International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2017)
R. Lindeman, G. Bruder, and D. Iwai (Editors)

VibVid: VIBration Estimation from VIDeo by using Neural Network

K. Yoshida[1] S. Inoue[1] Y. Makino[12] and H. Shinoda[12]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2]Graduate School of Frontier Sciences, The University of Tokyo, Japan

Abstract

Along with advances in video technology in recent years, there is an increasing need for adding tactile sensation to it. Many researches on models for estimating appropriate tactile information from images and sounds contained in videos have been reported. In this paper, we propose a method named VibVid that uses machine learning for estimating the tactile signal from video with audio that can deal with the kind of video where video and tactile information are not so obviously related. As an example, we evaluated by estimating and imparting the vibration transmitted to the tennis racket from the first-person view video of tennis. As a result, the waveform generated by VibVid was almost in line with the actual vibration waveform. Then we conducted a subject experiment including 20 participants, and it showed good results in four evaluation criteria of harmony, fun, immersiveness, and realism etc.

CCS Concepts

•***Human-centered computing*** → *Haptic devices;* •***Theory of computation*** → *Models of learning;* •***Hardware*** → *Haptic devices;*

1. Introduction

Due to recent developments in video cameras and the emergence of VR technology, various images and videos have been widely spread. In particular, first-person view videos taken with small action cameras [gop] or omnidirectional images supposed to be seen with HMD (Head Mount Display) enables to follow another person's experience. Adding tactile stimulus to such videos makes a viewing experience more immersive and realistic. There are already many reports that the results of subject experiments improved by actually adding tactile sensation to the videos [DFC*12].

An emerging challenge under such a background is to create the haptic signal from the recorded video automatically. In the previous research of [GNKT17], an attempt to automatically generate tactile signals from first-person view videos on a ride was conducted. This is a system that estimates the speed and acceleration of the vehicle itself and reproduces the vibration stimulus from the vehicle to the hand holding a handle.

In this research, we extend the tactile stimulus reproduction from video and sound into a case where the relationships between videos and appropriate tactile stimulus are not so obvious as in the case of the above [GNKT17]. As an example, we focus on the shot feeling at tennis. We, in advance, learn the relation among video, sound, and the acceleration data of tennis racket by applying the latest machine learning tools. Then, by using the model learned from such datas, we produce shot feelings at tennis matching the situation only from video and audio. Since we adopts a machine learning to generate vibration estimation model, that is, we have no assumptions about vibration reproduction of tennis, it can be applied to other kinds of videos.

2. Related Works

As mentioned in the previous section, it is reported that the quality of experience is improved by adding appropriate tactile stimuli to music or videos [HC14] [DFC*12].

There are some methods for generating tactile vibrations from image frames only. Kim et al. proposed a system in which a characteristic place in an image frame can be felt on user's back using nine transducers spatially arranged in 3×3 [KLC12]. It can be applied not only to artificial images but also to natural images, and it was reported that the results of questionnaires were improved in a subject experiment. Also, Gongora et al. proposed a method of estimating the motion of the camera horizontally and vertically based on the phase correlation of the continuous image frames, and converting them into vibratory stimulations [GNKT17]. The model converts vertical motion into vertical shock vibration and horizontal motion into steady vibration moving left and right. It enhances the fitness of the model to videos by limiting the target of the model to the first-person view videos on the vehicle such as ski, bike, and bicycle.

As a method of converting to the vibration by using the sound included in the video, Lee et al. proposed a method using the two values of the sound, loudness and roughness [LC13]. This method converts these two numerical values to the intensity and roughness

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.

DOI: 10.2312/egve.20171336

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

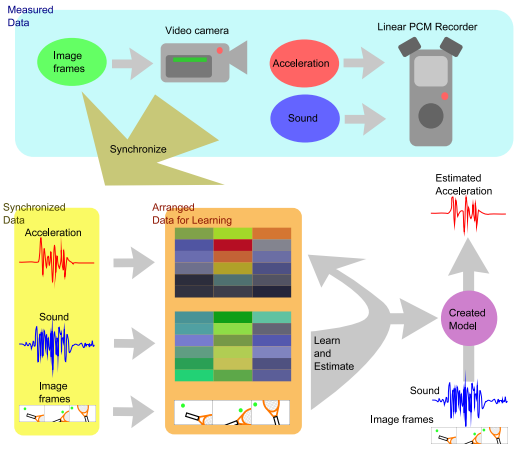

Figure 1: *System overview.*

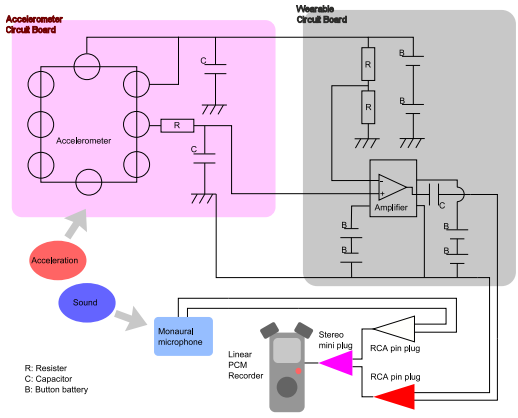

Figure 2: *Circuit diagram for recording acceleration and sound data.*

of vibration stimuli. The model shows particularly good results in games that emit specific strong sounds. But when the video contains rough sounds such as rock music or people's voice, this model might create undesired vibrations. These problems can be considered as limitations in the tactile estimation method based only on sound.

All of these conventional methods were proposals and studies based on heuristically designed model without using machine learning.

3. System Overview

The outline of the proposed system is shown in Figure 1. We use actually measured acceleration data to create the model. Image frames are recorded with a video camera, and sound and acceleration data are recorded with a linear PCM (Pulse-Code Modulation) recorder. Those image frames, sound, and acceleration data are synchronized later. Next, we create a model that estimates vibrations (acceleration data) from image frames and sounds. Before the estimation, we downsample the image frames in order to reduce the computational load, convert sounds and acceleration data to the frequency domain using STFT (Short Time Fourier Transform) to extract the feature at certain short time. In order to treat information by image and sound equally, the sound data is divided with the time corresponding to one image frame as one block.

3.1. Recording Acceleration Data

As mentioned above, it is necessary to measure the acceleration data together with the video in this method. There mainly two requests for recording this acceleration data. The first is that the recording device should be compact and wearable to avoid offending natural actions. The other is to have a sufficient sampling frequency. Generally, it is said that the frequency of tactile vibration that human can feel strongly enough is about 300 Hz or less, but it is recently reported that a sufficient refresh rate more than that is necessary to present various tactile stimuli such as hardness. Akahane et al. reported that a refresh rate of about 10k Hz is necessary to present rich tactile stimuli [AHKS05].

Therefore, in order to record acceleration data at a high frame rate, we use an analog output accelerometer and an audio recorder (linear PCM recorder). The circuit diagram is shown in Figure 2. In this circuit, a low pass filter for noise reduction, a capacitor for cutting the plug-in power voltage of the recording equipment, and an amplifier to obtain a zero average signal are integrated. Also, in order to perform wearable measurement, all the power is supplied with button batteries. By mounting the circuit board which is shown in gray in the Figure 2 in a wearable form such as a wristband, the player can record acceleration and sound while moving actively. The accelerometer circuit board and the wearable (supposed to be mounted on the player's body) circuit board are connected with two pole shielded wire, and the wearable circuit board and the sound terminal are connected with one pole shielded wire and one RCA pin plug. Since linear PCM recorder can record stereo sounds, one side is used for recording acceleration data and the other side is for recording sound by using monaural microphone. With this configuration, acceleration and sound recording are performed simultaneously.

3.2. Synchronization of Sound, Acceleration, and Image Frames

Since video and sound, acceleration data are separately recorded in this method, the synchronization is necessary. We can assume that the sound and acceleration data are synchronized because they are simultaneously recorded by a linear PCM recorder. Therefore, we synchronize sound and the image frame here. As shown in Figure 3, we use two conductive metal rods and a LED. A circuit is fabricated so that the LED emits light by contacting the bars with each other. Collision sound is detected from the recorded sound data, and LED light is detected from the image frame taken by the video camera. By making the detected metallic sound and the light from a LED at the same time, we can correct the deviation at the start of recording and synchronize all the measured data.

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.

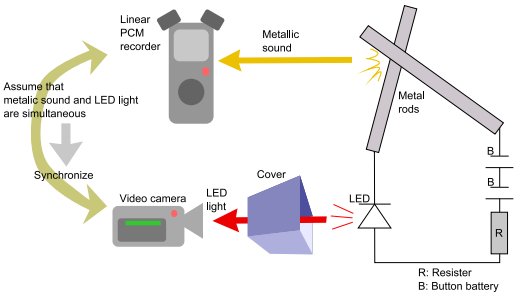

Figure 3: *Synchronization method overview.*

3.3. Estimation by Neural Network

In this method, we incorporate recorded sound, acceleration and image frames into machine learning inputs and labels in order to generate the acceleration estimation model. Therefore, a neural network is constructed whose inputs are sound data and image frames, and the output or the label is a corresponding acceleration. The outline of the proposed neural network is shown in Figure 4. For sound data, Fourier transform of audio samples for one image-frame time is used as the input of one block. Image data is downsampled into 32×32pixels square before being used for the input [TFF08a]. Since we used Fourier transform of acceleration data within a time corresponding to one image frame in the frequency domain as the label, those outputted by this model were also in frequency domain. Therefore, it is necessary to convert the outputted acceleration back to the time domain, which will be described later.

We propose two methods of model generation in this paper, Sound-image method and Sound-only method. In Sound-image method we use both sound data and image data for the inputs, and in Sound-only method we use only only sound data for the inputs. Since images and sounds are different in both the number of data and the amount of information, as shown in Figure 4, convolution layers and max pooling layers are added for image inputs, and the sound inputs are added later from multilayer perceptron to the neural network. By dividing the inputs like this, it is possible to reduce the enormous amount of information of the image to only the necessary feature quantity effectively [TFF08b]. In the Sound-only method, only the portion shown in yellow in Figure 4 is used.

We use the sounds and image frames in a time range wider (or the same) than one block which corresponds to one image frame to estimate acceleration in one block. Then as shown in Figure 4, the inputs include N blocks in time of sounds and image frames, and the output acceleration always includes one block in time (just the middle timing of the inputs of N blocks).

The back propagation method is used to advance learning of this neural network [MKB*10], and we use Adam as an optimizer [KB14] in this paper.

3.4. Conversion to Time Domain Waveform

What is estimated by the model is the power spectrum of the short-time acceleration data. Therefore, as shown in Figure 4, it is necessary to convert the output result into the time domain. Suppose a one-dimensional vector

$$\mathbf{v} = (v_1, v_2, \cdots v_i, \cdots v_M) \tag{1}$$

represents the power spectrum of the acceleration data for discrete frequency,

$$f_i = \Delta f i. \tag{2}$$

In the proposed method, the time domain signal was created from $\mathbf{v}$ as

$$V(t) = \sum_{i=0}^{M} v_i \times \sin(2\pi f_i t) \tag{3}$$

neglecting the phase information. In the experiment, we set the maximum frequency $f_M = 1080Hz$ where $\Delta f = 120Hz$.

4. Implementation

Figure 5 shows the prototypes of acceleration, video and sound recording devices. By mounting the circuit on the wristband as shown in A, B and C in Figure 5, the photographer is able to record acceleration simultaneously with active movements. This time we measured and reproduced the vibration transmitted to the tennis racket. The attachment of the accelerometer to the racket is shown in Figure 5 D, and the implementation for presenting vibration to the racket is shown in Figure 5 E.

This time ADXL001-70 (ANALOG DEVICES) is used for the accelerometer, and MM3C-LF (TactileLabs) is used as an vibrator to present vibrotactile stimuli to a person holding the grip. GoPro HERO5 Session (Woodman Labs) was used for the video camera, and LS-P2 (Olympus) was used as a linear PCM recorder for sound and acceleration recording

5. Experiment and Results

In order to evaluate the model generated by this method, we actually recorded a first-person view video of playing tennis and conducted an evaluation experiment on the reproduction of vibration. Evaluation experiments include the evaluation by reproducing the waveform and the evaluation by questionnaire when presenting the estimated vibrotactile stimuli to the subject.

A person hitting a ball and concurrently a photographer wore the devices as shown in Figure 6 A. The video taken at this time was 1280×720 pixels, 120 fps, and lasted for about two minutes. It included only forehand strokes, and the ball type was three kinds, top spin, slice, and lob shots. About ten seconds (all top spin) of this video was left for testing, and the other parts was used for the learning of the model.

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.

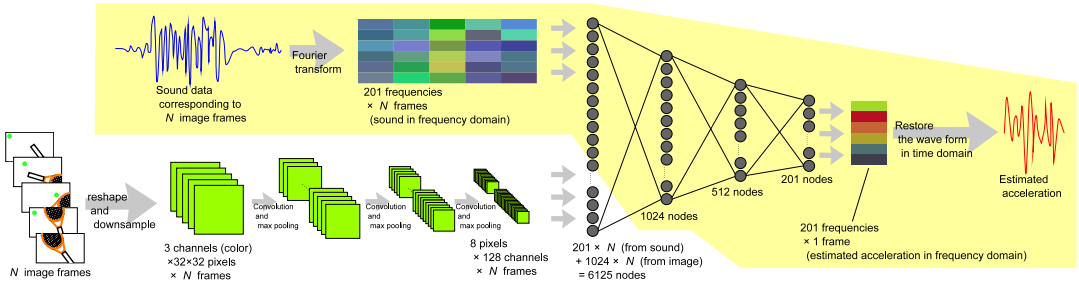

Figure 4: *Neural network overview. In the sound-only method, we used the part of the neural network shown in yellow. N in the figure represents the number of frames we use for the input of the model.*

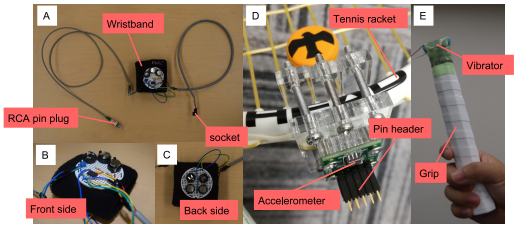

Figure 5: *A, B, C: acceleration circuit board implemented on a wristband, D: accelerometer on a tennis racket, and E: vibrator on a tennis racket grip.*

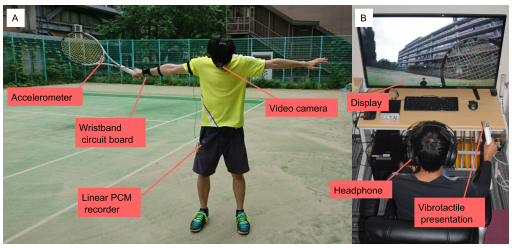

Figure 6: *A: a picture of wearing devices to record image frames, sounds, and accelerations in tennis, and B: a subject during the experiment.*

5.1. Vibration Reproduction

We applied the above video to the proposed method and learned the estimation model. Figure 7 and Figure 8 shows the learning curve during learning acceleration data in sound-only method and in sound-image method. The vertical axis represents the mean squared error, and the horizontal axis represents the epoch number. The blue line shows the error of the video used for learning, and the green line shows the error of the other part in the video for a verification of learning result. The part used for learning contains about 134 seconds, and the part used for the verification (not used for learning) contains about 9 seconds.

In both figures, since the video for testing is shorter than that for training, the variation of error for each epoch of the video for testing looks large. However, in the two lines appears to be descending in the same way. Similarly in both the sound-only method

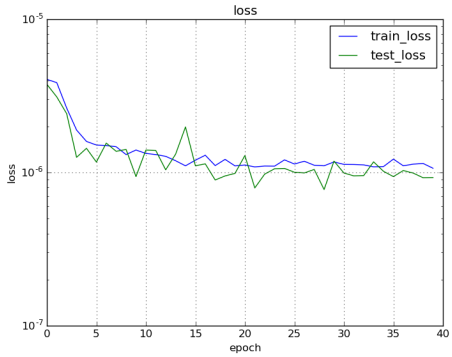

Figure 7: *Learning curve in Sound-only method.*

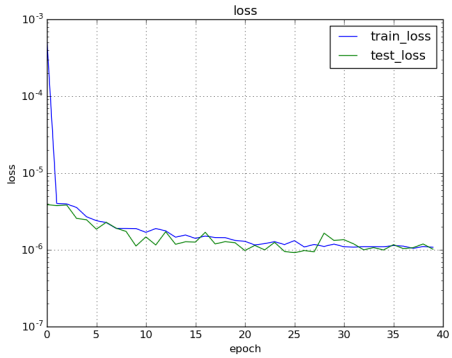

Figure 8: *Learning curve in sound-image method.*

and the sound-image method, the error value converges to about 1.0×10^{-6}, but these stabilities look different. It seems that the error converges more stably in sound-image method than in sound-only method.

The results of rendered wave forms of the acceleration data by the learning are shown in Figure 9 and Figure 10. The video for these estimation uses the part not used for learning. First, the black line shows the actual (measured) acceleration data. Next, the gray line indicates the restored data by Eq.(2) from the power spectrum of the measured data (the wave's phase information is lost for each

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.

block as mentioned in previous section). The blue line represents the recorded raw sound data. The remaining four are estimated wave forms. The yellow and the green one are estimated by sound-only method, and the orange and the red one are by sound-image method. The yellow and the orange one are the results of 10 epochs of the learning, and the green and the red one are of 40 epochs of the learning.

From the Figure 9, in the recorded sound (blue line), it seems that small waves like some noises when compared with the spike waves are larger than the others. This is conceivable to be due to the fact that sound waves are recorded by a linear PCM recorder up to high frequency bands such as human voices and wind sounds, whereas the acceleration sensor cannot record vibrations whose frequency is too high.

As for the learning, when comparing the results at 10 epochs (yellow and orange lines) and the results at 40 epochs (green and red lines) for both of the two estimation methods, it seems that the attenuation of the wave is better reproduced at the time of 40 epochs.

From the Figure 10, the start timing of the spike seems to be slightly earlier in the recorded acceleration and sound (black and blue line) than the other five. This is due to the fact that the phase information within one block is discarded when performing a Fourier transform for each one block.

Also, it is particularly noticeable around t = 0.3 in Figure 10 that the continuity of waves seen at actual acceleration was lost in the estimation result, because this method can not guarantee continuity of waves of each block in waveform generation.

The influence of these differences on people's senses will need to be verified by the subject experiments.

5.2. User Experiment

In order to evaluate vibrations estimated by this method as human senses, the following subject experiment was conducted.

5.2.1. Experiment Overview

20 persons participated in the experiment, and their information is shown in Table 1. Each participant was paid about 5 USD.

We assigned five vibrations from seven shown in Figure 9 or Figure 10 to the video. In the experiment, subjects firstly watched videos with a display and a headphone while being presented the vibrotactile stimuli as shown in Figure 6 B. The videos has 20 combinations of the above-described five kinds of vibrations (real, real-converted, sound-image method, sound-only method, and recorded sound) and four scenes (top spin shots NOT used for learning, top spin shots used for learning, slice shots used for learning, and lob shots used for learning), and the order was randomized for each participant. Before watching 20 videos for the experiment, subjects watched five videos including all the vibration patterns as a practice. After watching each video, the subject answered the following questionnaire: Harmony-"Did the vibrations match to the video?"; Fun / Satisfaction-"Did the vibrations makes the video fun?"; Immersiveness-"Did the vibrations help you be immersed in the video?"; Comfortableness / Fatigue-"Did the vibrations feel comfortable to enjoy the video?"; and Realism-"Did the vibrations played back with the video seem consistent with your real-world experiences?" To answer these questions, we adopted a method of having an answer from the position in the straight line segment, and thereby obtained a continuous value. The opposite expressions were labeled on both ends of the line segment. For example, there was"very comfortable" at the right end of the line segment, "very uncomfortable" at the left for comfortableness.

Table 1: *Subject's information*

ID	Sex	Age	Tennis experience	Dominant hand
A	Female	24	None	Right
B	Male	23	None	Right
C	Female	20	None	Right
D	Male	26	2 years	Right
E	Male	20	1- year	Right
F	Female	20	None	Right
G	Male	23	5+ years	Right
H	Male	23	None	Right
I	Male	21	None	Right
J	Male	24	5+ years	Right
K	Male	23	5+ years	Right
L	Male	24	None	Right
M	Female	23	5+ years	Right
N	Female	21	1- year	Right
O	Male	24	5+ years	Right
P	Male	24	5+ years	Right
Q	Male	23	5+ years	Right
R	Male	22	None	Right
S	Male	25	3 years	Right
T	Female	21	3 years	Right

5.2.2. Results

The average of the evaluation points of 20 subjects and their standard errors are shown in Figure 11, 12, 13, 14. In addition, multiple comparison was performed by Tucky's HSD tests, and significant parts were marked with "*" at the 5% significance level, and "**" at the 1% significance level. The two horizontal lines in the upper and lower parts in the graph indicate the positions of both ends that were the range of the answer, and one horizontal line in the middle indicates the middle of the line segment (the answer sheet).

First, when comparing Figure 11 and Figure 12, the results of top spin shots, there was hardly any difference between the video used for learning and the remaining video. In the video of top spin, the evaluation of the real-converted vibration tends to be the highest in almost all results, and the vibrations generated by our method were then the second most favorable. Finally, the evaluation of recorded sound was the lowest. Also, the three vibrations of real-converted, sound-image method, and sound-only method did not show a significant difference from the case where truly measured (real) vibration was presented.

In the results of the slice shots in Figure 13, they were

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.

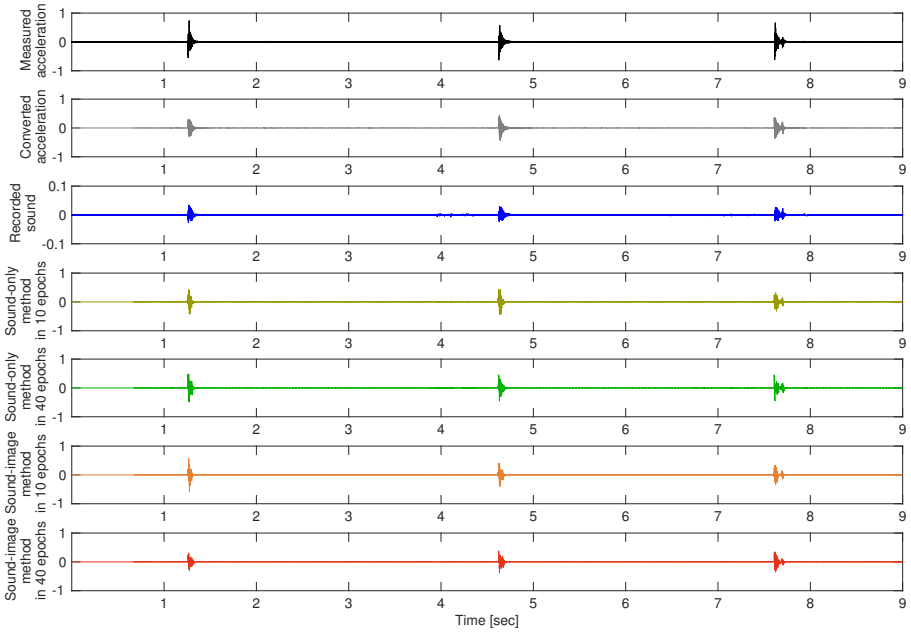

Figure 9: *Results of acceleration wave forms estimated by the system.*

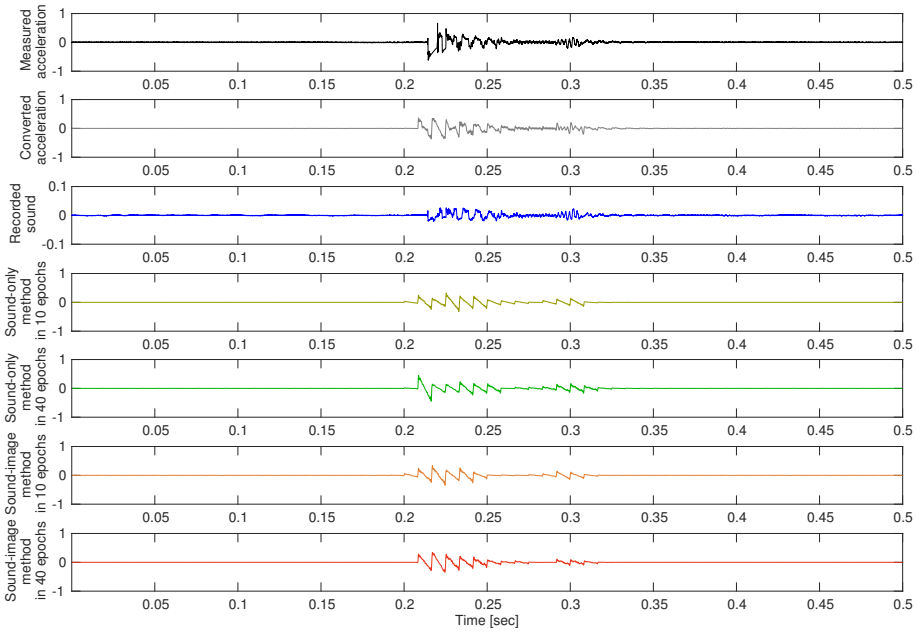

Figure 10: *Enlarged view of a single shot waveform for each method.*

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.

not significantly different among the four methods of real, real-converted, sound-image method, and sound-only method. The result of recorded sound (vibration following the recorded sound data) was significantly low similarly as the case of topspin.

As common features in Figure 11, 12, and 13, there was no significant difference among the five vibration patterns with the index of "Comfortableness", but in the other indices, the evaluation of recorded sound was statistically significantly lower.

Looking at the results of lob shots in Figure 14, unlike other results, the evaluation of recorded sound was significantly lower in all indices including Comfortableness.

6. Discussion and Future Work

In the reproduction of the tennis vibration waveform, it can be said that we could estimate the result that is close to the measured value by this method if the generated waveform is visually evaluated. However, the problem of wave discontinuity for each block also remains. So there is a possibility that the component of the frequency derived from the frame rate of the video influences the generated waveform. As a solution to this problem, we can mainly think of two. One is to match the amplitude position (and its derivative) of the waveform of each block, and the other is to add estimation results over multiple blocks through window functions while shifting them one by one. However in the latter case, if the Fourier transform of a long section is performed with the method of missing the phase information, it is conceivable that the deviation of the timing of a peak wave becomes large. Therefore, when Fourier transform is performed with a block of a longer time interval, there is a possibility that data used for learning also needs to include phase information.

Looking at the decrease in errors per epoch number of the learning, the error of the video used for learning and that of the video which is not so seemed to be almost the same from the viewpoint of converging value. Also in the subject experiments, there were no significant differences between the results of videos used for learning and those of videos not so. This means that, though the videos were all taken on the same day at the same place, they were able to respond to videos with a certain level of comprehension capability by machine learning. It is necessary to investigate how accurate acceleration can be estimated for unknown (unlearned) videos by increasing video data in the future.

In subject experiments, we performed comparative evaluation on five vibrations of real, real-converted, sound-image method, sound-only method, recorded sound. According to this, the evaluation of recorded sound was lower than others entirely. The reason for this may be that the waveform recorded up to the high frequency component by the microphone was applied to the vibrator that greatly vibrates at low frequency, resulting in only weak vibrations presented. So, only in evaluation criteria of Comfortableness, there was only one video where recorded sound was significantly lower.

Statistically significant differences were not observed for the four vibrations of real, real-converted, sound-image method, and sound-only method. However, the real tended to be slightly lower than the other three vibrations. Although this is an unexpected re-

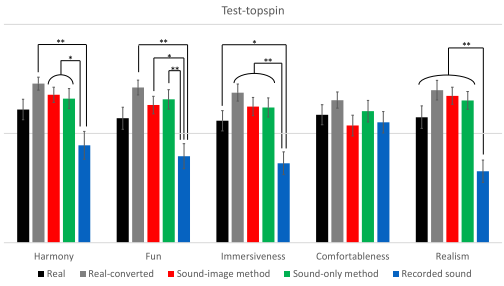

Figure 11: *Top Spin Shots NOT Used for Learning*

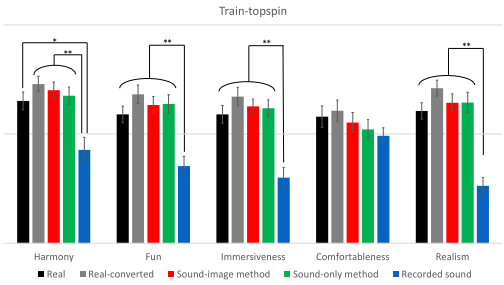

Figure 12: *Top Spin Shots Used for Learning*

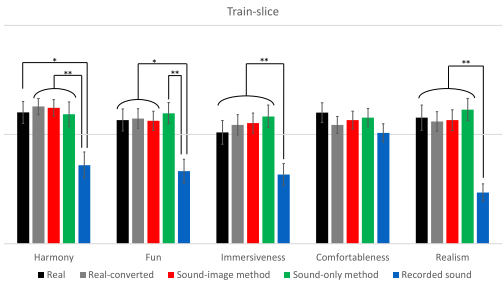

Figure 13: *Slice Shots Used for Learning*

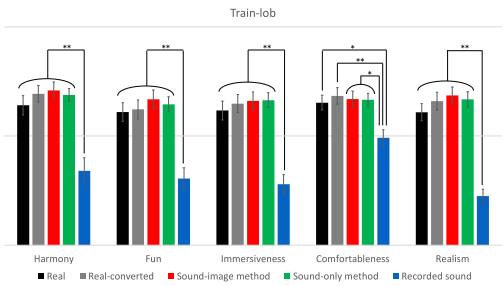

Figure 14: *Lob Shots Used for Learning*

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.

sult, the following three factors can be the reason for it by the processing such as Fourier transform included in our method: (1) phase information is lost, (2) the waves are discontinuous per a block, and (3) high frequency components have been removed. We will find out which factors have improved the evaluation by further comparison experiments. However, as a result of this experiment, it can be said that the present method can realize a viewing experience equal to or slightly higher than presenting actually measured acceleration as vibrotactile stimuli.

In addition, a reason why statistically significant difference was NOT found between the sound-image method and the sound-only method, seems that the video selected this time was noiseless clean tennis. In the video of tennis, the relationship between vibration and image is not so obvious, but it seems to be highly correlated with sound. As a result of that, we think that sufficient oscillation was generated even with the estimation result by only sound. If we compare them in the genre where vibration is not correlated with the image alone or the sound alone, we consider the results may change. Or, if noise is contained in the sound with high correlation between sound and vibration, there is a possibility that it can be made robust to noise by using not only sound but also image as input.

The difference due to the tennis ball type was hardly seen, but only in the lob shots, the recorded sound had a lower Comfortableness than the other. It is conceivable that this value is not low but Comfortableness against other vibrations has increased. The reason for this is that since the shot feeling contained in the image was soft, the degree of visual fatigue to the viewer may have been small.

The results of this experiment also seems to have a considerable influence of visual and auditory information, which is often called cross modal. In such a viewpoint, although we applied vibrotactile sense to ordinary videos this time, it is very meaningful to investigate the improvement of the immersive feeling by applying vibration to the 360-degree video for HMD. Also, we adopted the vibratory presentation method to vibrate only the grip part of the tennis racket this time, but if we can conduct the experiment while actually swinging the grip, some participants of the experiment said, that the result may improve more. It can be realized with a wireless and high power vibration device using a small battery and amplifier.

7. Conclusion

We proposed a method VibVid that estimates vibrotactile stimuli from the existing videos with sounds, using machine learning. As an example of that, we implemented and evaluated a video of first-person view playing tennis. As a result, we succeeded in generating a waveform close enough to the desired acceleration waveform. Also in a subject experiment, estimated waveform obtained comparable scores to the measured data especially in harmony, fun, immersiveness, and realism. There were statistically significant difference between those estimated acceleration waveforms and recorded-raw-sound vibrations in the tactile presentation experiment.

Theoretically, this method can be applied not only to first-person view videos of tennis but also to other genres' videos. This is because the method does not include the explicit assumption that the video is captured from first-person view or that the photographer holds the racket in the video.

In the future, we aim to expand the genres of videos applicable by the method and to generate higher quality vibrotactile waveforms. Also, this method may be applied not only to vibrations but also to another type of haptic sensation such as movements and inclinations with lower frequencies. In order to explore the scope of applicability of this method, we further such tactile measurement and presentation.

References

[AHKS05] AKAHANE K., HASEGAWA S., KOIKE Y., SATO M.: A development of high definition haptic controller. In *Proceedings of the First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems* (Washington, DC, USA, 2005), WHC '05, IEEE Computer Society, pp. 576–577. URL: http://dx.doi.org/10.1109/WHC.2005.5, doi:10.1109/WHC.2005.5. 2

[DFC*12] DANIEAU F., FLEUREAU J., CABEC A., KERBIRIOU P., GUILLOTEL P., MOLLET N., CHRISTIE M., LÉCUYER A.: Framework for enhancing video viewing experience with haptic effects of motion. In *2012 IEEE Haptics Symposium (HAPTICS)* (March 2012), pp. 541–546. doi:10.1109/HAPTIC.2012.6183844. 1

[GNKT17] GONGORA D., NAGANO H., KONYO M., TADOKORO S.: Vibrotactile rendering of camera motion for bimanual experience of first-person view videos. In *2017 IEEE World Haptics Conference (WHC)* (June 2017), pp. 454–459. doi:10.1109/WHC.2017.7989944. 1

[gop] Gopro. https://jp.gopro.com/. 1

[HC14] HWANG I., CHOI S.: *Improved Haptic Music Player with Auditory Saliency Estimation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 232–240. URL: https://doi.org/10.1007/978-3-662-44193-0_30, doi:10.1007/978-3-662-44193-0_30. 1

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2014). URL: http://arxiv.org/abs/1412.6980. 3

[KLC12] KIM M., LEE S., CHOI S.: *Saliency-Driven Tactile Effect Authoring for Real-Time Visuotactile Feedback*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 258–269. URL: https://doi.org/10.1007/978-3-642-31401-8_24, doi:10.1007/978-3-642-31401-8_24. 1

[LC13] LEE J., CHOI S.: Real-time perception-level translation from audio signals to vibrotactile effects. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2013), CHI '13, ACM, pp. 2567–2576. URL: http://doi.acm.org/10.1145/2470654.2481354, doi:10.1145/2470654.2481354. 1

[MKB*10] MIKOLOV T., KARAFIÁT M., BURGET L., CERNOCKÝ J., KHUDANPUR S.: Recurrent neural network based language model. In *INTERSPEECH* (2010), Kobayashi T., Hirose K., Nakamura S., (Eds.), ISCA, pp. 1045–1048. 3

[TFF08a] TORRALBA A., FERGUS R., FREEMAN W. T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell. 30*, 11 (Nov. 2008), 1958–1970. URL: http://dx.doi.org/10.1109/TPAMI.2008.128, doi:10.1109/TPAMI.2008.128. 3

[TFF08b] TORRALBA A., FERGUS R., FREEMAN W. T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 11 (Nov 2008), 1958–1970. doi:10.1109/TPAMI.2008.128. 3

© 2017 The Author(s)
Eurographics Proceedings © 2017 The Eurographics Association.