IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 June 2022, accepted 1 July 2022, date of publication 6 July 2022, date of current version 15 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3188852

RESEARCH ARTICLE

Intelligent Digital Human Agent Service With Deep Learning Based-Face Recognition

BUM-SOO KIM AND SANGHYUN SEO
School of Art and Technology, Chung-Ang University, Anseong 17546, South Korea

Corresponding author: Sanghyun Seo (sanghyun@cau.ac.kr)

This work was supported in part by the Culture Technology Research and Development Program 2021 through the Korea Creative Content Agency by the Ministry of Culture, Sports, and Tourism (A Specialist Training of Content Research and Development Based on Virtual Production) under Grant R2021040044; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant by the Korean Government through MSIT (Development of 5G-Based 3D Spatial Scanning Device Technology for Virtual Space Composition) under Grant IITP 2020000103001.

ABSTRACT This study proposes a framework for an intelligent agent information service using digital human and deep learning technology. The framework can recognize the identity of individuals using facial features and provide personalized services through a digital human. The personalized service is defined by a relevance graph based on personal data collected in advance. The proposed system can continuously evolve to recommend customized services using relevance graphs and dynamic data processing, gradually become more intelligent using additionally collected data. Moreover, it uses animation keyframe interpolation for natural and seamless digital human interaction and provides visual effects that are synchronized based on specific information collected for the intuitive service. The proposed system was tested on a school domain for two months, and a statistical domain feedback system based on a mathematical model that predicts service usage per unit time was developed using the recorded information. Additionally, we evaluate our system through user experience surveys.

INDEX TERMS Intelligent agent service, digital human, face recognition.

I. INTRODUCTION

With the development of the Internet, people can obtain the information they seek using various methods. The most common method of obtaining information is running a query on an Internet search engine; however, its effectiveness is strongly dependent on the query input. Digital signage technology is an emerging technique of information acquisition. It is a convergence platform that uses digital technology to display images and information on a display screen or projector and remotely manage it through a network.

The first designed digital signage simply displayed information based on a scenario. However, now the technology has developed into a two-way communication service that performs personal media functions using a touch panel. Hence, digital signage has recently been called "the 4th media" after television, personal computers, and mobile phones. This technology is connected to a network and can remotely control various content simultaneously. Moreover, based on its excellent ability of transmitting information, it can be an effective medium for providing advertisements and content.

A digital menu board is a typical example of applying digital signage in a commercial space. It displays menus with high visibility to simplify ordering food in fast food restaurants, cafes, and movie theaters. In addition, digital signage is actively used as advertising signage [1]–[3]. For example, a huge digital signage is being used as an advertising sign in Times Square, Broadway, New York, USA, demonstrating the effect of digital signage on attracting visitors by converting media art beyond the purpose of promoting advertisements.

Currently, most commercialized digital signage performs a one-way service that provides information based on text and images. However, in reality, they can provide advanced

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

services by reflecting user information or evolving into an intelligent form that offers services beyond the simple provision of scenario-based information. Recently, the utility of digital signage has been expanded by fusing it with other technologies such as near-field communication, haptic and gesture recognition, user behavior analysis, and face and gender recognition technology [4]–[9]. In the future, we would require an intelligent digital signage system that extends personalized services beyond the current simple information provision

Besides digital signage, natural language processing (NLP) technology offers an interactive information providing service by understanding text and spoken words; it is mainly embedded in mobile devices. For example, NLP in iPhone 6 (iOS8) allows Siri, a virtual assistant developed by Apple, to recognize voice commands and perform the corresponding specific tasks [10]. Recently, using personalized recognition, Siri has gained the ability to only recognize the characteristic voice of the user and not respond to other voices. Similarly, Bixby, a virtual voice assistant based on voice recognition technology developed by Samsung for their Galaxy S8 model, aims to simplify information acquisition to perform tasks. Such voice recognition-based services can provide a customized service using personal information available in the device. However, providing a similar visual service is challenging. Humans acquire 70% of their external information using their visual perception system; this sets a limit to the amount of information provided solely via voice data without visual information. Therefore, the existing personalization systems need to converge with digital signage to develop into an intelligent system.

In this paper, we propose an intelligent agent framework that combines personal identification and service provision technology. The proposed system, displayed in Figure 1, has the following technical characteristics:

- Adaptive data processing: For personalized service, data are processed dynamically; this includes preparation, manipulation, cleaning, and wrangling. It forms the basis of a progressive evolution system.
- Personalized face recognition: The system performs personalized face recognition based on deep learning technology with explainability. It has an expandable modular structure, which makes it easy to add recognition targets.
- Digital human control: As the main medium, the digital human provides information along with voice; this includes natural animation for a user-friendly service.
- Additional three dimensional (3D) virtual effects: The system provides data-synchronized 3D visual effects and changes, which improves visual intuition.
- Domain feedback system and self-evolution: When applied to a domain, the system continuously logs data to fit the feedback model. It gradually self-evolves, thereby inducing domain optimization.

The remainder of the paper is structured as follows. In Section 2, we discuss published works related to the proposed method. We describe the methods used for personalized face recognition in Section 3. In Section 4, the pipeline for data wrangling is discussed. We demonstrate the virtual character controls and additional visual effects in Section 5. In Section 6, we discuss the results of applying the proposed service to a domain and analyzing the feedback system. Finally, we discuss the conclusions of the study in Section 7.

II. RELATED WORK

Recently, a futuristic information-providing assistant-type service that helps users was proposed. At CES 2022, Samsung proposed a custom avatar life assistant named ‘‘Future Home,’’ which is a companion robot artificial intelligence (AI) avatar. MINDs Lab provides realistic AI virtual humans based on text-to-speech and speech-to-face technologies, which has been extended to various companies in the form of announcers, curators, and counselors to deliver an information provision service. As such, applications of intelligent agents are increasing rapidly in daily life.

We propose a digital human-based intelligent service that can provide personalized services. The proposed system uses personal identification, digital human control, and adaptive data wrangling techniques. Our system starts the service by recognizing a face. For personalized recognition, face recognition is performed using image-based deep learning. For personalized service, it is necessary to recognize the personal identification data (PID). Depending on the method of representing facial features, there are several approaches available for recognizing the face of an individual from an image. H. J. Mun, *et al.* proposed a method for finding the face of an individual by compressing the main features, such as eyes, nose, and mouth, in a face image, converting them into vectors, and matching them to the vectors of the target and input image [11], [12]. This method offers explainability by explicitly describing the main features. However, applying this method using other features is challenging. In addition, it requires a new descriptor to be defined for each feature of the target.

Similarly, there is a method of recognizing a face based on eigenface obtained using principal component analysis (PCA) that utilizes explicit feature definitions [13]–[15]. It creates an eigenface for a face image by carefully identifying the structural characteristics of that specific face. Thereafter, the eigenface is calculated as a Euclidean distance by matching the previously stored eigenface with the eigenface of the newly given image. This method recognizes the person in the image using the smallest Euclidean distance as a personalized face. However, it can provide accurate face recognition only when the structural characteristics of the object to be recognized are analyzed accurately; moreover, the structural characteristics of new objects must be analyzed again.

Feed forward backpropagation is an implicit method that uses deep neural networks (DNNs). This method does not define a descriptor or directly describe features [16], [17]. Instead, it extracts features and classifies data by itself.

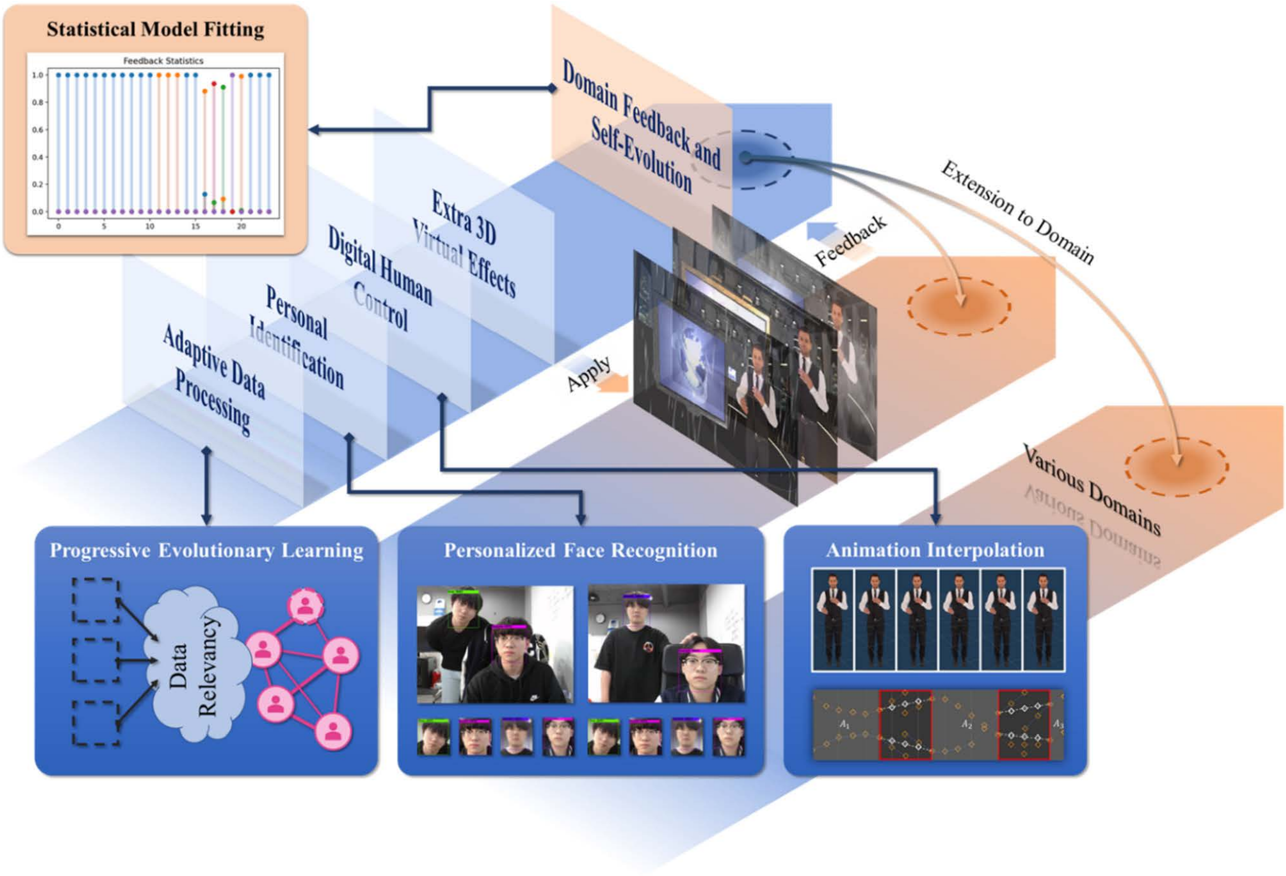


FIGURE 1. Overview of our proposed system.

With this mechanism, it is possible to classify the unique face of an individual by simply assigning each face a class number.

Digital human beings or virtual characters embedded in digital signage and kiosk machines are used for displaying digital information; they use intuitive communication to provide effective services. Developing human-friendly controls for digital humans allows us to maximize their advantages. Moreover, services using digital humans can be used for more intuitive applications. As controlling the animation of a digital human through non-verbal expression is an effective method of conveying information, developers enable a natural service centering on animation, which is a digital human control technology. This animation includes the natural action of a digital human, reaction through user interaction, and movement synchronized with utterance, and is controlled in synchronization with the service content. Digital human-based service is studied mainly in health care and medical domains. Easton and Katherine proposed the virtual agent service to support physical and mental comorbidities derive from individuals living [18]. This agent is used in health care domain with a format as chatbot. Loveys *et al.* proposed digital human service to report the mental stress and loneliness to younger and older adults during the COVID-19. This digital human is used remotely to support the individuals in society restriction [19]. Besides, research on avatar services combined with various virtual space and AI technologies is being conducted [20].

Recently, digital agents are being used in various fields as they converge based on AI technology. It is essential to develop intelligent agents optimized for domains rather than simply providing scenario-based information. In the mobile market, the agent mainly performs the requested service by recognizing a human voice based on voice recognition and NLP technology; representative examples include Siri by Apple and Bixby by Samsung. In the home Internet of Things field, Echo speaker by Amazon, NUGU candle by SK, and Google Home AI-speaker by Google perform integrated functions inside the house.

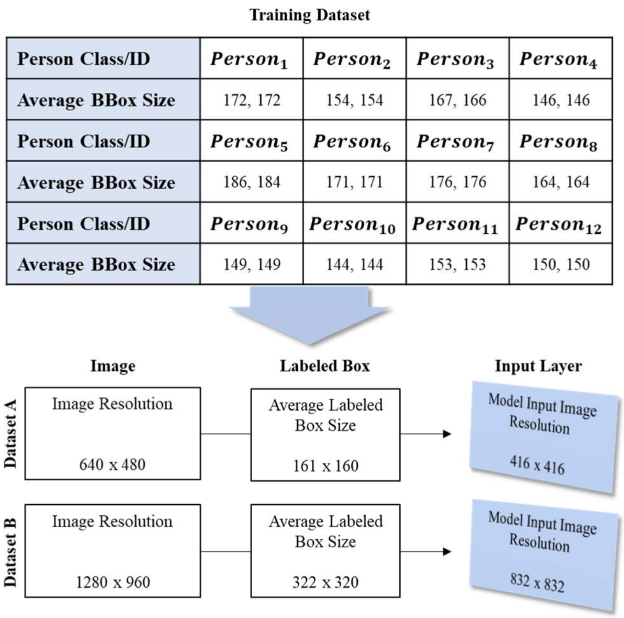
III. IDENTIFICATION OF PID FOR PERSONALIZED FACE

As identifying PID is critical for customized service, we used image-based deep learning for PID recognition. In this method, we must define a class name and number, and train the network by collecting face images in advance. After training, the network extracts the recognized face with a pre-defined class number through inference time.

For immediate response and real-time interaction, the proposed system required a bespoke model capable of real-time inference that included object localization. We used the You Only Look Once (YOLO) architecture [21], which is a real-time inference model. Specifically, we used the YOLOv4 lite model, denoted as Tiny, which can perform object detection at a speed over 400 FPS on RTX 2080Ti.

Using a lightweight deep learning model facilitated subsequent processing and avoided computational overload when information was provided.

The trained model calculated different training accuracies according to the meta data of the training dataset and learning method. Among several learning meta data, we compared recognition accuracy according to the input image of the network. For this experiment, a training dataset was built with images of two resolutions: 640 × 480, and 1280 × 960; we collected 2,000 face images of each individual in a group 12 people as the training dataset. Object detection or localization, which determines the position of the object in the image, was performed based on supervised learning. For localization, we annotated our custom dataset with a bounding box (BBox), hereinafter referred to as the ground truth (GT) BBox. To minimize human error in labeling, we used iterative pseudo-labeling using an open vision library, such as labeling-by-detection. After pseudo-labeling, we manually modified the position of GT BBox if it contained a completely different object or have part of an object in the images.



Training Dataset

Person Class/ID	$Person_1$	$Person_2$	$Person_3$	$Person_4$
Average BBox Size	172, 172	154, 154	167, 166	146, 146
Person Class/ID	$Person_5$	$Person_6$	$Person_7$	$Person_8$
Average BBox Size	186, 184	171, 171	176, 176	164, 164
Person Class/ID	$Person_9$	$Person_{10}$	$Person_{11}$	$Person_{12}$
Average BBox Size	149, 149	144, 144	153, 153	150, 150

	Image	Labeled Box	Input Layer
Dataset A	Image Resolution 640 x 480	Average Labeled Box Size 161 x 160	Model Input Image Resolution 416 x 416
Dataset B	Image Resolution 1280 x 960	Average Labeled Box Size 322 x 320	Model Input Image Resolution 832 x 832

FIGURE 2. Meta data of training dataset according to classes.

Figure 2 displays the meta data of the training image dataset focusing on BBox. Herein, pre-defined classes or people are capitalized, as *Person*$_{\#}$. The average BBox size for a 640 × 480 resolution image was 161 × 160 pixels; the BBox size changed because it was anisotropically rescaled in the input layer of the network. In addition, the statistics of the training dataset were calculated as mean = [0.444, 0.432, 0.428] and std = [0.305, 0.306, 0.301]. This varied from the statistics of ImageNet dataset, with mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225]. Results obtained according to the size of the input images are explained in Section 6. In addition, we compared image augmentation methods to determine the optimal augmentation for classification in similar objects, such as personalized face recognition.

IV. PROACTIVE DATA WRANGLING AND RELEVANCE GRAPH

A data processing pipeline is vital for dynamically acquiring service data; this includes data preparation, manipulation, cleaning, and wrangling. The detailed pipeline differs according to the purpose of data usage. The whole processing pipeline of general-purpose data, which are usually based on Internet data, is handled automatically; this includes weather data, news, and emergency alert messages. For example, weather data are obtained by requesting a query to the open application programming interface of a meteorological agency website. Thereafter, they are converted to an appropriate data form for the service in the pipeline. By contrast, custom data must be constructed in advance by humans.

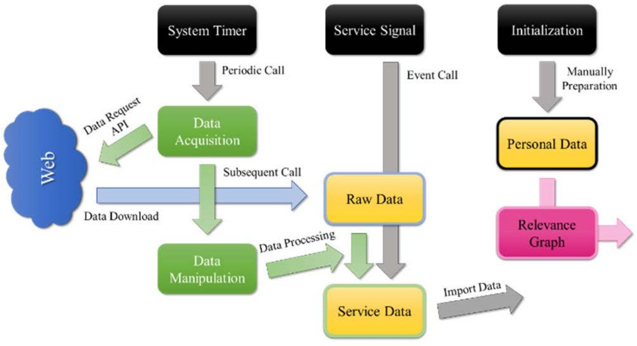


FIGURE 3. General purpose data processing pipeline.

The whole process is controlled by a system timer every unit time and uses synchronization for the renewal data. This ensures that the data are always up-to-date, and the updated data are loaded at service time. The pipeline of the general data processing is illustrated in Figure 3. The pipeline can be divided according to who controls the data. Data update is handled by a system timer and processing module, and data loading is handled by a system event call. The processing module is a scenario-based program suitable for data formatting and transformation of target data. This pipeline applies appropriate processing methods to audio, image, and text data.

Conversely, as personal data must be used as a knowledge-discovery system, they require another pre-analysis process. In addition, the data needs to be prepared manually before the pre-analysis. Personal data exhibit an object-semantic and hierarchical structure through interpersonal connections.

These connections are predefined as attribute-based relevance graphs. The initial relevance graph can evolve progressively into a large system if it is serviced continuously.

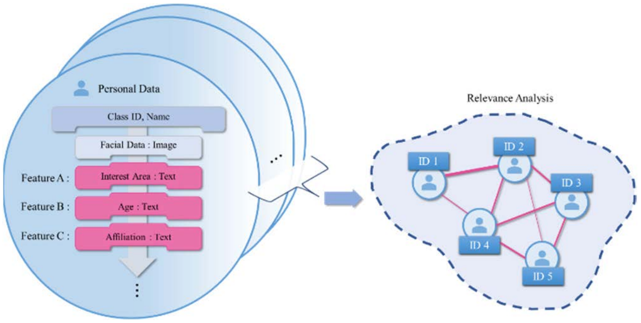


FIGURE 4. **Entities of personal data and relevance graph for custom service.**

Figure 4 displays the collected personal attribute structure and the transformed relevance graph. Personal attributes are extensible features, that is, personal data. For each feature notated to F, a relevance graph is constructed between predefined classes. If a person with ID 1 has similar data to a person with ID 2 for a particular feature, the connecting line represents a strong relation. Additional features can be added when specific information is needed in the applied domain; this can be used adaptively for relevance graphs.

A constructed relevance graph has the following characteristics: 1) easy scalability, 2) feature informatics estimation, and 3) gradual evolution. Relevance graphs consist of feature-based connections using personal data, which allows individuals with specific features to be easily added to an existing system. This is because the added person needs to match a predefined class only once to add relevance. A connection made with each related feature represents a human-to-human association. Two people with strong relation to features F_A, F_B, and F_C may have similar interests in feature F_D. By contrast, if two people have a weak association to features F_E, F_F, or F_G, they may have completely different tendencies. Relevance graphs imply an interpersonal relationship; therefore, they are used as a recommendation system to predict and provide information that has not been provided. These graphs become more stable and reliable as data accumulate, and the system gets bigger and serviceable. The scalability and informatics estimation functions allow the system to gradually expand itself to become more robust with generality.

We defined a system of relevance graphs between human informatics using personal data. The diagram and service process of this system are displayed in Figure 5. The graphs were pre-configured with the same number of features while training the model. Thereafter, the graph was rebuilt from the previous graph when a new person was added. When one predefined person was detected that connected with ID 01, 03, and 07, the property from those individuals was used to recommend specific information. In Figure 5, each node represents an individual, and each edge, which are lines connecting the nodes, indicates the degree of relevance between individuals. In case there was an abundance of interpersonal related data, the person was defined as a strongly related person. Conversely, a person with a lack of interpersonal related data was defined as a weakly related person. The graph on the left-hand side in Figure 5 displays the relevance expansion. If a new related graph needs to be defined, the relevance graph can be extended by adding new nodes while maintaining the existing graph. This allows the proposed systems to be more flexible for evolving. The graph on the right-hand side displays the process of providing recommendation information in the service for recognized people in real-time. First, when an identity was recognized, a valid related person was fetched from a pre-determined graph for each feature. Thereafter, the information was sorted through one final related graph based on that information. After weighting the sorted data, the data with the highest score were presented sequentially. This feature-based relevance is a progressively evolving algorithm with expanding and estimating operations.

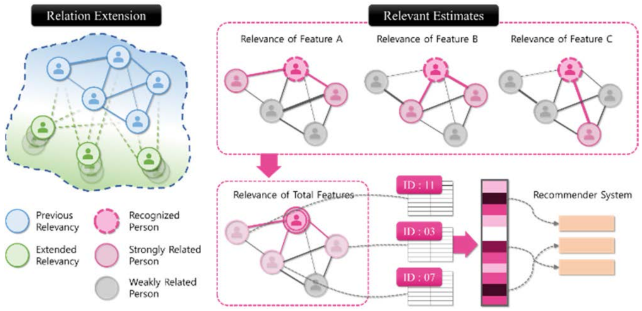


FIGURE 5. **Adaptive extension(left) and estimation (right) process of relevance between personalization data.**

V. VIRTUAL CHRACTERS WITH INTUTIVE VISUAL INFORMATION

For a visual naturalness that encompasses all states of a virtual character, we considered components that emphasized on animation. Commercially available digital humans usually provide simple animations, such as walking, talking, and laughing motions. However, these animations are presented in the form of unit animation blocks that are not linked to any other data. Therefore, when two or more block animations are operated continuously without other methods, the animation output displays discontinuous movement. To compensate for this unnaturalness, we used the animation keyframe interpolation method for continuous animation.

A. ANIMATION KEYFRAME INTERPOLATION

Natural movement in digital human allows easy usage of digital human-based services. To minimize unnatural animation, we used the animation keyframes interpolation method. The keyframe interpolation method connects the keyframes at the end of a previous animation and the beginning of the next animation such that they appear to be connected naturally using interpolation equation. In this study, we divided the interval to apply the interpolation method in the animation block. The main motion interval in an animation was assumed to be the middle part of the animation and was defined as an atomic interval where motion is guaranteed. Both ends, except for the atomic interval, were set as interpolable intervals.

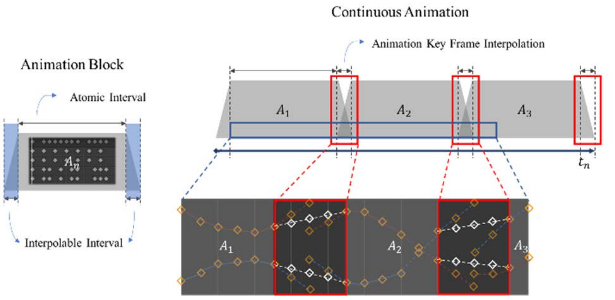


FIGURE 6. Divided animation block to atomic and interpolable interval and interpolation process.

Figure 6 displays the abbreviated interpolation process and segmented block animation interval. In order to define the proper interpolable intervals, the intervals were divided into several cases and evaluated for visual awkwardness. Animation interpolation was performed in increments of five frames. As a result, regardless of the duration, the animation was most natural when the interpolation interval was set to 40 frames at both ends. Atomic interval depends on the interpolable interval; therefore, it will have a different length for each animation.

Unlike curve fitting, interpolation is configured to go through all data points in an exacting manner. As a result, it is vulnerable to noise and exhibits low robustness. However, applying interpolation causes negligible error because there are no noise data in animation keyframes constructed discretely by humans. Interpolation methods include polynomial, rational, sinusoidal, and spline interpolation. We used first-order Lagrangian interpolation, which is the simplest form of polynomial interpolation, since connecting two animations does not require precise numerical estimation of each keyframe coordinates or defining a specific interpolation function. Even when a function such as natural cubic spline was applied using three or more keyframes, there was no difference in the visual movement of the virtual character. During interpolation, new points were linearly created in the interpolation interval by concatenating the two endpoints of the atomic interval. Therefore, keyframes between the two endpoints are discarded.

$$f_{n=1}(x) = \sum_{i=0}^{n=1} y_i l_i = \sum_{i=0}^{n=1} y_i \prod_{\substack{j=0 \\ j\neq i}}^{n=1} \frac{x-x_j}{x_i - x_j} \tag{1}$$

For interpolation we used Lagrange equation as (1). As we designed an order of polynomial to 1, interpolation equation f becomes linear function l_i since cardinal function is composed of $l_0 = x - x_1 / x_0 - x_1$ and $l_1 = x - x_0 / x_1 - x_0$.

$$P_{new}(k; A_N \cap A_{N+1}) = f_{n=1}\left(\frac{|x_1 - x_0|}{2I} k; A_N \cap A_{N+1}\right) \tag{2}$$

(A_n is n-th played animation Block; k is the index of generated point in the interpolable interval; f is Lagrangian interpolation function and l is cardinal function of f; x_0, x_1 represent the end point of the previous animation keyframe and beginning of next animation keyframe respectively; y_0, y_1 represent positional value in 3-dimensional virtual space about x_0, x_1. I is the length of the interpolable interval, which was set to 10; P_{new} represents the generated new points through interpolation.)

Our key-frame interpolation function in (2) was designed from basic Lagrange equation Eq (1). Both endpoints x_0 and x_1 used for interpolation were assigned by the previous animation A_n and the next animation A_{n+1}, respectively. Function $f_{n=1}$ was defined with two points, and P_{new} was generated in the interpolable interval between them. P_{new} was divided by the length of interval $2I$, which is equal to x_1 when A_{n+1} is reached. The increment was defined by the index k of the generated keyframe. When k is 0 and $2I - 1$, it was defined as x_0 and x_1, respectively; discrete value k increased from 0 to 19 based on the interval discrete length. Therefore, excluding the existing x_0 and x_1, a total of 18 keyframes were generated for interpolation using $f_{n=1}$. For natural animation, first-order Lagrange function was used to estimate the connection between two keyframes, which avoided the uncanny valley issues caused by the mechanical movement of virtual humans.

B. ADDITIONAL VISUAL EFFECTS SYNCHRONIZAED WITH DATA

For a more intuitive service, we used a particle system in the game engine to add visual effects, which displayed an appropriate visual effect based on the type of data being provided. For example, in the event of a rainy day, the visual effect of rain was the same as the current weather because the stored data reflected the latest rain information. As a result, with the rain falling effect, users could easily determine that it was raining.

Figure 7 displays the effect of snow and rain falling in the game engine. Weather effects were created using particle effects with different properties of textures depending on the attributes of a specific weather. Each visual effect was controlled in detail. For example, the number of rain unit particles increased based on actual meteorological agency data such as precipitation and rate over time. Other additional effects are listed in Figure 8.

Figure 8 displays the visual effect table of weather data according to the condition of effect being played. We defined four effects for weather: snowy, rainy, sunny, and cloudy/foggy. All these effects were played based on weather

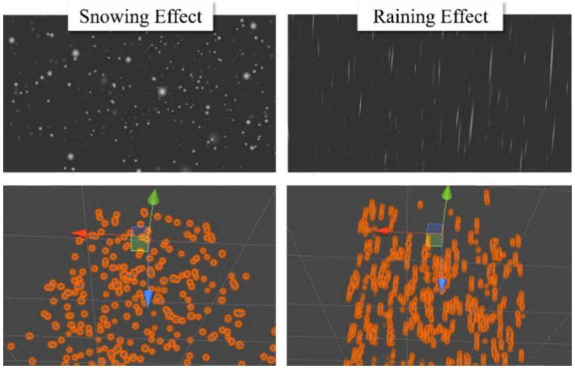
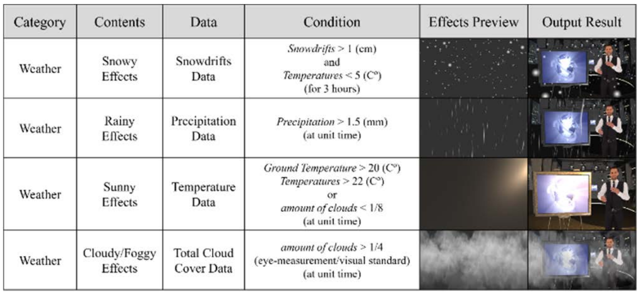


FIGURE 7. Visual effects to provide intuitive weather information about snowy and rainy days.



Category	Contents	Data	Condition	Effects Preview	Output Result
Weather	Snowy Effects	Snowdrifts Data	*Snowdrifts* > 1 (cm) and *Temperatures* < 5 (C°) (for 3 hours)		
Weather	Rainy Effects	Precipitation Data	*Precipitation* > 1.5 (mm) (at unit time)		
Weather	Sunny Effects	Temperature Data	*Ground Temperature* > 20 (C°) *Temperatures* > 22 (C°) or *amount of clouds* < 1/8 (at unit time)		
Weather	Cloudy/Foggy Effects	Total Cloud Cover Data	*amount of clouds* > 1/4 (eye-measurement/visual standard) (at unit time)		

FIGURE 8. Visual effects table for state of weather data and effects preview with resultant service scene.

data collected from the Korea Meteorological Administration website. If the precipitation variable of the meteorological data was 1.5 mm or more, the rain effect would automatically be played in the current scene with the condition check.

VI. EXPERIMENTAL RESULTS

To evaluate the efficiency of the proposed system, we expanded and built it in an educational facility, where it was in service for about two months. During validation, we recorded meta data for identified classes, dynamic data processing times, and types and times of information provided in text format. This recorded information was used to analyze the results for the applied services. The system was installed during non-vacation periods and operated all day.

A. ENVIRONMENT SETTINGS

The designed observation environment for the experiment is represented schematically in Figure 9, which consisted of a digital display device exposed in the aisle and an externally connected computing machine for real-time object detection and rendering. The system was located at the entrance to the facility. At the beginning of the service provision, a person who wants to use this service will approach the display more closely. Once their face is recognized, the service will begin with character animation and speech. Next, using the data collected from the applied domain, a mathematical model of visits and probabilities would be fitted, as described in Section 6.4.

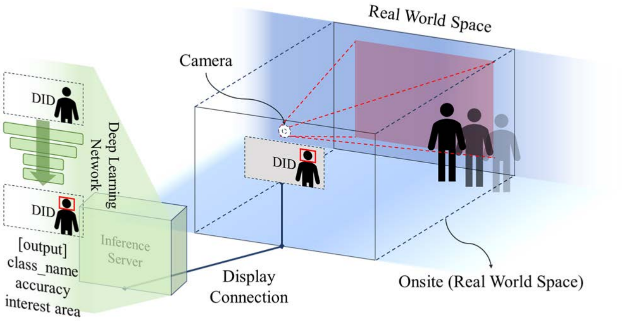


FIGURE 9. Experimental environment of our system with digital information display(DID).

B. TRAINING RESULTS

We performed several comparative experiments to account for personalized recognition. As personalized face detection is a subcategory of similar object classification, generic augmentation methods may not be useful. For example, Blur method, which softens the sharpness of an image, can confuse the model in training. This is because the model is strongly dependent on local features such as eyes, nose, and mouth for face recognition. Therefore, we trained a model using each of the image augmentation methods and compare the validation metrics.

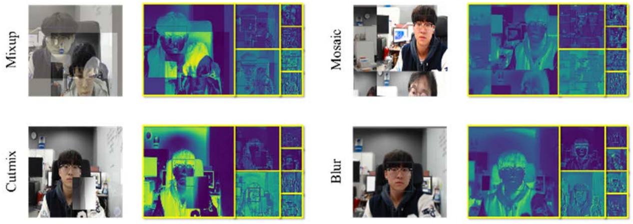


FIGURE 10. Augmented images and their feature map.

To evaluate our model according to augmentation method, we compared four simple image augmentation methods: MixUp [22], CutMix, Mosaic, and Blur. MixUp involves overlapping images with transparency, whereas CutMix involves replacing part of an image with another image. In Mosaic, four images are glued into one large image, and a specific rectangular area is extracted as the training image. Blur reduces the sharpness partial area in an image. Augmented images and feature map corresponding to these are shown as Figure 10. Feature maps are extracted from convolution layers after each pooling layers.

The abovementioned four augmentation methods were compared with the no augmented model notated to base model. We compared the results using three metrics: F1-score, mean average precision (mAP) at 75, false negative (FN). In particular, we extracted these metrics according to the resolution of the image; 640 × 480 and 1280 × 960. In training our YOLOv4-based Tiny model, we set the batch and max iteration as 64, 24,000 respectively. Other hyper-parameters are used as default YOLOv4 [21]. In training, we conducted transfer learning from pre-trained model which is trained using COCO dataset [23]. Our validation dataset was prepared along with the training dataset, which has the same statistics.

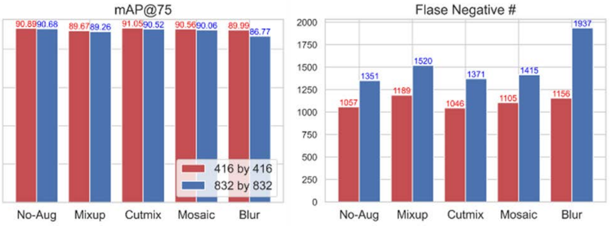


FIGURE 11. Validation results according to the augmentation method.

As per the comparison results shown in Figure 11, all models showed almost similar results compared to the base model in mAP. In the case of the Mixup model, the performance was rather low compared to the basic model. In the CutMix model, mAP was improved by 0.16 at 416 × 416 resolution, buy 0.16 lower at 832 × 832 resolution of input layers. The Mosaic model also showed lower performance at all resolutions like Mixup model. In particular, the Blur model also showed low performance with the lowest mAP value among all models. We can explain these results using the feature map shown in Figure 10. As an augmentation method, two images were superimposed on the feature map of the image augmented using MixUp. Because the feature components of the two faces overlap according to the feature map, we judged that face recognition accuracy of the model would have decreased. In the case of the Mosaic, since the feature maps only contain parts of the GT BBox in training, we through the validation accuracy would have decreased. Similarly, the Blur method of smoothing the image would have had a negative impact on the localization process where the model finds objects in the image. Because of this, it seems that mAP which evaluates the intersection ratio of the predicted BBox and GT BBox, was measured rather low. We can see that the feature map has a blurred background in Figure 10 when compared to other feature maps. Unlike the above methods, we conclude that the augmented image using CutMix model learns to replace a specific area with another image to increase the robustness of occlusion and noise data. When comparing the augmentation methods, we also extracted the validation metrics according to the input image resolution. As a result, the 832 × 832 image had slightly lower results, and the difference in accuracy of the model according to the augmentation method was similar to the calculated value for the 416 image.

The FN results shown in Figure 11 also showed similar results to those of mAP@75. The number of FN for each model and the base model were almost similar like the results of the mAP, and only the Blur model yielded 2,037 FN at 832 × 832 resolution of input layers. The model that performed better than the base model was the CutMix model, which yielded an FN of 1,146 at 416 × 416 input images. In the case of F1-Score, all models showed about 0.94 or 0.95 values. The Mixup model and Blur model at 832 × 832 resolution showed 0.93, 0.92 respectively.

C. FACE RECOGNITION PERFORMANCE

To explain recognition on the installed domain, we validated personalized detection accuracy by setting the identification measurement distance between the camera and individual as 50 cm, which is the most common distance used in the real world. With distance-based performance explanation, we can provide the location of an appropriate service installation for an applied domain.

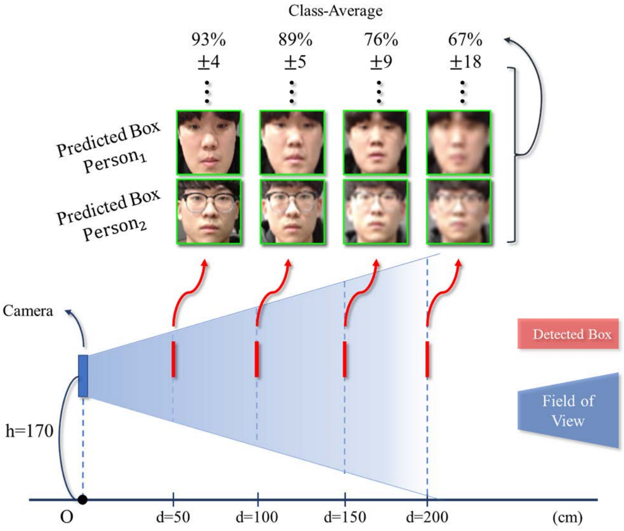


FIGURE 12. Face recognition accuracy according to service distance. In this test, facial dataset was acquired in our service applied space.

Figure 12 displays the inter-class average accuracy of our model and composition of measurement environment according to distance. As per our prediction, since the pixels per meter value was highest when the distance is low, the average face recognition exhibited highest accuracy when the distance was 50 cm. Predicted accuracy exhibited similar results for all predefined classes. The prediction result in Figure 12 was evaluated in an environment different from that of Figure 11. These changed environmental conditions included saturation, brightness, relative position of camera and individual, and position of light source. In addition, we recorded data in the wild scene to explain the statistics during service. Statistics for this dataset were: mean = [0.630, 0.513, 0.449] and std = [0.282, 0.248, 0.223]. The location where the domain was installed was equipped with strong lighting sources. Therefore, for a face image recognized in the test, the pixel mean was brighter than the training dataset. Even the average value of the R channel was obtained as 1.5 times higher than that of the training set; therefore, the statistics between the test and training datasets were distinctly different. In addition, our model was trained on a training dataset that was built for about three months before the service was installed. Therefore, if various evaluated data is continuously reflected in the domain model and additional learning is performed for each class, customized service according to personalization recognition is possible if contact distance is within 1.5 m.

D. DOMAIN FEEDBACK STATISTICAL MODEL

After installing the aforementioned service, we designed a feedback system as an optimal application to the domain; this involved creating a mathematical model that predicted the number of customer visits using service data. Service data has discrete events, with strong randomness and no causality about other events unless a specific event in considered. In addition, since the target events are countable, our feedback system model can be designed by utilizing Poisson distribution.

$$q\left(w;\lambda\right)=\frac{\lambda^{w}e^{-\lambda}}{w!} \tag{3}$$

Poisson distribution is a discrete probability distribution that represents the probability of an event occurring within a unit of time, with expectation and variance equal to λ in (3). It is mainly used for predicting under discrete data given specific unit time [24]–[26]. According to (2), given λ, which is the expectation of the number of times an event will occur within a given time, the event will occur w times. In our system, these probabilities were summed to predict the range wherein an event will occur within a unit period. In other words, it was equivalent to calculating the probability that an event will occur 5 to 10 times between 1:00 PM and 2:00 PM. It should be noted that the time unit notation followed for this study was the Korean Standard Time (KST) GMT +9 standard.

$$H_{h_i}(m)=\sum_{w=m}^{w=m+M}q\left(w;\lambda_{h_i}\right)=\sum_{w=m}^{w=m+M}\frac{\lambda_{h_i}^{w}e^{-\lambda_{h_i}}}{w!} \tag{4}$$

We set the unit time to 1 hour and defined a specific interval as the variable M to predict the range of times an event would occur. To predict an event as a range, (4) was designed using the Poisson distribution equation (3). Using the basic Poisson model (3), a single probability value was calculated for a specific single event rather than a range. Since predicting a single event is not useful, our system predicted a specific range, which is the sum of each probability within interval M and denoted as $H_h(m)$. This variable is dependent on the value of λ in unit time, the interval of interest of the random variable. Since λ signifies an average of the number of times an even will occur within a unit period, it must be recalculated whenever the unit time is changed. With data recorded from the domain, λ can be modified as time unit changes.

As per Figure 13, $H_h(m)$ was plotted on the y-axis and unit time h_i was plotted on the x-axis. Smaller M values offer higher predictability; however, calculating the probability of an event occurring 1–2 times or 3–4 times is not useful for a feedback system; this is because it provides extremely local information. Therefore, considering the characteristics of the experimental space, we set value of M as 5. Although the proposed system can adaptively evolve and become robust, we derived a statistical model using collected data for just two months for the experiment.

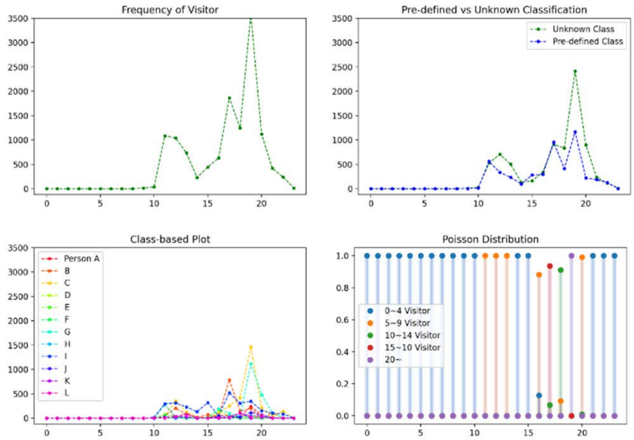


FIGURE 13. Analysis and prediction using our statistical model about visitation data obtained for two months.

The analysis result of the collected data is displayed in Figure 13. The top-left graph in same figure displays the data collected for two months. This graph represents raw data on the frequency of visitors. The x-axis and y-axis represent unit time and number of visits, respectively. The top-right graph displays the classification of frequencies based on whether the predicted class is a predefined or unknown class. If the inferred accuracy was below the threshold, the class was defined as an unknown class. The threshold was set to 0.4. In the graphical result, the frequencies for the unknown and predefined classes at specific times such as 11:00 AM and 5:00 PM exhibited similar values. Therefore, we concluded that a more accurate learning about identity recognition is required. For unseen data, the model needs to be trained using additional methods such as zero-shot learning [27], [28]. The bottom-left graph represents the classification of frequencies, which is number of visits to each predefined class. It displays the calculated probabilities for a range of visits using a probabilistic model fitted to the collected data. Since visit rate varied with period, we defined the model per unit time. A total of 24 models predicted probability values with different λ based on the collected data. As per the collected data (top-left), a relatively large amount of flow was obtained around 12:00 PM and 6:00 PM, which are lunch and dinner time, respectively. The model reflecting this also derived the result that there would be a lot of visits during lunch and dinner hours. Considering the amount of domain movement and predetermined the classes, we scaled the maximum visit value to 20. Therefore, if the model predicted a value of 20 or higher in a certain period, it meant that the number of individuals moving through the system would be higher than the number of people trained.

E. USER EXPERIENCE

To evaluate our system from a user experience perspective, we surveyed 30 individuals. These people included those who were pre-trained and those who were not, in a ratio of 4:6. For questionnaires, we grouped the survey categories by technology, intuition, and service. The details of questionnaire are shown in Table 1.

TABLE 1. Questionnaire about survey category.

Category	Question Coding	User	Question
Technology	Q1	Pre-trained User	Is the face recognition performance accurate? (1: not accurate ~ 7: very accurate)
Intuition	Q2	All	Can the digital human and extra effects help to you improve intuition? (1: not helpful ~ 7: very much intuitive)
Service	Q3	All	Are you thinking of utilizing our system for your domain? (1: not preferred ~ 7: very much preferred)
	Q4	All	Is the personalized service information useful? (1: not useful ~ 7: very much exciting)
Future Suggestion	S1	All	Please leave a comment if you have additional suggestion you would like to be added in the future.

For responses analysis, we set the scale of evaluation scores from 1 (very negative) to 7 (very positive). We determine a score of 5 or more as a positive evaluation and a score of 3 or less as a negative evaluation. From Q1 in Table 1, we asked the face recognition accuracy, which is the core technology of our system. Responses to Q1 were received only from in the pre-trained class. In Q2, we ask if it is intuitive to provide information, using a virtual environment with a digital human and visual effects. It is important part of our system about information provision service. From service category, we asked that our system is useful or not. Q3 is designed to ask if people are willing to use our system in the future or their domain. And it is asked in Q4 whether the personalized service provide helped them build their behavior or schedule. In addition, we gathered comments through S1 to accommodate several suggestions from reviewers for the future research and applications. In S1, each individual suggested an idea or improvement in text data format.

Figure 14 shows the results of responses through heatmap and box plot. From Figure 14 (a), we can find that many reviews indicated positive consensus (upper than 4), marked in dark blue, on a score range of 5 to 7, and there were very few negative reviews. Specifically, on each question, 75%, 93%, 93% and 63% people gave a positive rating. As an analytic result, from Figure 14 (b), each question, marked with a purple diamond, was rated positively with an average score of 5.25, 5.73, 5.93 and 4.93. In particular, the number of people who reviewed very much positive (7 score) in the Q3 response was 8 (26%), which is raw data and is scattered with green point. As a result of our survey, we found that, on average, about 81% of people rated our system positively for all questions asked.

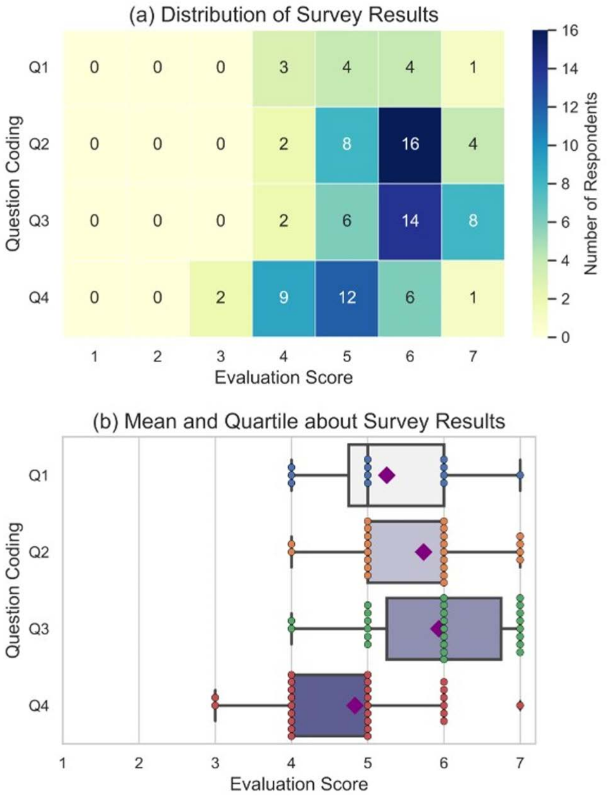


FIGURE 14. Analytics of survey results according to Q#.

In addition, reviewers suggested several application-level ideas and directions for improvement in S1. We were able to divide the suggestions into 3 groups: 1) It will be even more interesting if there are application contents such as various visual effects, virtual characters, and interactive elements based on domain themes, 2) It is hoped that the pipeline of our system will be well-established and standardized for efficient use by other content developers, 3) It is hoped that personalized data would be provided in a richer category and include various information. Through S1, we thought that people wanted more diverse information, which could be explained by the relatively low response to Q4 compared to other questions.

VII. CONCLUSION

In this study, we proposed an intelligent agent service that can provide customized services using digital humans.

The proposed system applied personalized face recognition using image-based deep learning technology, an adaptive data processing pipeline to provide customized information, and digital human animation interpolation technology for human-friendly services. The personalized face recognition was implemented using deep neural network. For a more accurate classification, the face recognition performance was compared to different augmentation methods using F1-Score, mAP, and FN. According to the results, only the CutMix method exhibited an accuracy improvement. To provide the information to be used in the services, we utilized an adaptive pipeline for data acquisition, processing, and transformation. It was designed to intelligently build data to enable customized services and evolve gradually. Furthermore, to provide intuitive communication, the service was represented by a digital human through a digital display. The animation interpolation method was applied for a more human-friendly service, which prevented the uncanny valley problem and helped users use the service more easily. In addition, visibility was improved by adding visual effects to certain services, which appeared adaptively as per the service content and helped users recognize the service immediately. The proposed system was integrated in the game engine, which made it easy to scale and be applied to various domains. To verify the capability of the intelligent agent service to provide customized service, the basic system was extended and applied to a school domain. This system was installed and operated at the training site for two months. Based on the contents serviced during this period, a feedback-enabled verification system was created. The service occurrence exhibited discrete characteristics and strong randomness; hence, it was modified using Poisson distribution for mathematical fitting. The fitted statistical model induced optimization in the applied domain by providing feedback information such as service occurrence frequency prediction, personalized face recognition accuracy, and unlearned classification. These results confirmed that the proposed service can be applied to a domain and expanded gradually.

In user experience, on average 81% of people rated our system as useful, and about 93% said they would like to utilize it in their own facility in Q3 as shown in Figure 14. Moreover, from Q1, we thought that personalized face recognition model should be retrained to be robust to the environment with additional dataset. And from Q4 and S1, we conclude that advanced algorithms such as GNN(Graph Neural Network) are needed to more analytically model interpersonal relevance graphs in the future works. According to the suggestions of the survey in S1, if additional functions specific to the application domain are included, our system can be utilized as an intelligent agent in various fields.

REFERENCES

[1] K.-C. Yin, H.-C. Wang, D.-L. Yang, and J. Wu, "A study on the effectiveness of digital signage advertisement," in *Proc. Int. Symp. Comput., Consum. Control*, Jun. 2012, pp. 169–172, doi: 10.1109/IS3C.2012.51.

[2] N. Kuratomo, H. Miyakawa, S. Masuko, T. Yamanaka, and K. Zempo, "Effects of acoustic comfort and advertisement recallability on digital signage with on-demand pinpoint audio system," *Appl. Acoust.*, vol. 184, Dec. 2021, Art. no. 108359.

[3] C. Bauer, N. Kryvinska, and C. Strauss, "The business with digital signage for advertising," in *Information and Communication Technologies in Organizations and Society* (Lecture Notes in Information Systems and Organisation), vol. 15, F. Ricciardi and A. Harfouche, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-28907-6_19.

[4] H. Limerick, R. Hayden, D. Beattie, O. Georgiou, and J. Müller, "User engagement for mid-air haptic interactions with digital signage," in *Proc. 8th ACM Int. Symp. Pervasive Displays*, New York, NY, USA, Jun. 2019, pp. 1–7, doi: 10.1145/3321335.3324944.

[5] H. Limerick, R. Hayden, D. Beattie, O. Georgiou, and J. Müller, "User engagement for mid-air haptic interactions with digital signage," in *Proc. 8th ACM Int. Symp. Pervasive Displays*, New York, NY, USA, Jun. 2019, pp. 1–7, doi: 10.1145/3321335.3324944.

[6] G. Alfian, M. F. Ijaz, M. Syafrudin, M. A. Syaekhoni, N. L. Fitriyani, and J. Rhee, "Customer behavior analysis using real-time data processing: A case study of digital signage-based online stores," *Asia Pacific J. Marketing Logistics*, vol. 31, no. 1, pp. 265–290, Aug. 2019, doi: 10.1108/APJML-03-2018-0088.

[7] A. Greco, A. Saggese, and M. Vento, "Digital signage by real-time gender recognition from face images," in *Proc. IEEE Int. Workshop Metrol. Ind. IoT*, Jun. 2020, pp. 309–313, doi: 10.1109/MetroInd4.0IoT48571.2020.9138194.

[8] S.-H. Lee, M.-K. Sohn, and H. Kim, "Implermentation of age and gender recognition system for intelligent digital signage," in *Proc. SPIE*, Dec. 2015, pp. 123–133.

[9] R. E. Abraham and K. M. Robert, "Intelligent digital signage system based on gender identification," in *Intelligent Embedded Systems* (Lecture Notes in Electrical Engineering), vol. 492, D. Thalmann, N. Subhashini, K. Mohanaprasad, and M. Murugan, Eds. Singapore: Springer, 2018, doi: 10.1007/978-981-10-8575-8_25.

[10] S. Team, "Hey Siri: An on-device DNN-powered voice trigger for Apple's personal assistant," *Apple Mach. Learn. Res.*, Oct. 2017. [Online]. Available: https://machinelearning.apple.com/research/hey-siri

[11] S. S. Z. Zhaoqing and L. I. Zhicheng, "Face images recognition research based on smooth filter and support vector machine," in *Proc. 29th Chin. Control Conf.*, 2010, pp. 2760–2764.

[12] A. Anggo and L. Arapu, "Face recognition using fisherface method," *J. Phys., Conf. Ser.*, vol. 1028, p. 012119, 2018.

[13] M. Ü. Çarıkçı and F. Özen, "A face recognition system based on eigenfaces method," *Proc. Technol.*, vol. 1, pp. 118–123, Jan. 2012, doi: 10.1016/j.protcy.2012.02.023.

[14] M. O. Faruqe and M. A. M. Hasan, "Face recognition using PCA and SVM," in *Proc. 3rd Int. Conf. Anti-Counterfeiting, Secur., Identificat. Commun.*, Aug. 2009, pp. 97–101.

[15] S.-J. Kim, "Face recognition using PCA and face direction information," *J. Korea Inst. Inf., Electron., Commun. Technol.*, vol. 10, no. 6, Dec. 2017, pp. 609–616, doi: 10.17661/JKIIECT.2017.10.6.609.

[16] K. Teoh, R. Ismail, S. Naziri, R. Hussin, M. Isa, and M. Basir, "Face recognition and identification using deep learning approach," *J. Phys., Conf.*, vol. 1755, no. 1, Feb. 2021, Art. no. 012006.

[17] M. Arsenovic, S. Sladojevic, A. Anderla, and D. Stefanovic, "FaceTime—Deep learning based face recognition attendance system," in *Proc. IEEE 15th Int. Symp. Intell. Syst. Informat. (SISY)*, Sep. 2017, pp. 53–57.

[18] K. Easton, S. Potter, R. Bec, M. Bennion, H. Christensen, C. Grindell, B. Mirheidari, S. Weich, L. de Witte, D. Wolstenholme, and M. S. Hawley, "A virtual agent to support individuals living with physical and mental comorbidities: Co-design and acceptability testing," *J. Med. Internet Res.*, vol. 21, no. 5, May 2019, Art. no. e12996, doi: 10.2196/12996.

[19] K. Loveys, M. Sagar, I. Pickering, and E. Broadbent, "A digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: Randomized pilot trial," *JMIR Mental Health*, vol. 8, no. 11, Nov. 2021, Art. no. e31586, doi: 10.2196/31586.

[20] M. Korban and X. Li, "A survey on applications of digital human avatars toward virtual co-presence," 2022, *arXiv:2201.04168*.

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[22] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[23] T. Lin, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 8693, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, doi: 10.1007/978-3-319-10602-1_48.

[24] G. Y. Kim, H. T. Joo, and H. J. Kim, "A study on probability of heavy snowfall disaster using Poisson distribution," *Crisis Emergency Manage.*, vol. 13, no. 5, pp. 135–151, Apr. 2017, doi: 10.14251/crisisonomy.2017.13.5.135.

[25] C. Choi, "Regional risk analysis based on Poisson distribution and its implications," *J. Korea Planning Assoc.*, vol. 8, no. 5, pp. 319–331, 2013.

[26] T.-J. Kim, H.-H. Kwon, and Y.-S. Shin, "Frequency analysis of storm surge using Poisson-generalized Pareto distribution," *J. Korea Water Resour. Assoc.*, vol. 52, no. 3, pp. 173–185, Mar. 2019.

[27] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning—The good, the bad and the ugly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3077–3086, doi: 10.1109/CVPR.2017.328.

[28] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. M. J. Wu, "A review of generalized zero-shot learning methods," 2020, *arXiv:2011.08641*.

BUM-SOO KIM is currently pursuing the bachelor's degree with the College of Art and Technology, Chung-Ang University. His research interests include computer vision, computer graphics, and artificial intelligence.

SANGHYUN SEO received the B.S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 1998, and the M.S. and Ph.D. degrees from the GSAIM Department, Chung-Ang University, in 2000 and 2010, respectively. He was a Senior Researcher with G-Inno Systems, from 2002 to 2005. He was a Postdoctoral Researcher with Chung-Ang University, in 2010, and the LIRIS Laboratory, Lyon 1 University, from February 2011 to February 2013. He has worked at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, from May 2013 to February 2016. He has also worked at Sungkyul University, from March 2016 to February 2019. He is currently a Faculty Member with the College of Art and Technology, Chung-Ang University. His research interests include computer graphics, non-photorealistic rendering and animation, real-time rendering using GPU, VR/AR, and game technology. He has been a program committee member of many international conferences and workshops. He has been a Reviewer of *Multimedia Tools and Applications* (MTAP), *Computers and Graphics* (Elsevier), U.K., the *Journal of Supercomputing* (JOS), and *The Visual Computer* (Springer). He has edited a number of international journal special issues as a Guest Editor, such as the *Journal of Real-Time Image Processing*, the *Journal of Internet Technology*, and *Multimedia Tools and Applications*. He has been an Associate Editor of the *Journal of Real-Time Image Processing*, since 2017.