Generating Video from Single Image and Sound

Yukitaka Tsuchiya[1] Takahiro Itazuri[1] Ryota Natsume[1]
Shintaro Yamamoto[1] Takuya Kato[1] Shigeo Morishima[2]
[1]Waseda University [2]Waseda Research Institute for Science and Engineering

Abstract

In this paper, we propose a method of generating a video linked to sound from a single image and a few seconds of sound while maintaining the appearance of the image. Conventional video generation methods from sound require key points extraction related to the sound in each object, such as the mouth in speech and arms in musical instrument performance. They can not be applied to objects whose shape changes significantly like fireworks. The proposed method can generate a video without extracting specific key points from images. We experimented not only the mouth shape and body pose of human treated in the conventional ways, but also fireworks and sea waves where it is difficult to design key points.

1. Introduction

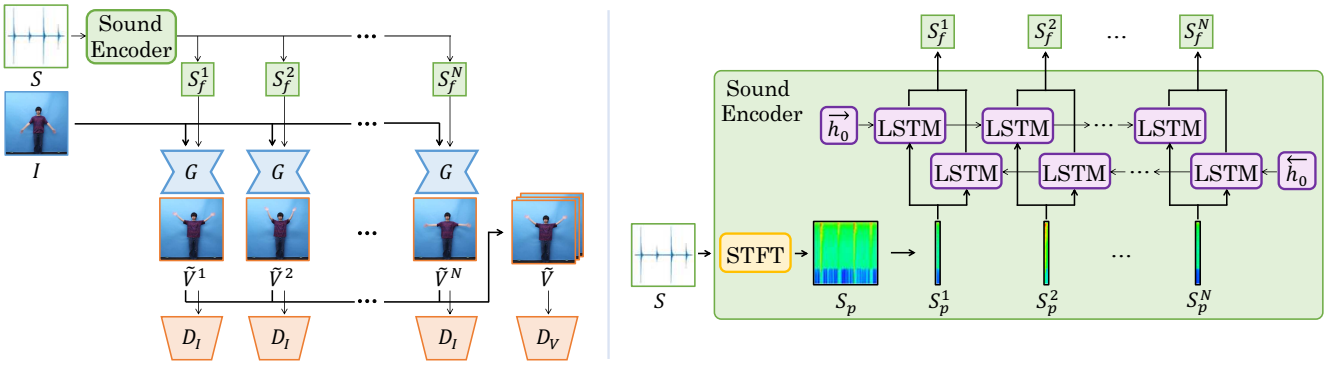
There are many phenomena in which movement and sound are linked to each other, like human speech, musical instrument performance, fireworks, or sea waves. There is a demand for making a video that motion and sound are synchronized. This is because it is recognized in combination by sight and sound. In recent years, focusing on the relationship between sound and motion, methods of generating a video using sound has been proposed. Suwajanakorn et al. [10] proposed a method to generate continuous images with realistic movement of the mouth by estimating the mouth shape from human voice. Shlizerman et al. [8] predicts how bones move from a sound by learning the relationship between instrumental performance and human hand movement. Then they applied predicted bone information to the 3D avatar to generate a video. These methods can be divided into a mechanism for predicting motion from sound and a mechanism for generating an image or 3D avatar for motion by using key points such as mouth shape or bone for correlating sound and motion. Therefore, it is possible to generate realistic videos. When applying these methods to non-human objects, designing key points is inevitable for each target. In addition, it can not be applied to objects whose key points are difficult to design, such as fireworks and sea waves whose shape is greatly deformed.

Some methods [12, 7, 11] have been proposed to generate a video using GAN [2] without key points. In these methods [12, 7], since one latent variable corresponds to the video one by one, the video has a fixed length, and each corresponds to another latent variable even if the speed is different with the same motion. Tulyakov et al. [11] proposed a method for generating a video by separating video component into the motion and the content. By separating the latent variable space by the content and the motion, it is possible to fix the motion and change the content, and vice versa. Although GAN-based methods can generate a continuous video without key points, the generated video does not contain sound, as sound information is not considered.

In this paper, we propose a GAN-based video generation method for general objects that move in conjunction with sound. Our proposed method generates a video in which the image region corresponding to the input sound moves according to the sound. The feature of the movable part corresponding to the sound is learned, by making the feature of temporal change of sound correspond to the image for each frame. Correlating not the sound but the temporal change of the sound with the image enables to generate a video for the phenomenon with many silent sections such as pulse like sound. In addition, learning the appearance of the image without defining key points specific to the target makes it possible to create a video that retains the appearance of the input image, even for objects whose key points design is difficult.

2. Proposed Method

We introduce a method uses GAN to generate videos by using visual information obtained from a single image and information on temporal change in sound obtained from a few seconds of sound. As shown in Figure 1, the proposed network consists of Sound Encoder (SE), Generator (G), and Discriminator (D_I, D_V). First, SE extracts the feature of temporal change of sound with the sampling rate of the frames. Then, G generates an image for each frame from the obtained features from the SE and the input image. D_I discriminates the spatial naturalness between the generated image and the real image. Finally, D_V discriminates whether the temporal change of the generated continuous frames is natural. The proposed method does not require feature points specified to certain object, which al-



(a) Outline of the proposed network (b) Sound Encoder

Figure 1. Proposed Network.

lows sound and target in general to be applicable.

2.1. Sound Feature Extraction

In order to generate a video synchronized with sound, it is necessary to associate sound with motions. In the proposed method, sound information is associated with each image by Fourier transformation of the sound waveform according to the sampling rate of the frames. For videos with a sound sampling rate of F_S [Hz] and an image sampling rate of F_I [fps], short-time Fourier transform (STFT) is performed for F_I/F_S [s], and we get the spectrogram ($S_p = \{S_p^t\}_{t=1,\dots,N}$). The spectrogram represents the intensity of each frequency of sound over time. The obtained S_p contains a silent interval, and the spectrogram in the silent interval does not contain necessary information to predict motion. Therefore, it is not appropriate to associate S_p directly with the images. In this paper, it is possible to extract information for predicting motion even in silent sections, by focusing on the temporal change of the spectrogram. The feature ($\{S_f^t\}_{t=1,\dots,N}$) for temporal change of sound is defined for each frame, by inputting the obtained S_p^t into Bidirectional LSTM (BLSTM) [3] for each F_I/F_S [s]. We use BLSTM as SE because the entire input sound is given in advance. BLSTM is a network that propagates information bidirectionally from the past to the future and from the future to the past. By using information in both directions, it becomes possible to extract the feature of movement until the sound is emitted, not only for continuous sound but also for single sound such as pulse like sound. The initial values of hidden state of BLSTM $\overrightarrow{h_0}$ and $\overleftarrow{h_0}$ are set with 0.

2.2. Image Generation

An image $\tilde{V}^t$ is generated by G from a single input image I and a feature S_f^t. We use U-Net [6] for G to generate a video that retains the appearance of the input image. In order to generate natural video, spatial naturalness of each image and temporal naturalness in continuous images are important. U-Net can generate an image including spatial information of the input image by referring to the information in the middle layer during encoding and decoding. The generated images are discriminated by two D as Tulyakov et al. [11] proposed. D_I discriminates one image and D_V discriminates the naturalness of multiple images. Therefore, it can be divided into D_I specializing in space and D_V specializing in time. By using two D, the naturalness of static image feature and the naturalness of motion feature can be determined separately, and then more realistic video can be generated.

2.3. Training

In this paper, we define the following loss function $\mathcal{L}$

$$\mathcal{L} = \mathcal{L}_{GAN} + \mathcal{L}_{rec} \tag{1}$$

G and SE are trained to minimize $\mathcal{L}$. D_I and D_V are trained to maximize $\mathcal{L}$. $\mathcal{L}_{GAN}$ is defined by the following equation.

$$\mathcal{L}_{GAN} = \lambda_{adv\text{-}I}\mathcal{L}_{adv\text{-}I} + \lambda_{adv\text{-}V}\mathcal{L}_{adv\text{-}V} \tag{2}$$

$\lambda_{adv\text{-}I}$ and $\lambda_{adv\text{-}V}$ are the weights of how D_I and D_V are considered. $\mathcal{L}_{adv\text{-}I}$ and $\mathcal{L}_{adv\text{-}V}$ are defined by following equations.

$$\begin{aligned}\mathcal{L}_{adv\text{-}I} = {} & \mathbb{E}_{V\sim p_{data}}\left[\log D_I(\Gamma_1(V))\right] \\ & + \mathbb{E}_{\tilde{V}\sim p_{\tilde{V}}}\left[\log(1 - D_I(\Gamma_1(\tilde{V})))\right]\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{adv\text{-}V} = {} & \mathbb{E}_{V\sim p_{data}}\left[\log D_V(\Gamma_T(V))\right] \\ & + \mathbb{E}_{\tilde{V}\sim p_{\tilde{V}}}\left[\log(1 - D_V(\Gamma_T(\tilde{V})))\right]\end{aligned} \tag{4}$$

V means a video in the dataset, $\tilde{V}$ means the generated video, and $\Gamma_T(\cdot)$ represents T images of continuous images

($\Gamma_1(\cdot)$ means one image of continuous images). $\mathcal{L}_{rec}$ is defined by the following formula. The reconstruction error is measured at $L1$ loss for all frames of the generated video and the real video.

$$\mathcal{L}_{rec} = \|V - \tilde{V}\|_1 \tag{5}$$

3. Experiment

To show the effectiveness of our method, two factors are required to verified. One is if it is applicable for the target used in conventional method using key points. The other is if it is applicable for the target in which the key points are difficult to design. Therefore, we selected human mouth and hands that can acquire key points, and fireworks and sea waves that cannot define key points. In this experiment, four types of targets are trained separately.

3.1. Dataset

The existing video datasets [9, 1, 4] contain sounds that are not related to motion, so it is not suitable for verifying the effectiveness of the proposed method. Therefore, we construct the dataset of videos with correspondence between sound and motion. There are three conditions for data.

- The sound and the image are corresponding each other.
- There is no sound unrelated to the movement.
- The camera is fixed.

We shot videos in which five people pronounced the vowels "a", "i", "u", "e", and "o", and videos of seven people clapping their hands above their heads as they moved their hands up and down. We collected fireworks and sea videos published on YouTube. The sound sampling rate is $F_S = 44,100$ [Hz], and the frame sampling rate is $F_I = 30$ [fps]. We created data for each video by the following procedure. First, we randomly cut out four seconds from a video. Next, videos are separated into 120 frames and four seconds of sound. After cropping the image to be square, it is resized to 286×286. The total number of seconds and frames for each video of the constructed dataset is shown in the Table 1.

3.2. Experiment setting

We select 100 continuous images from 120 continuous images of the dataset for data augmentation. Then we randomly crop an area of image size 286×286 to 256×256, then resize to 64×64. SE extracts the characteristics of temporal change by inputting four seconds of sound. After that, learning is performed on the part corresponding to 100 continuous images. In this paper, we generated a video with image size 64×64, frame sampling rate 29.97 [fps], sound sampling rate $44,100$ [Hz]. The learning was conducted on a computer with an Intel Core i7-5930K 3.50 GHz CPU, NVIDIA GTX TITAN X GPU, and 64 GB of RAM. During training, the weight of the loss function is $\lambda_{adv\text{-}I} = 1.0$, $\lambda_{adv\text{-}V} = 1.0$, respectively. We used ADAM [5] for optimization, set the initial learning rate to 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$. The batch size is two.

Table 1. Details of the datasets used in our experiment.

	Mouth	Hand	Fireworks	Sea
Seconds	912	2,352	960	5,280
Frames	109,440	282,240	115,200	633,600

4. Results and Discussion

The Figure 2 shows the results of the video of human mouth, human hands, fireworks and sea waves. In this paper, the amplitude of the sound is normalized to $[-1.0, +1.0]$. The input image is a green frame. The generated video is shown every 20 frames with orange frame. The original video image corresponding to each frame is a red frame. The input sound is recorded in the original video. As shown in Figure 2-(a), the generated mouth images look similar to that of the original video. Although the method of Suwajanakorn et al. [10] requires key points to generate a video from speech, the proposed method archived to generate a video of the mouth moving without designing key points. In addition, it is possible to generate images that reflect the appearance of the face and background of the input image as well as the mouth part. We confirm that the proposed method can be applied to small movements of the mouth, which is an area that emits sound. As shown in Figure 2-(b), the arms are generated to correspond to the motion of the original video, but the yellow color of the input image can not be maintained. It is considered that the created dataset did not contain enough clothes colors for learning. Unlike the human mouth or hands, we use fireworks and sea waves as examples of targets for which it is difficult to design key points. Fireworks is a partial change of the image, whereas the sea is a whole object of the change. The result of fireworks (Figure 2-(c)) shows that the generated fireworks light is delayed by 20 to 40 frames compared to the original video. Not only the brightness but also the shape of the fireworks changes. From the result of the sea (Figure 2-(d)), it can be seen that the proposed method is generating white waves in the same section ($t \in [20, 80]$) as the original video. The generated video shows that the white waves contained in the input image disappear as the sound gets smaller ($t \geq 100$).

5. Conclusion

In this paper, we proposed a method to generate a video from a single image and a few seconds of sound without key

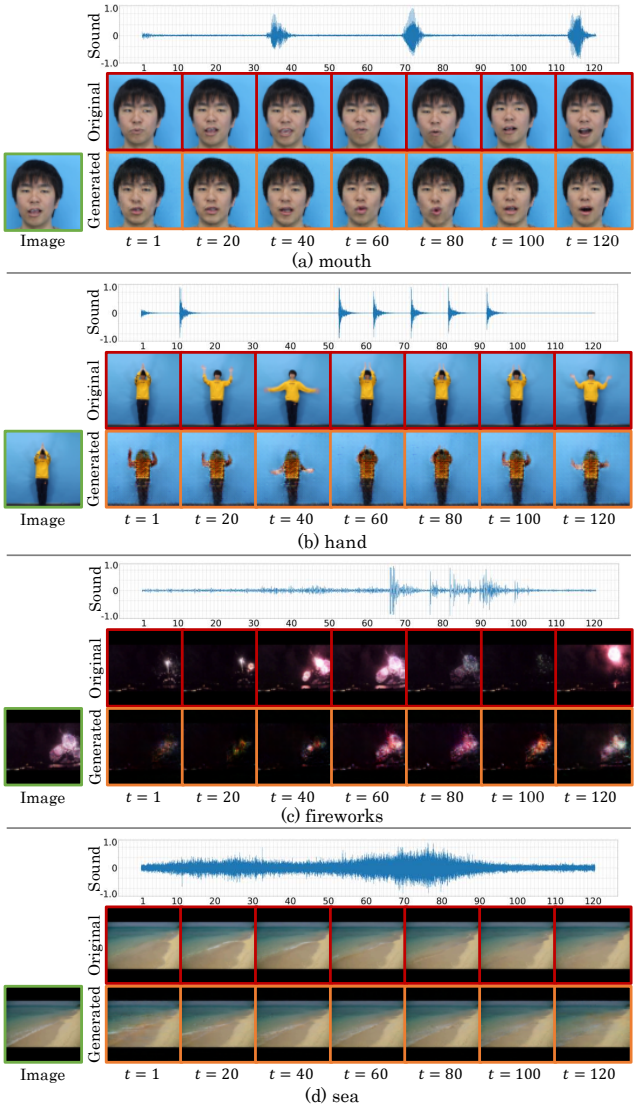


Figure 2. Comparison of generated video and original video for input image and input sound.

points. Since it is not necessary to define key points for each object, learning can be performed for each object simply by changing the dataset. In order to verify the effectiveness of our proposed method, experiments were conducted on four types of objects. Our experimental results show that it is difficult to generate a realistic video when the moving part according to the sound is large or when the motion is not uniquely determined for the sound. Extension to the case where the object moves largely from the input image is a future task. In the proposed method, a frame is generated one by one from the corresponding feature of temporal change of sound and the input image. In this case, the relationship between the generated image and the previous image can not be taken into consideration. We consider that referring recursively to the previous frame information makes it robust.

Acknowledgements

This work is supported by the JST ACCEL Grant Number JPMJAC1602, JSPS KAKENHI Grant Numbers JP17H06101 and JP19H01129.

References

[1] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Neural Information Processing Systems (NIPS)*. 2016. 3

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Neural Information Processing Systems (NIPS)*, 2014. 1

[3] A. Graves, A. rahman Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. *Computing Research Repository*, 2013. 2

[4] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *Computing Research Repository*, 2017. 3

[5] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computing Research Repository*, 2014. 3

[6] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 2

[7] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[8] E. Shlizerman, L. M. Dery, H. Schoen, and I. Kemelmacher-Shlizerman. Audio to body dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[9] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *Computing Research Repository*, 2012. 3

[10] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 2017. 1, 3

[11] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[12] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Neural Information Processing Systems (NIPS)*. 2016. 1