# Audio-guided Video Interpolation via Human Pose Features

Takayuki Nakatsuka[1,2][a], Masatoshi Hamanaka[2] and Shigeo Morishima[3]
[1] *Waseda University, Japan*
[2] *RIKEN, Japan*
[3] *Waseda Research Institute for Science and Engineering, Japan*

Keywords: Video Interpolation, Pose Estimation, Signal Processing, Generative Adversarial Network, Gated Recurrent Unit.

Abstract: This paper describes a method that generates in-between frames of two videos of a musical instrument being played. While image generation achieves a successful outcome in recent years, there is ample scope for improvement in video generation. The keys to improving the quality of video generation are the high resolution and temporal coherence of videos. We solved these requirements by using not only visual information but also aural information. The critical point of our method is using two-dimensional pose features to generate high-resolution in-between frames from the input audio. We constructed a deep neural network with a recurrent structure for inferring pose features from the input audio and an encoder-decoder network for padding and generating video frames using pose features. Our method, moreover, adopted a fusion approach of generating, padding, and retrieving video frames to improve the output video. Pose features played an essential role in both end-to-end training with a differentiable property and combining a generating, padding, and retrieving approach. We conducted a user study and confirmed that the proposed method is effective in generating interpolated videos.

## 1 INTRODUCTION

Composing music, like any creative work, is a continuous process of trial and error to pursue the desired music. Like Wolfgang Amadeus Mozart, who is a famous composer, said that "Music should never be painful to the ear but should flatter and charm it, and thereby always remain music," crafted music should be well-thought-out and enticing. We have expanded the Generative Theory of Tonal Music (GTTM) (Lerdahl and Jackendoff, 1996) to deepGTTM (Hamanaka et al., 2016) (Hamanaka et al., 2017) that enabled a time-span tree of a melody to be automatically acquired based on the GTTM. Besides, we developed an interactive music system called the "Melody Slot Machine (Hamanaka et al., 2019)," which is an interactive music system that provides an experience of manipulating and controlling a music performance to musical novices. The melodies used in the system are divided into multiple segments, and each segment has multiple variations of melodies, from intense ones with many notes to calm ones with few notes, so that users can explore the melody they want. We prepared an AR display showing a performer so that the result of the operation can be visually understood as well as aurally.

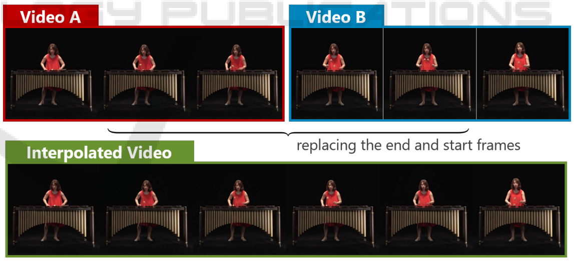

Figure 1: Interpolated Video as the result of the interpolation between Video A and Video B. To fill in the gaps between the two videos, we replace the end frames of the first video and the start frames of the second video with interpolated frames. The length of interpolation is changed by pose similarity between the two videos.

While the strong background of the music theory and the adequate quality of generated music, the video, which is accompanied to the music, lowers the satisfaction of playing the system. This is because merely joining an audio recording with a video does not produce a pleasant result. In the case of a discontinuous audio signal, to remove some noises (e.g.,

[a] https://orcid.org/0000-0003-3181-4894

Nakatsuka, T., Hamanaka, M. and Morishima, S.
Audio-guided Video Interpolation via Human Pose Features.
DOI: 10.5220/0008876600270035
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 27-35
ISBN: 978-989-758-402-2; ISSN: 2184-4321

clip noises, crackle noise) will improve the quality of sounds reasonably. Editing the joint parts of videos, however, is still challenging. The main problems lie in video resolution and temporal coherence. To improve the experience of the interaction, we addressed these problems by generating frames and interpolating videos using aural information via human pose features.

In this study, we propose an audio-guided video interpolation method for videos of a musical instrument being played. We adopted two-dimensional pose features as the intermediate representation that bridges aural and visual information in our workflow for efficient and effective video interpolation. The subjective evaluation shows that our interpolated videos appear as natural as an original video and satisfy the participants. Our primary contributions are the following:

- A new method of using audio to generate an interpolated video.
- A novel algorithm that adopts two-dimensional pose features to make in-between frames natural from the viewpoint of resolution and temporal coherence of videos.

# 2 RELATED WORK

## 2.1 Audio-driven Animation

**Facial Animation.** Several studies succeeded in inferring facial animation from speech audio. Suwajanakorn *et al.* synthesized lip synchronized videos using a recurrent neural network that learns the mouth shapes from raw audio features (Suwajanakorn et al., 2017). Karras *et al.* proposed a deep neural network that learns 3D vertex coordinates of a face model with its expression from raw audio (Karras et al., 2017). They applied the formant analysis to input speech audio to extract audio features. Visemes are also useful to estimate the facial animation from speech (Zhou et al., 2018). Cudeiro *et al.* used revised DeepSpeech (Hannun et al., 2014) features and implemented animator controls that preserve identity-dependent facial shape and pose (Cudeiro et al., 2019).

**Motion Synthesis.** Synthesizing motion from input audio has been the subject of various research. Fan *et al.* and Ofli *et al.* synthesized dance motion from music using a statistical and example-based approach (Fan et al., 2012) (Ofli et al., 2012). Shlizerman *et al.* generated the motion of playing an instrument using a Long Short-Term Memory (LSTM) network (Shlizerman et al., 2018).

## 2.2 Video and Image Processing

**Video Interpolation.** Recently, interpolation for a short interval between two subsequent frames in high resolution has been successfully performed using deep learning (Niklaus et al., 2017) (Niklaus and Liu, 2018) (Meyer et al., 2018) (Jiang et al., 2018). These methods help create high frame rate videos from ordinary ones. However, they are not useful in interpolating a long interval of frames because of the low correspondence between the frames that often leads to low-fidelity of generated videos. Chen *et al.* and Xu *et al.* introduced a bidirectional predictive network for generating in-between frames for an extended interval (Chen et al., 2017) (Xu et al., 2018). Bidirectional constraint propagation, which ensures the spatial-temporal coherence between frames, succeeds in generating plausible videos. Denton *et al.* proposed a video generation model with a learned prior that represents the latent variable of the future video frames (Denton and Fergus, 2018). Li *et al.* suggested a deep neural network with three-dimensional convolutions, which guarantee the spatial-temporal coherence among frames, for generating in-between frames directly in the pixel domain (Li et al., 2019). Wang *et al.* presented a skip-frame training strategy that enhances the inference model to learn the time counter implicitly (Wang et al., 2019). These methods, however, still lacked in terms of video resolution (the generated video resolution was 64 [px] × 64 [px]).

**Conditional Image Generation.** Conditional image generation is the task of generating new images from a dataset by setting specific conditions. In this task, a generative adversarial network (GAN) can improve the fidelity of the generated images. Wang *et al.* and Brock *et al.* proposed GANs that generated high-fidelity images from sparse labels and annotations (Wang et al., 2018b) (Brock et al., 2019). Wang *et al.* expanded an image-to-image synthesis approach to video-to-video one using the generative adversarial learning framework with a Spatio-temporal adversarial objective function (Wang et al., 2018a). Pumarola *et al.* implemented facial animation from a single image-conditioning scheme based on action-units annotations (Pumarola et al., 2018).

We propose a fusion approach of video frame generation, padding, and retrieval to achieve long-interval interpolation and produce in-between frames with high resolution.

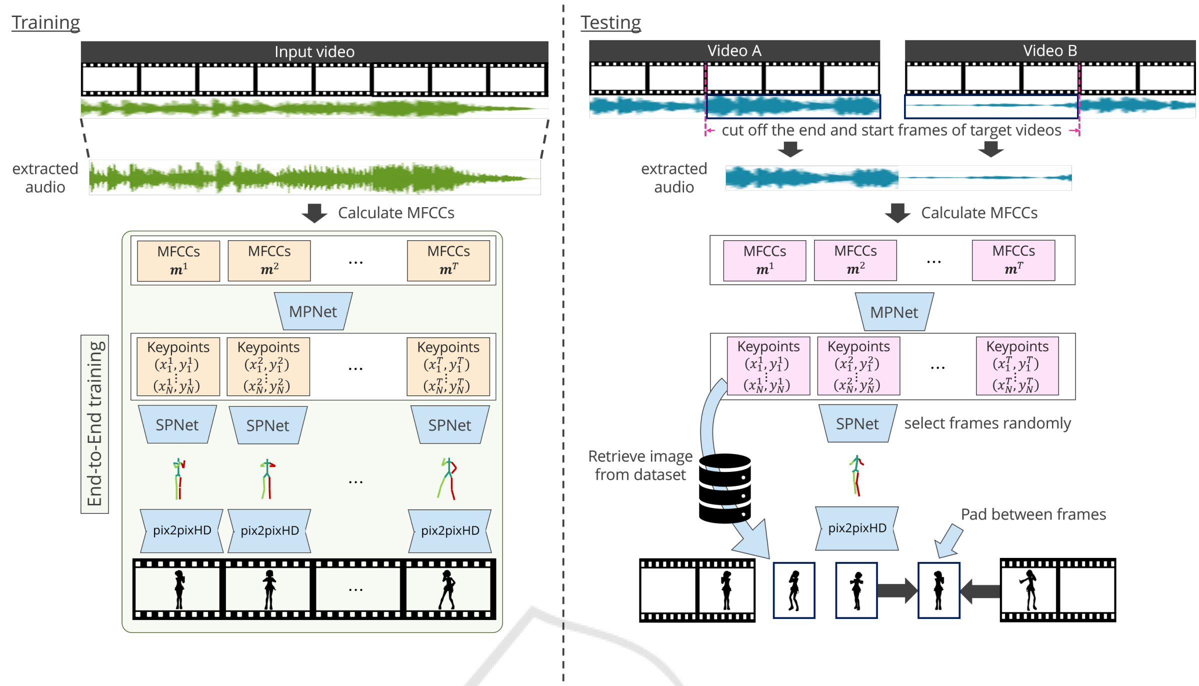

Figure 2: Overview of our proposal. (left) **Training:** Our model estimates pose features from the input audio and generates a video from pose stick figures. To fill in the gap between the pose features and the pose stick figures, we propose a differentiable line drawing from the points. This technique enables a deep neural network to train the model by end-to-end learning. (right) **Testing:** We adopt a combined approach to generate an interpolated video. The trained model is used for the pose features estimation from the audio and video frame generation from the pose stick figure in the interpolation section.

## 3 METHOD

Our goal is to interpolate two target performance videos using audio features. To tackle this challenge, we divided our workflow into three stages for the training: (i) audio feature extraction from the input audio, (ii) pose estimation from the audio feature, and (iii) video interpolation between the target videos by generating in-between frames. Our framework is trained in an end-to-end manner. Using the trained model, we introduce a novel algorithm that interpolates the two target videos with high resolution. The significant point is that our method combines generating, retrieving, and padding methods to produce a seamless and natural video.

### 3.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficient (MFCC) is one of the most popular methods to extract audio features from a signal waveform. MFCC is widely used for not only speech analysis but also music analysis (Foote, 1997) (Logan and Chu, 2000) (Logan et al., 2000). In this paper, we used 20-dimensions of MFCC calculated from the input audio. The parameters to calculate MFCCs are: the sampling rate is 44.1 [kHz], the window size is 2,048, the hop size (window overlaps) is 512, and the type of a window is the hann window. The hann window is described by the following equation:

$$w(x) = 0.5 - 0.5\cos 2\pi x, \text{ if } 0 \leq x \leq 1. \qquad (1)$$

To fit the number of elements to frame per second of videos, a linear interpolation (Lerp) is applied to MFCCs. Finally, we obtained 20-dimensions of MFCC per video frame.

### 3.2 Pose Features from MFCCs

Let $MP : \mathbf{m} \rightarrow \mathbf{z}$ be the function that learns a mapping from MFCCs $\mathbf{m} \in \mathbb{R}^{\mathbf{T}\times\mathbf{20}}$ to pose features $\mathbf{z} \in \mathbb{R}^{\mathbf{T}\times\mathbf{25}\times\mathbf{2}}$, where the $T$ is the length of the input MFCCs, and the pose features $\mathbf{z}$ were described as 25 joint positions (50 dimensions) on each video frame. For the sequence-to-sequence transfer, we developed a deep neural network with a recurrent structure using three layers of the bidirectional Gated Recurrent Unit (BiGRU) (Chung et al., 2014) to learn pose features $\mathbf{z}$ from the MFCCs $\mathbf{m}$. We also added two weights sharing fully connected layers to all BiGRU outputs. Fig.

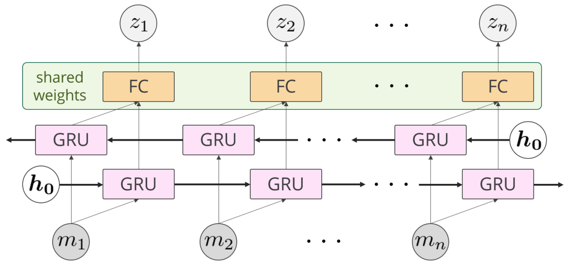

Figure 3: The architecture of our network for pose estimation from MFCCs. The $m_i$ is the $i$-th frame of MFCCs (input) and the $z_i$ is the $i$-th frame of estimated pose features (output).

3 shows the schematic architecture of our network. The objective function $\mathcal{L}_{pose}$ to train our network is following:

$$\mathcal{L}_{pose} = \frac{1}{T}\sum_{i}^{T} \|z_i - MP(m_i)\|_2 \tag{2}$$

The network parameters that we used were the followings: the input length of MFCCs $T$ was set at 50, and the hidden state dimension of BiGRU was set at 512. Fully connected layers consisted of two layers, in which hidden states of layers were 1024 and 512 dimensions, respectively. We used the Adam optimizer (Kingma and Ba, 2014) to train the network. The learning rate was 2.5e-4, $\beta_1$ was 0.9, $\beta_2$ was 0.999, and $\varepsilon$ was 1.0e-8.

## 3.3 Stick Figure from Pose Features

To apply an image-to-image transfer method, which enables a deep neural network to generate a natural image from a sparse annotation, pose features have to be rendered to an image. Let $SP : \mathbf{z} \rightarrow s$ be the function that maps from pose features $\mathbf{z}$ to stick figure $s$. We introduced a differentiable line drawing to realize end-to-end learning from audio to images. Let $A$ and $B$ be joint positions. Problem setting involves drawing line segment $\overline{AB}$ between points A and B. Let $I_0$ be the image matrix where all pixels have the value one. To draw a line, we calculated the distance $d$ between any point $P$ and line $AB$ using the following:

$$d = \frac{|(B-A)\times(P-A)|}{\|B-A\|}. \tag{3}$$

We then obtain line $AB$ using the following:

$$AB = I_0 \times \left(1+\frac{d^2}{\nu}\right)^{\left(-\frac{\nu+1}{2}\right)}. \tag{4}$$

In this paper, we used $\nu = 1$. To cut off line $AB$ at points $A$ and $B$, we calculated the angle $\theta$ between line segments $\overline{AB}$ and $\overline{AQ}$ for any point $Q$, and the angle $\theta'$ between line segments $\overline{BA}$ and $\overline{BQ'}$ for any point $Q'$ using the followings:

$$\cos\theta = \frac{(B-A)\cdot(Q-A)}{\|B-A\|\|Q-A\|}, \tag{5}$$

$$\cos\theta' = \frac{(A-B)\cdot(Q'-B)}{\|A-B\|\|Q'-B\|}. \tag{6}$$

We finally obtained line segment $\overline{AB}$, which is rendered on image $I_0$ with differentiable properties, using the following:

$$\begin{aligned}&\overline{AB}\\ &= AB \times \left(1+e^{-\alpha\cos\theta}\right)^{-1} \times \left(1+e^{-\alpha\cos\theta'}\right)^{-1}.\end{aligned} \tag{7}$$

Fig. 4 shows a concrete example of Eq. (7). We used $\alpha = 5$ to make line segment $\overline{AB}$ sharp. A differentiable line drawing with colors was used to render the stick figure $s$ from pose features $\mathbf{z}$, as shown in Fig. 5.

## 3.4 Video Generation

We employed (Wang et al., 2018b) for the image-to-image transfer. In this study, the input image was the stick figure $s \in \mathbb{R}^{3\times H\times W}$, which was rendered from pose features $\mathbf{z}$ (Section 3.3), and the output was a natural image of a person $x \in \mathbb{R}^{3\times H\times W}$, where the $H, W$ are the height and width of an image, respectively. Let $G : s \rightarrow x$ be the generator, and $D$ be the discriminator. The objective functions for training the network are followings:

$$\begin{aligned}&\mathcal{L}_{image}\\ &= \min_{G}\left(\max_{D_k}\sum_{k=1}^{3}\mathcal{L}_{GAN}(G,D_k) + \lambda\sum_{k=1}^{3}\mathcal{L}_{FM}(G,D_k)\right),\end{aligned} \tag{8}$$

$$\begin{aligned}&\mathcal{L}_{GAN}(G,D)\\ &= \mathbb{E}_{(\mathbf{s},\mathbf{x})}[\log D(\mathbf{s},\mathbf{x})] + \mathbb{E}_{\mathbf{s}}[\log(1-D(\mathbf{s},G(\mathbf{s})))],\end{aligned} \tag{9}$$

$$\begin{aligned}&\mathcal{L}_{FM}(G,D_k)\\ &= \mathbb{E}_{(\mathbf{s},\mathbf{x})}\sum_{i=1}^{L}\frac{1}{E_i}[\|D_k^{(i)}(\mathbf{s},\mathbf{x}) - D_k^{(i)}(\mathbf{s},G(\mathbf{s}))\|_1],\end{aligned} \tag{10}$$

where $\lambda$ is the weight, $D_k^{(i)}$ is the $i$-th layer feature extractor of multi-scale discriminator $D_k$, $L$ is the total number of layers, and $E_i$ indicates the number of elements in each layer. The network was fine-tuned with the target videos of a musical instrument being played. We used the Adam optimizer (Kingma and Ba, 2014) to fine-tune the network. The learning rate was 2.0e-4, $\beta_1$ was 0.9, $\beta_2$ was 0.999, and $\varepsilon$ was 1.0e-8.

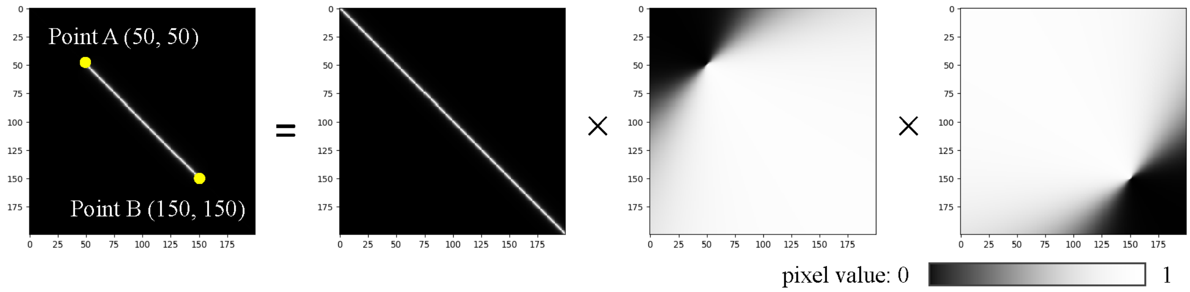

Figure 4: An example of a differentiable line drawing. Let $A = (50,50)$ and $B = (150,150)$ be points. (a) shows line segment $\overline{AB}$ which we want to calculate. (b) shows line $AB$ calculated by Eq. (4). (c) and (d) indicate endpoints of line $AB$ calculated by the second and third term on the right hand of the Eq. (7), respectively.

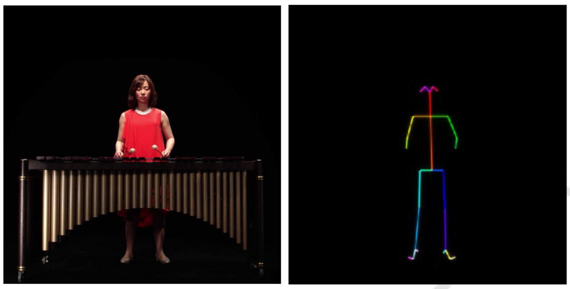

Figure 5: **(right)** An original image. **(left)** An example of the stick figure. We used these images as the pair data to train the network which learns the image-to-image transfer.

## 3.5 Video Interpolation

We proposed a novel algorithm that combined approaches to generating, padding, and retrieving images for interpolating the target videos with high resolution while testing. Algorithm. 1 shows the order of processing. Pose similarity $S_p$ between two target videos is calculated using the Object Keypoint Similarity (OKS) (Ruggero Ronchi and Perona, 2017) for several end frames of the first video and start frames of the second video, and defined as following:

$$S_p = \sum_i^F w_i \mathrm{OKS}_i \Big/ \sum_i^F 1, \qquad (11)$$

where $F$ is the window size, $w$ is the weight. We used $F = 10$ for the experiments, and $\mathbf{w} \in \mathbb{R}^{10}$ follows the Gaussian distribution. The interpolation length $L_p$ depends on pose similarity $S_p$. The higher the pose similarity is, the shorter the interpolation length is. The interpolation length is empirically defined in three lengths (9, 13, and 17 frames) by sorting pose similarity. Then, we connected the sequence of pose features in the target videos and that of generated ones by Lerp.

To generate natural images, we adopted both the frame padding method (Niklaus et al., 2017) and the image-to-image transfer method (Wang et al., 2018b). We retrieved images from personal performance videos, where the sequence of original pose features was highly matched to that of generated ones. On the other hand, we used the image-to-image transfer method (Wang et al., 2018b), where the sequence of generated pose features was not included in the performance videos. To fill the gaps in the original and retrieved/generated images, we applied video frame padding method (Niklaus et al., 2017) to them, and connected the frames seamlessly.

Algorithm 1: Pseudo code for video interpolation (Fig. 2: bottom). While testing, Original Pose features $\mathbf{Z}_{pose} \in \mathbb{R}^{\mathrm{DATA}\times 25\times 2}$, Original Images $\mathbf{X} \in \mathbb{R}^{\mathrm{DATA}\times 3\times H\times W}$, which were used in the training, were utilized for video interpolation. *DATA* denotes the size of a dataset.

```
Input: Audio segment A
Output: in-between frames $\mathbf{X}' \in \mathbb{R}^{F\times C\times H\times W}$
  Calculate MFCCs $\mathbf{m} \in \mathbb{R}^{F\times 20}$ from Audio segment A (Section 3.1)
  Calculate pose features $\mathbb{R}^{F\times 25\times 2} \ni \mathbf{z} \Leftarrow \mathrm{MP}(\mathbf{m})$ (Section 3.2)
  Calculate pose similarity $S_p$ (Section 3.5)
  Determine the length of the interpolation length $L_p$ (Section 3.5)
  Calculate Lerp of pose features $\overline{\mathbf{z}}$ (Section 3.5)
  Randomly sample pose features $\overline{\mathbf{z}'} \subset \overline{\mathbf{z}}$ (Section 3.5)
  if $|\mathbf{Z}_{pose} \ni z_{pose} - z' \in \overline{\mathbf{z}'}| <$ threshold then
    Retrieve frame $\mathbf{X}' \ni x' \Leftarrow \mathbf{X}(z_{pose})$ (Section 3.5)
  else
    Generate frame $\mathbf{X}' \ni x' \Leftarrow G(SP(z))$ (Wang et al., 2018b) (Section 3.3, 3.4)
  end if
  for $\overline{\mathbf{z}''} \subset \overline{\mathbf{z}} \setminus \overline{\mathbf{z}'}$ do
    Pad frames $\mathbf{X}' \ni x' \Leftarrow \mathrm{Padding}(\mathbf{X},\mathbf{X}')$ (Niklaus et al., 2017) (Section 3.5)
  end for
  in-between frames $\mathbf{X}'$
```

## 4 DATA PREPROCESSING

For the experiments, we prepared personal performance videos (audio sampling rate: 44100 [Hz], video definition: full HD, frame rate: 30 [frames/s]). The scores played by the performer were generated by the melody morphing (Hamanaka et al., 2017) of two different scores (Horn Concerto No. 1/Mozart and La Gioconda: Dance of the Hours/Ponchielli). We prepared eleven scores, including two original scores and nine morphed scores, for the experiments. Fig. 6 shows a part of scores that we used in the experiments. Videos are trimmed around the target performer and re-sized to 480 [px] × 480 [px]. The length of the videos was 374 [sec] in total. We then calculated 20 dimensions of MFCCs from the audio corresponding to each video frame. To obtain pose features from the recorded videos, we applied 2D pose estimation (Cao et al., 2019) to all frames and removed noisy estimations by the following equation:

$$Z_t = \frac{\sum_{i=-(u-1)/2}^{(u-1)/2} Z_{t+i} C_{t+i}}{\sum_{i=-(u-1)/2}^{(u-1)/2} C_{t+i}}, \qquad (12)$$

where $Z$ is the pose feature, $C$ is the confidence of the CNN (Cao et al., 2019) output, $t$ is the $t$-th frame of the video and $u$ is the window size.

## 5 EVALUATION

We evaluated whether the change in the melody and the video were seamless in the Melody Slot Machine (Hamanaka et al., 2019). We asked the participants to answer three questions after watching a video using a seven-point Likert scale (1: strongly disagree; 7: strongly agree) as follows.

**Q1:** *"Did you feel that the melody was as natural as the original melody?"*

**Q2:** *"Did you feel that the video was matched to the melody?"*

**Q3:** *"Did you feel that the video was as natural as the original video?"*

The participants were twelve young adults (age: 21–25), and four of them had experienced playing an instrument. We compared two conditions for the melody (original melody and the melody with the transition of melody morphing level (Hamanaka et al., 2017)) and five conditions for the video (original video, our method, merely combined video, (Niklaus et al., 2017), and (Wang et al., 2018b)) using a Wilcoxon signed-rank test (Wilcoxon, 1992) with 1% and 5% levels of significance. Tests 1–4 were selected from videos that had a difference in pose similarity (Test 1 was a video that interpolated two videos with a high degree of pose similarity, and Test 4 interpolated two videos with a low one). Fig. 7 shows that all of the melodies in the four tests were natural regardless of the difference conditions. Fig. 8 shows that the quality of the video depended on the joint parts of the videos. There was a mismatch between melody and video when the video was interpolated or combined. Fig. 8 indicates that improving video interpolation, which is matched to the melody, is important for user experience. Our method is superior to any other method in most cases. Fig. 9 shows that our method makes the produced a video appear more natural compared to those produced using other methods. Our method was effective in Test 2 and Test 3 because there are significant differences between our method and the others. In particular, in Test 2, the original video and the one produced with our method show no significant differences.

Fig. 10 and Fig. 11 show a comparison of (a) two target videos, (b) our method, (c) (Wang et al., 2018b) and (d) (Niklaus et al., 2017). In Fig. 10 (c) and Fig. 11 (c), the generated images become blurred on the mallet. The blur is caused by the lack of annotations. To avoid this problem, specific annotations on the instruments will be required. In Fig. 10 (d), the generated faces are corrupted, and in Fig. 11 (d), the generated bodies are corrupted because the method (Niklaus et al., 2017) was not adequate for interpolating the target videos with significant difference in the frames. This stems from the limitations of the short-interval interpolation method. On the other hand, our method, which uses a fusion approach of generating, padding and retrieving images, as shown in Fig. 10 (b) and Fig. 11 (b), generates a seamless and series of natural video frames. We show some results in Fig. 12.

## 6 METHOD LIMITATION

Our method is suitable for videos that include a clear target with a high correspondence to audio. Then, it is not adequate to apply our method to the videos without any target. Moreover, it is challenging to generate in-between frames in the case of mute, constant sounds, and repeated sounds, which are the audio without any character. In addition, the quality of the interpolation is low when the pose similarity is quite different between the two videos. By the request of the "Melody Slot Machine (Hamanaka et al., 2019)," the interpolation length was restricted up to 20 frames because the switching process of the videos will not

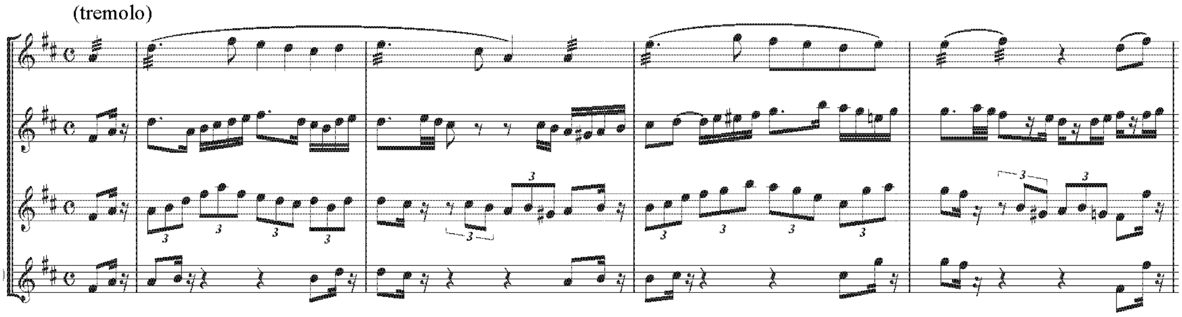

Figure 6: A part of the scores prepared for the experiments. The top score is the beginning part of the melody of Horn Concerto No. 1/Mozart, and the bottom score is the one of La Gioconda: Dance of the Hours/Ponchielli. The middle scores are the ones of morphed melodies of them generated by (Hamanaka et al., 2017).

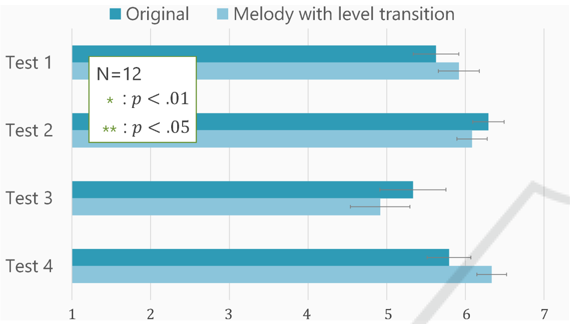

Figure 7: The subjective evaluation score for Q1.

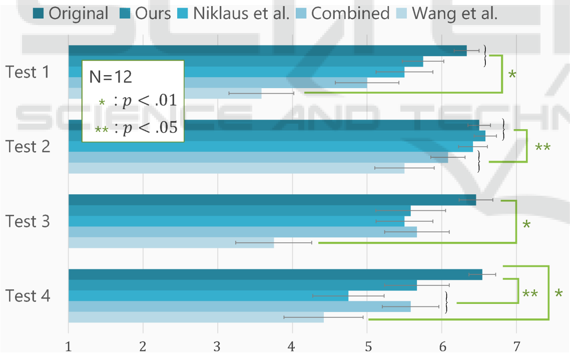

Figure 8: The subjective evaluation score for Q2.

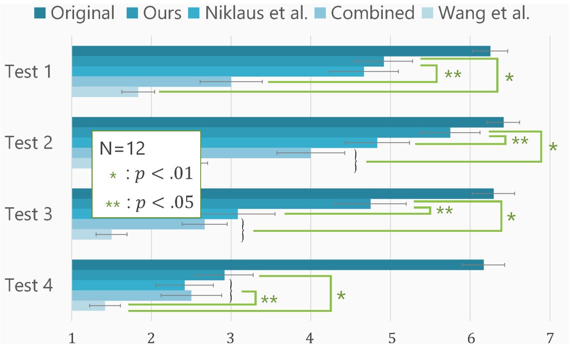

Figure 9: The subjective evaluation score for Q3.

make it in time. It is possible to generate in-between frames when the pose similarity is low, but the interpolated video appears strange.

## 7 CONCLUSION

We proposed a novel framework of video frame interpolation using audio to improve the interaction experience in the "Melody Slot Machine (Hamanaka et al., 2019)," which enables the enjoyment of manipulating music. Our method could produce the interpolating video with a long interval of frames and high resolution. We confirmed that our interpolated videos appear as seamless and natural as an original video, and satisfy the participants. We are planning to improve the quality of the interpolation when pose similarity is quite different between the two videos.

## ACKNOWLEDGEMENTS

This work was supported by the Program for Leading Graduate Schools, "Graduate Program for Embodiment Informatics (No. A13722300)" of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, JST ACCEL Grant Number JPMJAC1602, and JSPS KAKENHI Grant Numbers JP16H01744, JP17H01847, JP17H06101, JP19H01129 and JP19H04137.

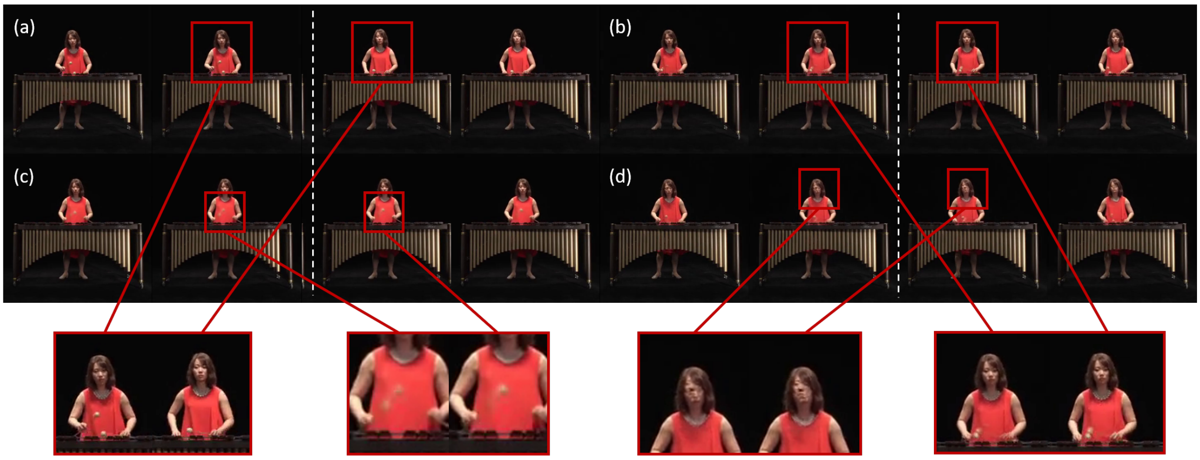

Figure 10: Comparison of (a) two target videos, (b) our method, (c) Wang *et al.* (Wang et al., 2018b) and (d) Niklaus *et al.* (Niklaus et al., 2017). The broken line denotes the boundary of between the two target videos.

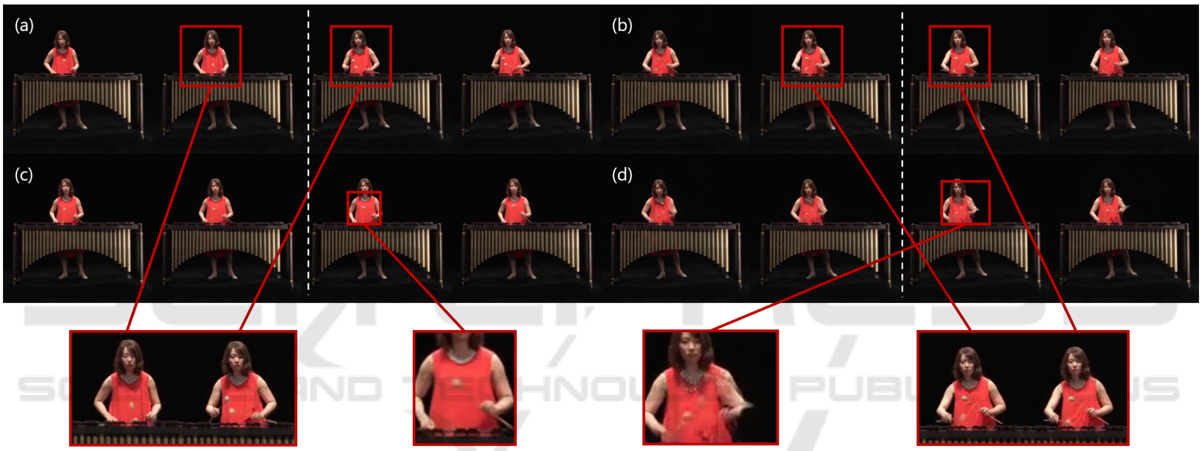

Figure 11: Another comparison of (a) two target videos, (b) our method, (c) Wang *et al.* (Wang et al., 2018b) and (d) Niklaus *et al.* (Niklaus et al., 2017). The broken line denotes the boundary of between the two target videos.

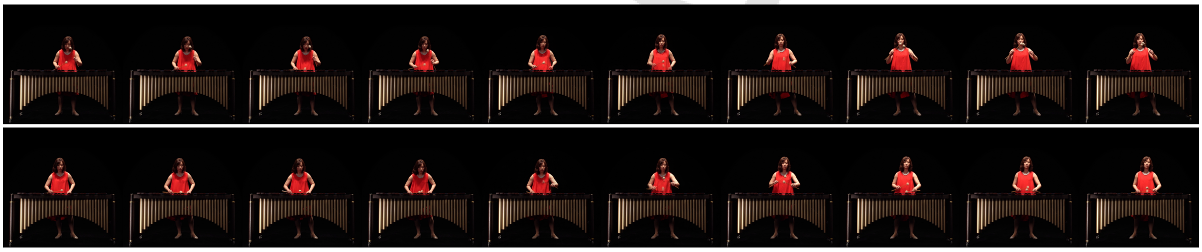

Figure 12: Results of our method. Our method succeeded in generating in-between frames with audio-guide.

## REFERENCES

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–1.

Chen, X., Wang, W., and Wang, J. (2017). Long-term video interpolation with bidirectional predictive network. In *IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014).

Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M. (2019). Capture, learning, and synthesis of 3D speaking styles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111.

Denton, E. L. and Fergus, R. (2018). Stochastic video generation with a learned prior. In *International Conference on Machine Learning (ICML)*.

Fan, R., Xu, S., and Geng, W. (2012). Example-based automatic music-driven conventional dance motion synthesis. *IEEE Trans. Visualization and Computer Graphics (TVCG)*, 18(3):501–515.

Foote, J. T. (1997). Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II*, volume 3229, pages 138–147. International Society for Optics and Photonics.

Hamanaka, M., Hirata, K., and Tojo, S. (2016). deepgttm-i&ii: Local boundary and metrical structure analyzer based on deep learning technique. In *International Symposium on Computer Music Multidisciplinary Research*, pages 3–21. Springer.

Hamanaka, M., Hirata, K., and Tojo, S. (2017). deepgttm-iii: Multi-task learning with grouping and metrical structures. In *International Symposium on Computer Music Multidisciplinary Research*, pages 238–251. Springer.

Hamanaka, M., Nakatsuka, T., and Morishima, S. (2019). Melody slot machine. In *ACM SIGGRAPH Emerging Technologies*, page 19. ACM.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Jiang, H., Sun, D., Jampani, V., Yang, M.-H., Learned-Miller, E., and Kautz, J. (2018). Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9000–9008.

Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graphics (TOG)*, 36(4):94.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lerdahl, F. and Jackendoff, R. S. (1996). *A generative theory of tonal music*. MIT press.

Li, Y., Roblek, D., and Tagliasacchi, M. (2019). From here to there: Video inbetweening using direct 3d convolutions. *arXiv preprint arXiv:1905.10240*.

Logan, B. and Chu, S. (2000). Music summarization using key phrases. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages II749–II752. IEEE.

Logan, B. et al. (2000). Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval (ISMIR)*.

Meyer, S., Djelouah, A., McWilliams, B., Sorkine-Hornung, A., Gross, M., and Schroers, C. (2018). Phasenet for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 498–507.

Niklaus, S. and Liu, F. (2018). Context-aware synthesis for video frame interpolation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1710.

Niklaus, S., Mai, L., and Liu, F. (2017). Video frame interpolation via adaptive separable convolution. In *IEEE International Conference on Computer Vision (ICCV)*.

Ofli, F., Erzin, E., Yemez, Y., and Tekalp, A. M. (2012). Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Trans. Multimedia (TOM)*, 14(3-2):747–759.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., and Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision (ECCV)*, pages 818–833.

Ruggero Ronchi, M. and Perona, P. (2017). Benchmarking and error diagnosis in multi-instance pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 369–378.

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

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Trans. Graphics (TOG)*, 36(4):95.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. (2018a). Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018b). High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, T.-H., Cheng, Y.-C., Hubert Lin, C., Chen, H.-T., and Sun, M. (2019). Point-to-point video generation. In *IEEE International Conference on Computer Vision (ICCV)*.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.

Xu, Q., Zhang, H., Wang, W., Belhumeur, P. N., and Neumann, U. (2018). Stochastic dynamics for video infilling. *arXiv preprint arXiv:1809.00263*.

Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., and Singh, K. (2018). Visemenet: Audio-driven animator-centric speech animation. *ACM Trans. Graphics (TOG)*, 37(4):161.