hwan Kim
School of Electrical Engineering
KAIST
Daejeon 305–701, Republic of Korea
Email: {jypark, wrko, johkim}@rit.kaist.ac.kr

***Abstract*—Multimedia recommendation technology has been developed in various fields these days. In order to provide multimedia in addition to dialog, it is essential to select appropriate multimedia associated with a certain situation for more delivery effects of digital storytelling, which enables story telling agents to share their stories with users using digital multimedia in an effective way. For this purpose, we propose a multimedia recommendation system for software agents of smart devices to select multimedia that is appropriate to the given situation in storytelling to users or interacting with users. The fusion ART network is employed for the multimedia recommendation system that selects an appropriate digital media file for individual multimedia features. The system is learned incrementally based on feedback from users. The proposed system is purposed to select multimedia to be conveyed in addition to the dialog between the user and the digital creature on a smartphone. The applicability is verified through experiments with a smartphone application implemented for demonstration.**

I. Introduction

Multimedia recommendation technology has been developed in various fields these days. Smartphone applications for recommending music or movie according to genres that users prefer has been developed. For the advanced application of the technology, we propose a multimedia recommendation system for digital storytelling of the digital creature on a smartphone.

Digital storytelling enables story tellers to share their stories with listeners in a powerful way by using digital multimedia [1], [2]. With the digital tools such as narratives and music, story tellers can convey their stories more effectively. We propose to employ digital storytelling to the digital creature that is developed for the project, Technology Development of Virtual Creatures with Digital Emotional DNA. The purpose of the project is to develop a digital human that is a software agent providing emotional services based on the digital Deoxyribonucleic Acid (DNA) extracted from the personal information including application usage, location and music preference, collected from a user's smartphone. The digital creature, based on the digital DNA that expresses the characteristic of voice, appearance, action and personality extracted from users, is developed to have a conversation with the user using interactively generated sentences. If the digital creature can communicate with the user using additional multimedia as well as the generated sentences, the user can understand what the digital creature has intended more effectively. In order to provide multimedia in addition to dialog, it is essential for the digital creature to select appropriate multimedia associated with a certain situation for more delivery effects of digital storytelling. Hence, we develop a multimedia recommendation system to select an appropriate digital file for individual multimedia features such as images and audio for a given situation. Fig. 1 shows a conceptual diagram of the project, and the application of the proposed multimedia recommendation system for digital storytelling.

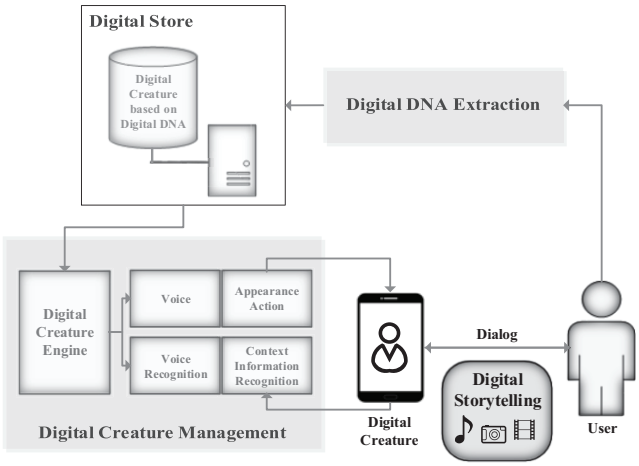


Fig. 1: The conceptual diagram of the project, Technology Development of Virtual Creatures with Digital Emotional DNA of Users. The dialog from the digital creature has more delivery effects with the proposed multimedia recommendation system for digital storytelling.

There have been several researches on recommendation system [3], [4]. There exist various approaches for recommendation system depending on the factor to be considered. Collaborative Filtering (CF) recommender systems recommend the items that people preferred before [5]. Content-based recommender systems recommend the items that have similar

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

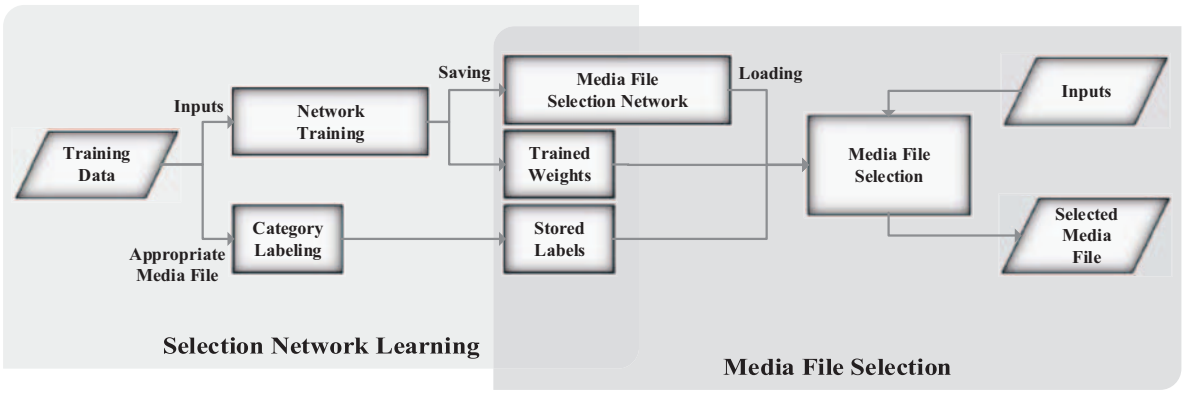


Fig. 2: Overview of the media file selection algorithm including the selection network learning process and the media file selection process.

content to the items that users preferred [6]. Hybrid recommender systems combine several recommendation methods [7].

In this paper, we develop a kind of hybrid recommendation system. As the method to select and recommend digital multimedia files, we employ a neural network, Fusion Adaptive Resonance Theory (Fusion ART) [8], [9], which is usually applied to pattern recognition [10] or memory system [11]-[13]. The fusion ART network is one of the powerful methods to classify inputs into classified categories. The proposed multimedia recommendation system selects and recommends a digital media file for individual multimedia features using the fusion ART network. For initial training of the system, the training data that contains pairs of preferences of users for multimedia selection in various situations is provided by the designer beforehand. By accumulative learning based on feedback from users, the system recommends similar multimedia that the users prefer. As the criteria for selecting a digital media file for individual multimedia features to be recommended, our system employs the keywords of the generated sentences from the digital creature and the context information, which is defined as any information representing the situation such as a person, place or object [14], inferred from the sensor data collected from a user's smartphone. A fusion ART network is trained for the training set that contains pairs of a specific media file and a given situation that is represented with the keywords of dialog and context information of users for each type of multimedia such as images and audio. We employ two fusion ART networks respectively for images and audio in this paper. Based on the trained networks for all multimedia features used, the proposed system recommends multimedia for a certain situation.

The significant feature of the fusion ART network is that the number of category nodes representing the classified categories is increased with coming inputs of a new category without retraining the network. Using the feature, our system gets improved based on feedback from users. If a user is not satisfied with the recommended multimedia and wants to change it, the user can inform a more preferred media file for a certain multimedia feature to the multimedia recommendation system as feedback. With the feedback, the fusion ART network for the corresponding media type is trained accumulatively, and becomes a more personal model.

When recommending multimedia with different types of media, it is important to decide the types of media to be provided at the same time. Since some media files should not be provided because of the characteristics of media, for example if an audio file and a video file are played simultaneously, the sound will be overlapped, the media type selection process is performed to decide certain types of media at the same time. In our system, a single network is used to select a media file for each multimedia feature. The selection process decides which media type files to be provided based on the characteristics of media.

The rest of this paper is organized as follows. Sections II and III present the media file selecting algorithm and the proposed multimedia recommendation system, respectively. Section IV describes the experimental results. Finally, concluding remarks follow in Section V.

II. Media File Selection

The proposed multimedia recommendation system selects and provides a digital media file using the fusion ART network for individual multimedia features such as images and audio as shown in Fig. 2. For each type of multimedia, the ART network is trained for a training set which contains pairs of a specific media file and given inputs such as the keywords of conversations and context information of the user. Based on encoded inputs of keywords and context information, the trained network selects a media file for individual types of multimedia for a certain situation.

The structure of the fusion ART network used for the multimedia recommendation system is shown in Fig. 3. It consists of an input field F_1 and category field F_2. The F_1 field receives the keywords of conversations and context information of the user as inputs. In this paper, the keywords

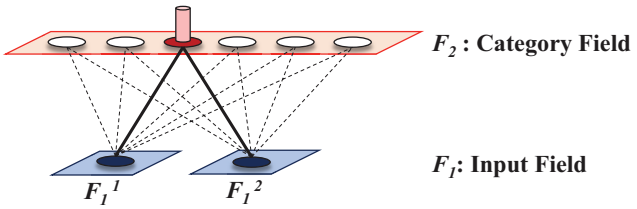


Fig. 3: The fusion ART network structure used to select a media file for individual multimedia features for the proposed multimedia recommendation system.

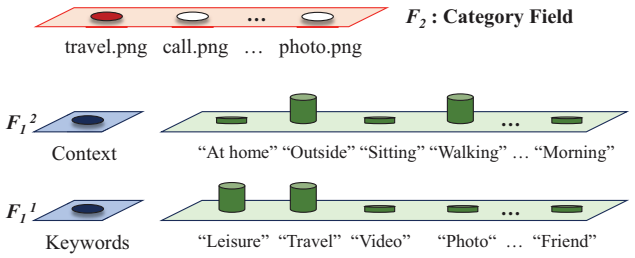


Fig. 4: The inputs encoded into two input channels and the category field of the fusion ART network for the case of recommending the image related to travel.

and context information are encoded as binary vectors into two input channels F_1^1 and F_1^2, and the selected media file is represented as a node in the F_2 field as shown in Fig. 4, respectively.

A. Fusion ART

A fusion ART network is employed for individual multimedia features to select a media file. The fusion ART network is trained with a set of universal computational processes for the encoding, recognition, and reproduction of patterns, and encodes the pattern of association between given inputs and a media file as the weighted connections between the inputs from multiple channels and the corresponding category. Each category is characterized by an activity value. The process to match and activate each category node from the input vector is described in the following five stages.

1) Complement coding: The vector received for each channel of the input layer, $\mathbf{X}^k = (x_1^k, x_2^k, \cdots, x_{2n}^k) = (\mathbf{I}^k\ \bar{\mathbf{I}}^k)$ is the concatenated form of the input vector, $\mathbf{I}^k = (I_1^k, I_2^k, \cdots, I_n^k)$ such that $I_i^k \in [0,1]$ for $k = 1, 2, \cdots, n$ and its complement vector $\bar{\mathbf{I}}^k = (\bar{I_1}^k, \bar{I_2}^k, \cdots, \bar{I_n}^k)$ such that $\bar{I_i}^k = 1 - I_i^k$.

2) Code activation: The activity value of the jth output node associated with the received vector $\mathbf{X}^k$ is determined as follows:

$$T_j = \sum_{k=1}^{n} \gamma^k \frac{|\mathbf{X}^k \wedge \mathbf{W}_j^k|}{(\alpha^k + |\mathbf{W}_j^k|)} \tag{1}$$

where $\alpha^k \geq 0$ is a choice parameter, $\gamma^k \in [0,1]$ is a contribution parameter, $\mathbf{W}_j^k$ is a weight vector, the fuzzy AND operator $\wedge$ is defined as $(\mathbf{A} \wedge \mathbf{B})_i \equiv min(a_i, b_i)$, and the norm $|\cdot|$ is defined as $|\mathbf{A}| = \sum_i a_i$.

3) Code competition: The Jth node of the largest activity value among all activity values derived during the stage of code activation is selected as follows:

$$T_J = max\left\{T_j : \text{for all } F_2^{fusion} \text{ node } j\right\} \tag{2}$$

As a node is selected as a category, an output value of 1 is set for the chosen node. The other nodes have output values of 0.

4) Template matching: Each selected Jth node is checked according to its resonance value. If the resonance value is larger than a vigilance parameter $\rho^k \in [0,1]$, the Jth node is selected finally. If not, a new node is committed to be activated. Template matching is performed as follows:

$$m_J^k = \frac{|\mathbf{X}^k \wedge \mathbf{W}_J^k|}{|\mathbf{X}^k|} \geq \rho^k \tag{3}$$

5) Template learning: If template matching is accomplished, the weight vector is then updated as follows:

$$\mathbf{W}_J^k = (1 - \beta^k)\mathbf{W}_J^{k(old)} + \beta^k(\mathbf{X}^k \wedge \mathbf{W}_J^{k(old)}) \tag{4}$$

where $\beta^k \in [0,1]$ is the learning rate.

B. Selection Network Learning

The fusion ART network employed for selecting a media file can be incrementally trained for continuously given inputs. When the pairs of inputs representing a certain situation such as the keywords of conversations and context information of user and an appropriate media file are given as the training set, the fusion ART network is trained as the weights connecting the input channels and the selected category node are updated.

In this paper, the keywords of conversations and context information of the user are used as inputs. The number of bits in the binary vector representing the context information is constant as 18, but the number of keywords of conversations is not constant. So we provide the keyword category containing all the keywords given as inputs to be used for generating the binary bit vector of the keywords.

The process of training the fusion ART network shown in the selection network learning part in Fig. 2 is described in the following. As each pair in the training set is encoded, a category node is activated by the five stages of processes of fusion ART. The trained weights are stored as the trained network. The category nodes in the fusion ART network are labeled as the corresponding media file name, and the file name labels are also stored. For individual types of multimedia, the fusion ART network is trained individually.

C. Media File Selection

Before selecting a media file, the corresponding trained network of the type of media is loaded. The stored weights from the selection network learning process are applied to the initialized fusion ART network, and the category nodes in the category field are created and labeled as the stored media

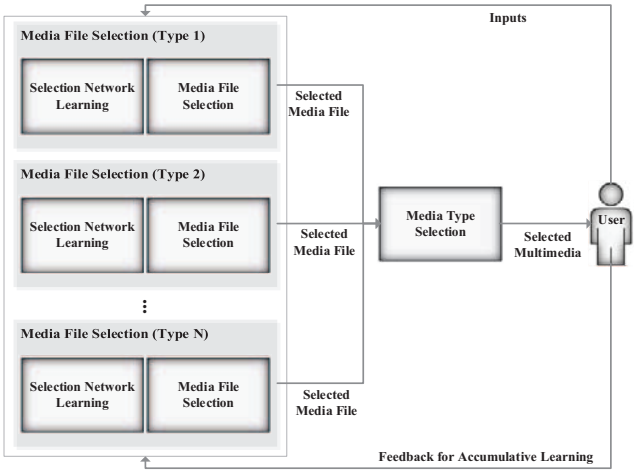


Fig. 5: Overview of the proposed multimedia recommendation system.

Algorithm 1 Feedback-based Accumulative Learning

1: **given** inputs representing a situation and a preferred media file as feedback
2: select a node J in F_2 that has the largest activity value among all activity values based on input vectors from F_1 through the code competition process in fusion ART with $T_J = max\left\{T_j : \text{for all } F_2^{fusion} \text{ node } j\right\}$
3: let the activity value of the selected node as y_J
4: **if** y_J is larger than a vigilance parameter ρ^k **then**
5: label the selected activity node J as the preferred media file as feedback
6: **else**
7: create a new category node and select as J
8: label the new activity node J as the preferred media file as feedback
9: **end if**
10: perform the template learning process with $\mathbf{W}_J^k = (1-\beta^k)\mathbf{W}_J^{k(old)} + \beta^k(\mathbf{X}^k \wedge \mathbf{W}_J^{k(old)})$

file names. From the loaded network, a digital media file is selected based on the inputs such as the keywords and context information.

When the inputs are encoded as binary vectors, the corresponding category node is activated by the template matching process in fusion ART. The category node which has the largest activity value among those that exceed the vigilance value is selected. Then the media file among all the stored media file names that is labeled to the activated node is selected. The process is shown in the media file selection part in Fig. 2.

III. Multimedia Recommendation System

The proposed multimedia recommendation system is described in Fig. 5. When inputs representing a certain situation of users such as the keywords of conversations and context information are given, a media file is selected through the fusion ART network by the media file selection process for individual multimedia features. Among all the selected media files, the media type selection process selects certain types of media and provides the corresponding media files to the user with the dialog by the digital creature on a smartphone. If the user is not satisfied with the selected media file for a certain type of media, the user may inform a preferred media file as feedback for network learning.

A. Media Type Selection

Before recommending the selected media files to users, the process of selecting the types of media to be conveyed should be performed. Some types of media should not be provided at the same time because of the characteristics of media, for example audio and video are overlapped each other by the sound. Thus, if there exist overlapped media files, the activity values of selected category nodes are compared and the media file that has the largest activity value is chosen to be recommended finally through the media type selection process.

B. Feedback-based Accumulative Learning

Our multimedia recommendation system proceeds accumulative learning based on the feedback from users as summarized in Algorithm 1. When the selected media file is provided to the user, the user can inform preferred media file as feedback if the user is not satisfied with it.

For the case that the associated inputs have already made resonance with the existing category node, the label of the category node changes into the preferred media file provided as the feedback. On the other hand, if the associated inputs are not resonant with the existing category nodes, a new category node is created and the node is labeled as the preferred media file from the feedback. Based on the feedback from users, the fusion ART network to select a media file for a certain type of media is trained incrementally. In other words, the accumulative learning process is performed based on the feedback from users.

IV. Experiments

A. Experimental Setup

In the experiments, we used the keywords of conversations and context information of the user as the inputs in the input field. The multimedia recommendation system is purposed to be applied for the digital creature on a smartphone developed for the project, Technology Development of Virtual Creatures with Digital Emotional DNA of Users as shown in Fig. 6.

The digital creature can talk with the user using interactively generated sentences according to the current dialog. In order to convey the meaning of sentences in a more understandable way, the multimedia related to the dialog is helpful. Thus, we used the keywords of the generated sentences as an input. A binary vector which has the same number of bit as the number of all the used keywords was used to represent the keywords as the bits associated with the given keywords were set to 1,

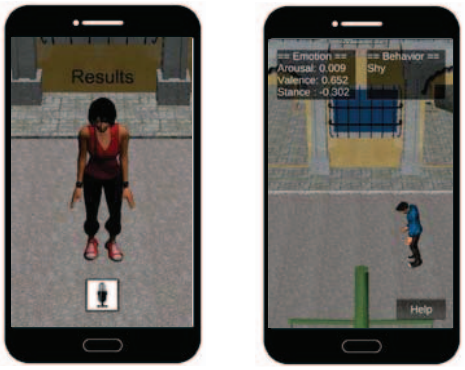


Fig. 6: Examples of the digital creature developed for the project, Technology Development of Virtual Creatures with Digital Emotional DNA of Users.

and the other bits were set to 0. In this experiment, we used 36 keywords.

In addition to the keywords of conversations with the digital creature, the context information of the user was also used as an input. We defined 18 features categorized into 4 different classes to describe the context information as shown in Table I. The context information is inferred based on the data collected from the sensors on a smartphone using the algorithm, Quantum-inspired Evolutionary Algorithm-based classifier algorithm [15]. The context information was represented as a binary vector with 18 bits as the bits associated with the given context information were set to 1, and the other bits were set to 0.

The proposed multimedia recommendation system was implemented on the android smartphone application for testing. A certain situation was provided with the keywords of conversations and context information of the user as the corresponding words separated by a comma. For the given inputs, the file names of image and audio were selected respectively from the two trained fusion ART networks. Based on the selected file names, the recommended image and audio were shown and played, respectively. Since images and audio do not interrupt each other, both media files were always provided at the same time.

Two fusion ART networks were built and trained on the android application for image and music recommendations. The training set was provided as a text file containing the pairs of the keywords, context information and corresponding media file. In this paper, we trained both networks for 15 different situations as listed in Table II. The trained networks were stored as text files containing the values of weights connecting the F_1 and F_2 fields in each network. Based on the selected media files, the corresponding image and audio were shown and played on the android application, respectively.

TABLE I: A list of the context information categorized into four classes.

Class	Context
Place	Inside, Bus, At home, Outside, Restaurant
User state	Lying down, Sitting, Standing, Walking, Running, Stair
Smarthone state	On desk, In hand, In pocket
Time	Morning, Noon, Evening, Night

TABLE II: A list of the keywords and context information along with the corresponding media file name.

Keyword	Context	Media file
Leisure, Stress, Travel	Outside	Travel
Knowledge, Search, Correct	Inside	Search
Video, Youtube	Bus	Video
Music, Sound	Sitting	Music
Photo, Camera	Standing	Photo
Human, Someone	Walking	Human
Sleep, Night, Evening	At home, Lying down, Evening, Night	Sleep
Morning, Noon, Alarm	Morning, Noon	Morning
Rain, Umbrella	Running	Rain
Spring, Bird, Romance	Stair	Spring
Web, Comments	On desk	Web
Conversation, Talk, Call	In hand	Call
Work, Company, Job	In pocket	Work
Weather	Morning	Weather
Restaurant, Eat	Restaurant	Eat

B. Experimental Results

For the demonstration of our multimedia recommendation system on the android smartphone application, we set 10 cases of different situations with keywords and context information, and provided the results as shown image and played audio to 10 users. The snapshots of the android smartphone application for multimedia recommendation are shown in Fig. 7.

1) User satisfaction: In this experiment, each user was requested to give a score in 7-point Likert scale about the appropriateness of the selected multimedia for each case of situation. And, if the user wanted, the user informed a preferred media file for each multimedia feature as feedback. After learning the two fusion ART networks incrementally based on the feedback, the user was led to give a score again.

Before users gave feedback to the system, the average score of appropriateness was 4.99 in 7-point Likert scale. After feedback was given, the average score was 5.65. The average score of 0.66 increased through the network adaption process based on feedback. The increase in the score of appropriateness verified that the feedback-based accumulative learning process improved the network as compared to the network only trained for the training data. Furthermore, each user gave feedback differently, so the network became a more personal model.

2) Trained networks: The trained networks are visualized in Fig. 8. In Figs. 8(a) and 8(b), the weights of fusion ART network before (after) accumulative learning based on

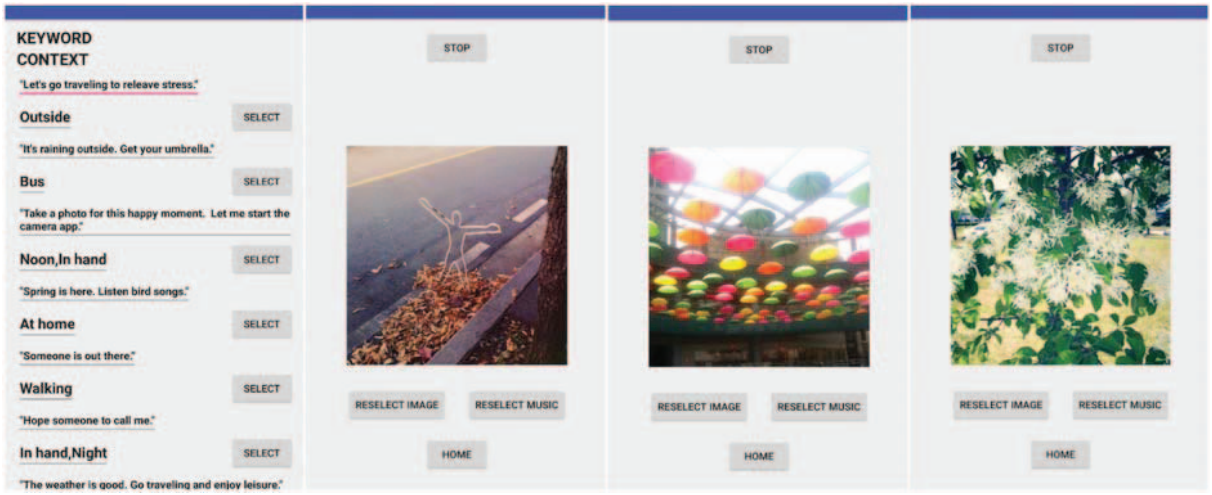
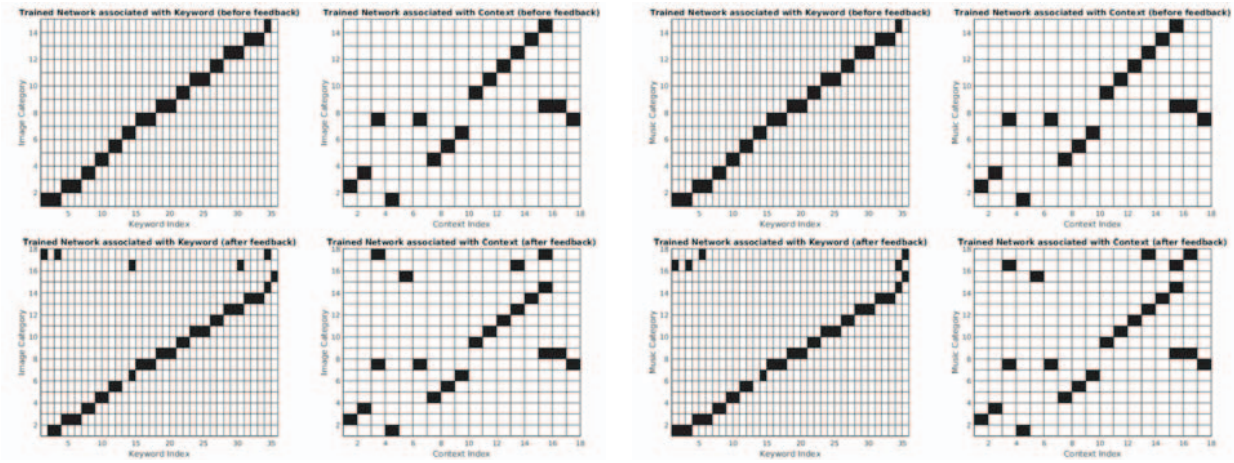


Fig. 7: The snapshots of the android smartphone application for testing the multimedia recommendation system.



(a) (b)

Fig. 8: The grid maps of each trained network. The weights of fusion ART network before (after) accumulative learning based on feedback from one user for the input of 36 keywords and for the input of 18 context information are shown in the upper-left and upper-right (lower-left and lower-right) parts, respectively, each for (a) image recommendation and (b) music recommendation.

feedback from one user for the input of 36 keywords and for the input of 18 context information are shown in the upper-left and upper-right (lower-left and lower-right) parts, respectively.

The weights in fusion ART updated by the process of template learning using (4) represent the associated input, so the weights in each fusion ART network shown in Fig. 8 have values of 0 or 1. A black grid represents 1, and a white grid represents 0. Since the input representing all the resonant inputs associated with a specific category node becomes the weights, even if the network is not trained with a large number of training data, the network can select the media file for all situations.

As shown in the grid maps, some weights connecting the existing category node and the input changed, and the number of category nodes were increased creating the weight values. The number of category nodes increased from 15 to 18 for both image and music recommendations. These changes in each network demonstrated that the network was trained incrementally based on feedback creating a new category node to represent a newly given situation if needed. The fact that the changes in the network were made differently for each user also demonstrated that each network was learned to be a more personal model that follows user preferences.

V. Conclusion

In this paper, we proposed a novel multimedia recommendation system using fusion ART networks. For multimedia recommendation, the keywords of conversations and context

information of the user were encoded as the inputs into two input channels, and the selected image and audio files were represented as the activated nodes in the category field in each fusion ART network. Based on feedback from the user including preferred media file, the fusion ART network is learned accumulatively. The proposed system is purposed to select multimedia to be conveyed in addition to the dialog between the digital creature on a smartphone with the user in terms of digital storytelling. The experimental results showed that the android smartphone application selected and recommended specific image and audio files appropriate to the certain situation based on the keywords and the context information of the user, and users were more satisfied with the recommended media files after accumulative learning based on feedback. The trained networks were shown to be learned incrementally through user's feedback and become more personal models. As a future work, we will include the emotion of user as an input in our system, as the emotion is also a criterion in selecting a proper multimedia. We will also compare the performance with other methods using a larger dataset.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea MSIP. (No.B0101-15-0551, Technology Development of Virtual Creatures with Digital DNA).

REFERENCES

[1] J. Lambert, *Digital storytelling: Capturing lives, creating community*, Routledge, 2012.
[2] K. Lundby, *Digital storytelling, mediatized stories: Self-representations in new media*, vol. 52, Peter Lang, 2008.
[3] F. Xia *et al.*, "Mobile multimedia recommendation in smart communities: a survey," *IEEE Access*, vol. 1, pp. 606–624, 2013.
[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no.6, pp. 734–749, 2005.
[5] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances Artificial Intell.*, 2009.
[6] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," *Adaptive Web*, Springer, Berlin Heidelberg, pp. 325–341, 2007.
[7] R. Burke, "Hybrid web recommender systems," *Adaptive Web*, Springer, Berlin Heidelberg, pp. 377–408, 2007.
[8] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern-recognition machine," *Comput. Vis., Graph., Image Process.*, vol. 37, pp. 54–115, Jan. 1987.
[9] A. Kaylani *et al.*, "An adaptive multiobjective approach to evolving ART architectures," *IEEE Trans. Neural Netw.*, vol. 21, no. 4, pp. 529–550, Apr. 2010.
[10] T. H. Oong and N. A. M. Isa, "Adaptive evolutionary artificial neural networks for pattern classification," *IEEE Trans. Neural Netw.*, vol. 22, no. 11, pp. 1823–1836, Nov. 2011.
[11] W. Wang *et al.*, "Neural modeling of episodic memory: Encoding, retrieval, and forgetting," *Trans. Neural Netw. Learning Syst.*, vol. 23, no. 10, pp. 1574–1586, 2012.
[12] W. Wang *et al.*, "A self-organizing multi-memory system for autonomous agents," in *Proc. Int. Conf. Joint Neural Netw. (IJCNN)*, pp. 1–8, Brisbane, Australia, Jun. 2012.
[13] Y.-H. Yoo and J.-H. Kim, "Procedural memory learning from demonstration for task performance," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, pp. 2435–2440, Hong Kong, Oct. 2015.
[14] A. K. Dey, "Understanding and using context," *Personal Ubiquitous Computing*, vol. 5, no. 1, pp. 4–7, 2001.
[15] K.-H. Han and J.-H. Kim, "Quantum-inspired Evolutionary Algorithm for a Class of Combinatorial Optimization," *IEEE Trans. Evol. Computation*, vol. 6, no. 6, pp. 580–593, Dec. 2002.