Generative Personal Assistance with Audio and Visual Examples

Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino

Abstract

The rapid progress of deep learning is affecting the world we live in. Media generation (i.e., image and audio generation) is a typical example of this progress, and impressive research results are being reported around the world. In this article, we first overview this very active research field. Then we introduce our efforts in developing a generative personal assistance system with audio and visual examples. Specifically, we explain how our new deep-learning approach will overcome the limitations encountered in existing studies on personal assistance systems. Finally, we discuss future directions in the media generation field.

Keywords: deep learning, media generation, personal assistance

1. Rapid progress in deep learning

The rapid progress achieved recently in deep learning is having an impact on the world we live in. Media generation (i.e., image and audio generation) is a typical example of this progress. Many studies are being done in this field, and amazing research results are being reported around the world.

For example, StackGAN [1] can automatically synthesize photorealistic images only from a text description. Neural style transfer [2] can convert arbitrary images to arbitrary-style ones (e.g., Gogh-style or Monet-style images) without any manual processing.

Until a few years ago, the main objective in deep learning was to improve accuracy for comparatively easily defined tasks such as image classification and speech recognition. However, more complex tasks have recently been tackled by introducing new models and theory. Also, a major breakthrough was recently achieved in the media generation field due to the emergence of deep generative models. Consequently, expectations for media generation are rising as a tool to embody various wishes.

2. Generative personal assistance

When we do new things (e.g., try to throw a ball faster, try to pronounce English more fluently, or try to acquire an appropriate facial expression according to the time, place, or occasion), it is sometimes difficult to know what the most effective methods for doing so are, and this can cause frustration. One conceivable solution is to find a teacher and have him or her teach us; another is to search via the Internet or books by oneself.

The former approach is useful because it provides detailed instructions, but it is not always easy to find a suitable teacher. The latter approach is useful because it does not require help from anyone else, but it is not easy to find the optimal solution to fit the individual. To solve this problem, we aim to develop a generative personal assistance system as a means of learning new things. It is advantageous in that it offers individuality and concreteness provided in the former approach above, as well as automaticity provided in the latter approach.

3. Generative personal assistance with audio and visual examples

The concept of our proposed generative personal assistance system providing audio and visual examples is illustrated in **Fig. 1**. We achieve individuality by analyzing the relevant data based on the data

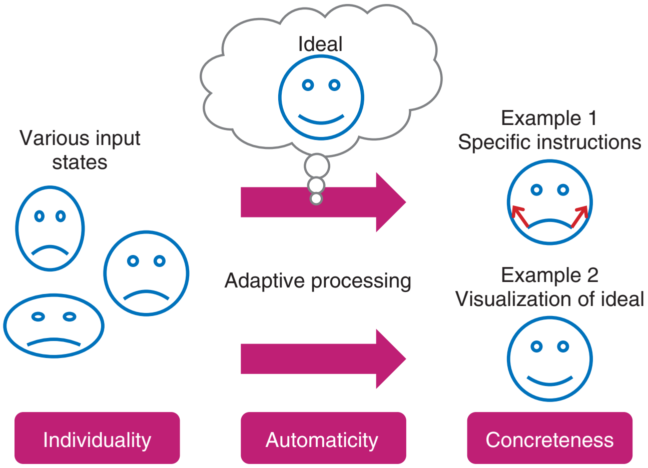

Fig. 1. Generative personal assistance with audio and visual examples (e.g., facial expression improvement).

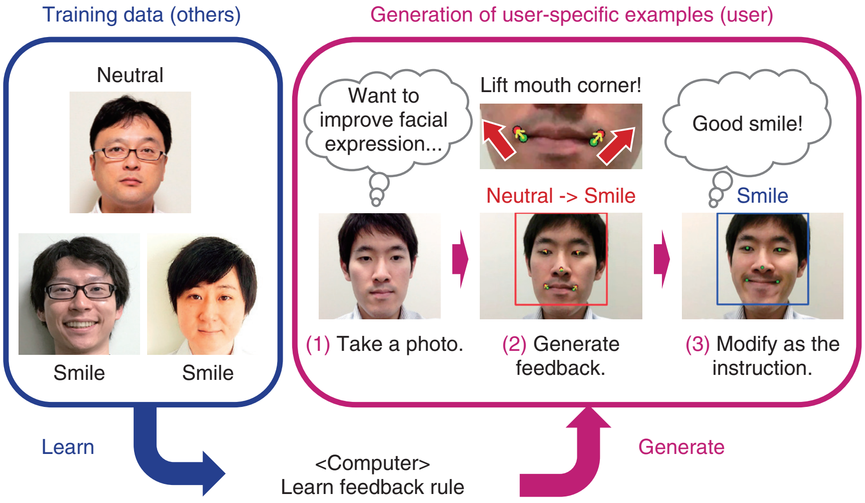

Fig. 2. Operation example of proposed feedback system.

provided by the user (i.e., images and audio). We achieve concreteness by visualizing or auralizing the specific instructions or the ideal state. We achieve automaticity by ensuring the above process is conducted automatically.

An operation example of the system [3] developed to provide visual feedback to someone who wants to improve their facial expression is shown in **Fig. 2**. In this system, we learn the feedback rule using the pre-prepared data. We assume that the data are collected not from the actual user but from the general public. Moreover, we assume that we can only collect one-shot data (i.e., one facial expression) for one person. This condition enables us to achieve automaticity without requiring a large data collection effort.

In generating feedback, we achieve individuality by analyzing and calculating feedback based on the user-provided image. We then display improvement guidelines with arrows. This feedback is concrete and interpretable, enabling the user to achieve the target state by following the instructions. This example targets facial expression improvement, but the same idea can be applied to other tasks (e.g., improving pitching in baseball or improving pronunciation).

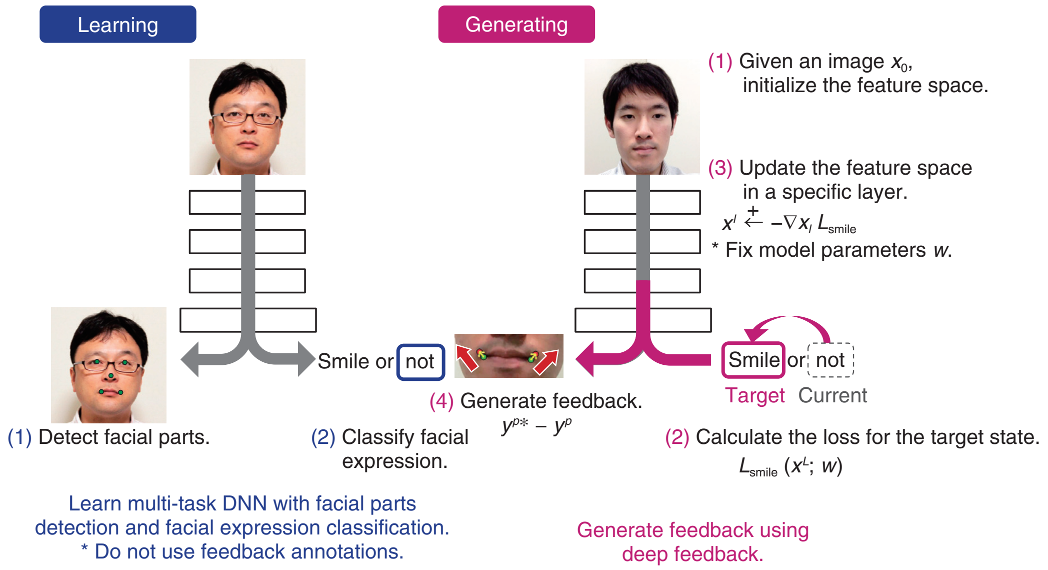
Learning
Generating
(1) Given an image x_0,
initialize the feature space.
(3) Update the feature space
in a specific layer.
$x^l \overset{+}{\leftarrow} -\nabla x_l \, L_{\text{smile}}$
* Fix model parameters w.
Smile or not
(4) Generate feedback.
$y^{p*} - y^p$
Smile or not
Target Current
(1) Detect facial parts.
(2) Classify facial expression.
(2) Calculate the loss for the target state.
$L_{\text{smile}} \, (x^L; w)$
Learn multi-task DNN with facial parts detection and facial expression classification.
* Do not use feedback annotations.
Generate feedback using deep feedback.

Fig. 3. Feedback generation with deep feedback.

4. Input variation and output concreteness

Two challenges need to be overcome in order to achieve this system. The first is how to apply it to various input data. This system is aimed for use by a diverse range of people with different ages, genders, and races. Therefore, it is necessary to adapt the system to provide the optimal feedback for each person. The main approach used in previous feedback systems is rule-based, in which the feedback rule is defined manually. This imposes a large rule-creation cost when there is a large amount of variation data. We propose to solve this problem by using a learning-based approach.

The second challenge is not only to recognize the current state but also to generate detailed feedback based on it. Previous systems also used a learning-based approach that can be applied to various inputs, but they can only recognize the current state; for example, they can only classify facial expressions, estimate the degree of a smile, and detect the different parts of the face. It is not easy to extend this approach to generate feedback because the typical learning scheme requires the pair data of input and output, that is, the user image and correct feedback in this case. In general, feedback annotations require professional knowledge. Therefore, it is not easy to collect a large amount of data. This limitation makes it difficult to develop a feedback system using a learning-based approach.

5. Feedback generation with deep feedback

To overcome these two challenges, we propose a new deep neural network (DNN)-based method that can learn and generate feedback without the annotations of correct feedback. The process flow of this method is shown in **Fig. 3**. In the learning step, we train a multitask DNN in facial parts detection and facial expression classification. This model also requires annotation data—that is, facial expression annotations and facial parts annotations—for training, but they are easier to collect than feedback annotation because they do not require professional knowledge. In the resulting model, the feature space has relevance to input images, facial parts, and facial expression classes that are key factors for generating feedback.

In the generating step, we apply a novel propagation method called deep feedback to extract the feedback information from the feature space. (1) Given an image, the system initializes the feature space and estimates the current position of facial parts and the facial expression class. (2) For the classification output, the system calculates the loss of the target state. (3) The system reduces this loss by optimizing the feature space in a specific layer using back-propagation. In the experiment, we found nonlinear transformation through deep back-propagation is useful for adapting feedback to various inputs. Note that in general training the model parameters are updated while fixing the

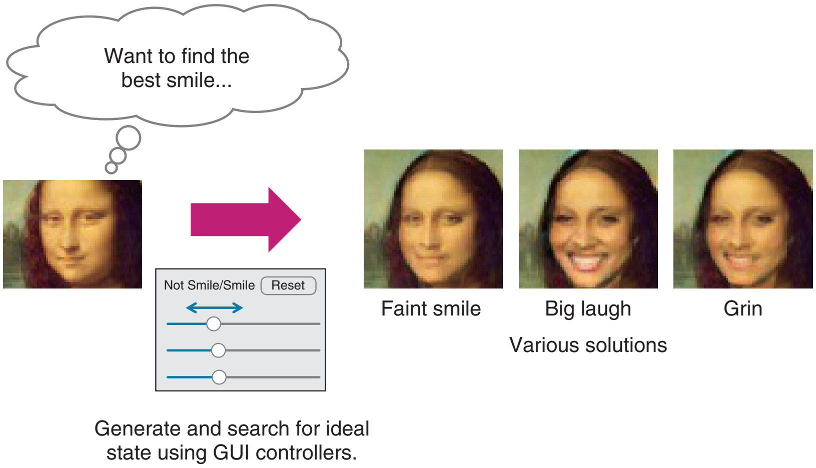
Want to find the best smile...
Not Smile/Smile
Reset
Generate and search for ideal state using GUI controllers.
Faint smile
Big laugh
Grin
Various solutions

Fig. 4. Interactive operations for generating and searching for ideal state.

feature space, but in the deep feedback process the feature space is updated while fixing the model parameters. (4) The optimal facial parts position is calculated based on the optimized feature space using forward propagation. Finally, arrows indicating the difference between the current and optimal facial parts position are displayed.

6. Interactive operations for generating and searching for ideal state

In the above framework, an algorithm finds and visualizes only one solution for smiling, but there are actually various solutions for smiling (e.g., a grin, faint smile, or big laugh). In such a case, it is not easy for a computer to select and generate an optimal one without the user's interaction. We were motivated by this to develop a novel system [4] where a user can generate and search for his/her ideal state interactively using typical GUI (graphical user interface) controllers such as radio buttons and slide bars. A conceptual image of this system is shown in **Fig. 4**.

7. Representation learning using deep attribute controller

To develop this system, we need to obtain a representation space that is disentangled, expressive, and controllable. First, attributes, for example, smiling, and identity need to be disentangled in the representation space in order to change the attributes independently from the identity. Second, the representation space needs to be expressive enough to represent the attribute variations. Third, controllability is important because our goal is to enable a user to intuitively control attributes. One of the challenges in learning such a representation space is to learn the space without a detailed description of attributes. This constraint is important because it is not easy to obtain a detailed description of attributes due to the complexity and difficulty in defining a rule for organizing them.

To solve this problem, we propose a novel method called deep attribute controller to learn disentangled, expressive, and controllable representations only from the binary indicator representing the presence or absence of the attribute. In particular, we propose a conditional filtered generative adversarial network (CFGAN), which is an extension of the generative adversarial network (GAN) [5]. The GAN is a framework for training a generator and discriminator in an adversarial (min-max) process.

The generator maps a noise to data space and is optimized to deceive the discriminator, while the discriminator is optimized to distinguish a generated and real sample and *not to be deceived* by the generator. The CFGAN incorporates a filtering architecture into the generator input, which associates an attribute binary indicator with a multidimensional latent variable, enabling the latent variations of the attribute to be represented. We also define the filtering architecture and training scheme considering controllability, enabling the attribute variations to be intuitively controlled. The CFGAN architecture is shown in **Fig. 5**.

The above explanation describes the case where the CFGAN is applied to learn smile variations, but the same scheme can be used to learn other attribute

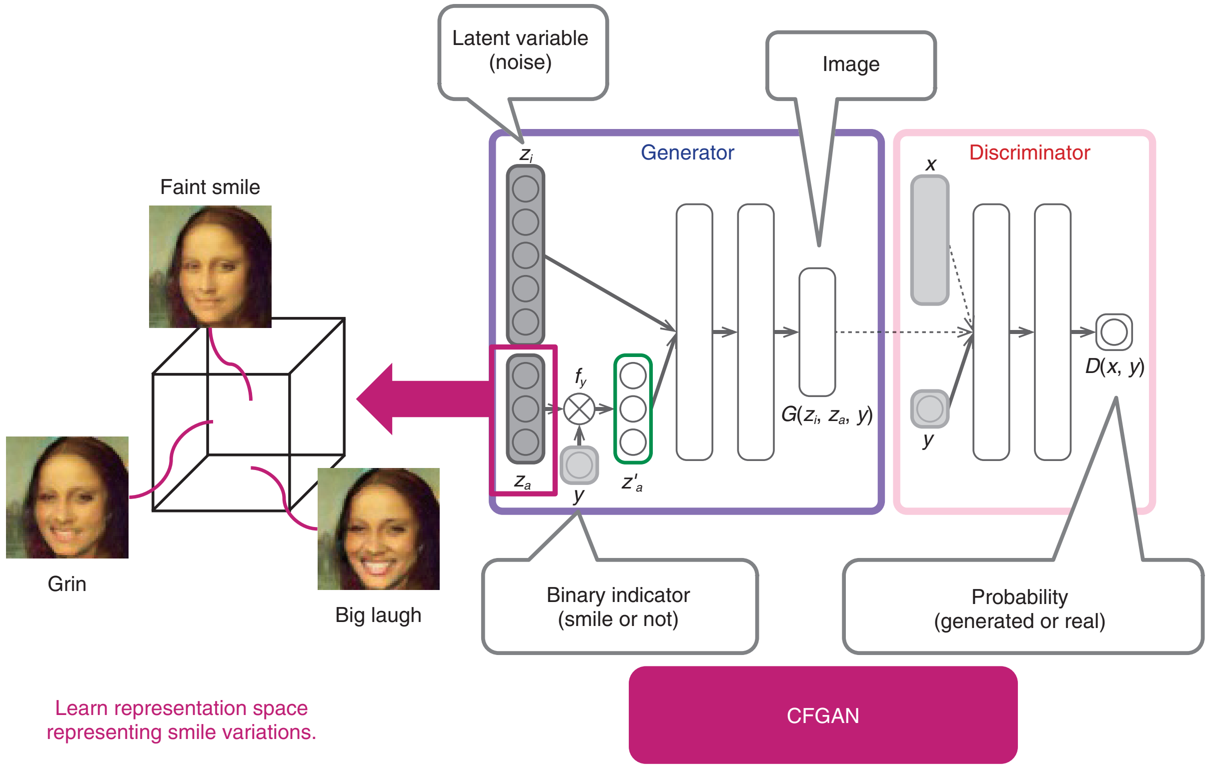

Fig. 5. Representation learning using deep attribute controller.

variations such as ages, hair styles, and genders. Moreover, it can be applied not only to images but also to audio data. For application to audio, we also developed some essential technologies consisting of realistic speech synthesis [6, 7] and realistic voice conversion [8] methods.

8. Future direction

The key objective of our approaches is to use media generation to develop an affinity with—that is, get close to—users. These technologies are essential for not only personal assistance but also for embodying users' wishes through concrete media information. In the future, we aim to establish the technology to generate exceedingly high quality media to meet any expectation. To achieve this, we are working to cultivate the imagination, knowledge, and experience of media generation.

References

[1] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," Proc. of the 16th International Conference on Computer Vision (ICCV 2017), pp. 5907–5915, Venice, Italy, Oct. 2017.

[2] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," Proc. of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), pp. 2414–2423, Las Vegas, NV, USA, June/July 2016.

[3] T. Kaneko, K. Hiramatsu, and K. Kashino, "Adaptive Visual Feedback Generation for Facial Expression Improvement with Multi-task Deep Neural Networks," Proc. of the 24th ACM International Conference on Multimedia (ACMMM 2016), pp. 327–331, Amsterdam, Netherlands, Oct. 2016.

[4] T. Kaneko, K. Hiramatsu, and K. Kashino, "Generative Attribute Controller with Conditional Filtered Generative Adversarial Networks," Proc. of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), pp. 6089–6098, Honolulu, Hawaii, USA, July 2017.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," Proc. of Advances in Neural Information Processing Systems 27 (NIPS 2014), pp. 2672–2680, Montreal, Quebec, Canada, Dec. 2014.

[6] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative Adversarial Network-based Postfilter for Statistical Parametric Speech Synthesis," Proc. of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), pp. 7910–4914, New Orleans, USA, Mar. 2017.

[7] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative Adversarial Network-based Postfilter for STFT Spectrograms," Proc. of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), pp. 3389–3393, Stockholm, Sweden, Aug. 2017.

[8] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence Voice Conversion with Similarity Metric Learning Using

Generative Adversarial Networks," Proc. of 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), pp. 1283–1287, Stockholm, Sweden, Aug. 2017.

Takuhiro Kaneko

Researcher, Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.E. from the University of Tokyo in 2012 and 2014 and began Ph.D. studies at the University of Tokyo in 2017. He joined NTT Communication Science Laboratories in 2014, where he studies computer vision, image processing, speech processing, pattern recognition, and machine learning. His interests include image generation, speech synthesis, and voice conversion using deep generative models. He received the Hatakeyama Award from the Japan Society of Mechanical Engineers in 2012 and the ICPR2012 Best Student Paper Award at the 21st International Conference on Pattern Recognition in 2012. He received the Institute of Electronics, Information and Communication Engineers (IEICE) ISS (Information and Systems Society) Young Researcher's Award in Speech Field in 2017. He is a member of IEICE and the Information Processing Society of Japan (IPSJ).

Kaoru Hiramatsu

Senior Research Scientist, Supervisor, and Leader of Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.S. in electrical engineering and an M.S. in computer science from Keio University, Kanagawa, in 1994 and 1996, and a Ph.D. in informatics from Kyoto University in 2002. He joined NTT Communication Science Laboratories in 1996 and has been working on the Semantic Web, sensor networks, and media search technology. From 2003 to 2004, he was a visiting research scientist at the Maryland Information and Network Dynamics Laboratory, University of Maryland, USA. He is a member of IPSJ and the Japanese Society for Artificial Intelligence (JSAI).

Kunio Kashino

Senior Distinguished Researcher, Head of Media Information Laboratory, NTT Communication Science Laboratories.

He received a Ph.D. from the University of Tokyo in 1995. He has been working on media information processing, media search, and cross-modal scene analysis. He is also an adjunct professor at the Graduate School of Information Science and Technology, the University of Tokyo, and a visiting professor at the National Institute of Informatics. He is a senior member of IEEE (Institute of Electrical and Electronic Engineers) and IEICE, and a member of the Association for Computing Machinery, IPSJ, the Acoustical Society of Japan, and JSAI.