Automatic Generation of Spatial Tactile Effects by Analyzing Cross-modality Features of a Video

Kai Zhang
Stanford University
kzhang3@stanford.edu

Lawrence H. Kim
Stanford University
lawkim@stanford.edu

Yipeng Guo
Stanford University
guo322yp@stanford.edu

Sean Follmer
Stanford University
sfollmer@stanford.edu

ABSTRACT

Tactile effects can enhance user experience of multimedia content. However, generating appropriate tactile stimuli without any human intervention remains a challenge. While visual or audio information has been used to automatically generate tactile effects, utilizing cross-modal information may further improve the spatiotemporal synchronization and user experience of the tactile effects. In this paper, we present a pipeline for automatic generation of vibrotactile effects through the extraction of both the visual and audio features from a video. Two neural network models are used to extract the diegetic audio content, and localize a sounding object in the scene. These models are then used to determine the spatial distribution and the intensity of the tactile effects. To evaluate the performance of our method, we conducted a user study to compare the videos with tactile effects generated by our method to both the original videos without any tactile stimuli and videos with tactile effects generated based on visual features only. The study results demonstrate that our cross-modal method creates tactile effects with better spatiotemporal synchronization than the existing visual-based method and provides a more immersive user experience.

CCS CONCEPTS

• **Human-centered computing → Haptic devices**.

KEYWORDS

Haptic, Audiovisual Content, Vibrotactile Display, Computer Vision

ACM Reference Format:
Kai Zhang, Lawrence H. Kim, Yipeng Guo, and Sean Follmer. 2020. Automatic Generation of Spatial Tactile Effects by Analyzing Cross-modality Features of a Video. In *Symposium on Spatial User Interaction (SUI '20), October 31-November 1, 2020, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3385959.3418459

SUI '20, October 31-November 1, 2020, Virtual Event, Canada

ACM ISBN 978-1-4503-7943-4/20/10…$15.00
https://doi.org/10.1145/3385959.3418459

1 INTRODUCTION

Vibrotactile feedback is effective in making multimedia content more expressive, vibrant, and realistic [6]. 4D movies are one of the successful attempts to integrate tactile feedback for an improved visual-audio content delivery [9, 26]. In this context, it is very important to present tactile stimuli that are aligned with the visual and auditory stimuli as well as their semantics [6]. Any conflict between the sensory cues will compromise the overall user experience. Thus, a key challenge for spatial tactile effects is the spatiotemporal synchronization between the audio-visual content and vibration patterns rendered on a tactile display [6].

To create the haptic content for vibrotactile displays, researchers have developed many manual authoring tools [4, 20, 31]. However, manual editing is a laborious process especially for long videos. To address this, researchers have explored various ways to automatically generate tactile effects. Visual saliency calculated from the object's motion is used to generate spatial tactile content by Kim et al. [17]. This method performs well for simple scenes but due to its lack of understanding cross-modal information, complex scenes with rich audio and visual events cannot be translated in the most meaningful way. For instance, generating temporally synchronized tactile effects for a moving truck that honks sporadically is impossible with visual features alone. Audio signals without any visual information (e.g., music or special sound effects) have also been translated to tactile stimuli [5, 22]. However, the resulting haptic content is not spatially aligned with the visual content. While spatial audio channel could be used to generate spatial haptic content, many online videos (e.g., YouTube) do not have the necessary spatial audio information.

This work focuses on automatically generating the haptic content for a spatial vibrotactile display based on the audiovisual information from a video. Contrary to previous work which only utilized either the visual or audio channel, we use cross-modal information from both channels to improve the spatiotemporal synchronization between the tactile stimuli and the audiovisual content. This design decision is supposed to improve the user experience compared to when solely relying on a single modality. In order to extract audio and visual features, we first use neural networks to pull out diegetic audio signals (i.e., sounds from objects that are visible or can be implied from the scene) [28] from the sound track. The diegetic sound is then used to decide the intensities of the tactile stimuli and localize the sound source within the scene [36], which is then mapped to a spatial heatmap of the tactile effects.

In order to evaluate the performance and user experience of our cross-modal approach, we conducted a user study to compare videos with tactile effects from our method against videos without any tactile stimuli and videos with tactile stimuli generated by a modified version of the visual saliency-based approach from prior work [18]. For our study, we used a custom designed tactile display, which consists of an array of haptic actuators mounted on a backrest of a chair, to display the tactile stimuli to the back of the participants. To evaluate the user experience, we measured the Sensory, Distraction, Novelty and Immersion on a Likert Scale. Videos downloaded from YouTube were used for the study. Based on the study results, we discuss advantages of extracting cross-modal information for automatic generation of tactile stimuli compared to using visual channel alone. The novelty and technical contribution of this paper center around its use of both audio and visual content to automatically generate spatial tactile effects. While we used algorithms from prior work [28, 36] for parts of our pipeline, none to our knowledge have developed a complete pipeline that utilizes both the audio and visual information nor compared its performance to visual-only algorithms.

2 RELATED WORK

2.1 Haptics to Enhance Multimedia Experience

Researchers have investigated the importance of haptics for improving the multimedia experience. Both kinesthetic [2, 27] and tactile [20, 23, 31] haptic effects can complement the audiovisual delivery. Tactile effects can be displayed by attaching vibrotactile actuators to different fixtures including hand-held game controller [35], gloves [20], mobile phone [12], and jackets [23]. Tactile effects have been used for various purposes such as making more realistic physical contact [3], enhancing audio effects [33], and simulating motion [11]. An array of tactors is commonly used to display tactile effects [6], which can present spatial correspondence between visual information and tactile stimuli. Israr et al. [11] and Kim et al. [19] further discussed the design space of tactile effects with the spatial correspondence. Compared with traditional plain audiovisual content, enhanced multimedia content with physical output can help users perceive information in a more immersive way [25]. Our work focuses on generating haptic patterns for tactile arrays as such devices are being increasingly used in commercial theaters and are more affordable than kinesthetic devices.

2.2 Authoring Haptic Effects

One of the key requirements for generating haptic effects is the synchronization with the audiovisual content. There has been mainly three ways to achieve this: capturing physical sensor data during filming [7], having haptic content designers manually add haptic features [31, 37], and automatically generating haptic effects based on the video or audio content [17, 21, 22].

2.2.1 Sensor-based Authoring. The most comprehensive approach is to capture the exact forces and motion by placing the appropriate sensors directly in the scene. For instance, piezo-electric sensors and accelerometers have been used to monitor contact forces [30] and motion profiles [1]. However, capturing these physical sensor data can dramatically increase the setup time and the overall cost as additional technicians and equipment are required.

2.2.2 Manual Authoring. Another approach involves designers manually creating haptic content. To support this, researchers have developed various graphical editing tools that allow designers to set the amplitude of a specific actuator [20, 31] or only require the description of what the user should feel [7, 37]. Although manual editing tools can be very effective for a small number of videos, it is not a scalable solution.

2.2.3 Automatic Authoring. More recently, researchers have explored automatic generation of haptic effects. By extraction of specific features (i.e. motion [17]), these automatic generation methods can translate video content to haptic effects. However, since these methods only rely on visual features, they can often generate contextually inapt tactile effects. For example, these methods will always generate strong tactile stimuli whenever there is a large object moving across the scene. Audio signals have also been used to generate haptic effects by analyzing their frequency characteristics [5] or perception-level intensity [22]. However, simply translating an audio signal to a haptic signal alone without understanding its semantic correspondence with visual content could potentially lead to inadequate haptic effects. For example, directly translating an off-screen narrator's voice in a video to spatially distributed tactile stimulus can be confusing to the audience.

Building an end-to-end pipeline that can automatically generate spatial vibrotactile effects to a large number of videos is still an unsolved challenge. We will demonstrate that utilizing cross-modal information instead of single-modal information can be a significant step toward addressing this problem.

3 FRAMEWORK FOR AUTOMATICALLY GENERATING SPATIAL TACTILE EFFECTS

In this section, we provide an overview of our framework to automatically generate spatial tactile effects. Our pipeline utilizes both the visual and the audio signals to determine the spatial distribution and the intensities of the tactile effects.

As shown in Fig. 1, the audiovisual content is first separated into the visual and audio content. Since the relation between the nondiegetic sounds (i.e., the sound that is not visible on the screen or whose source cannot be implied by action of the film) and the haptic signals is not well-understood [8], we extract only the diegetic sounds from the audio content and discard the non-diegetic audio information. With the recent progress in neural-network-based methods, many researchers have applied neural networks to separate diegetic and non-diegetic audio channel. For our pipeline, we used the most relevant on/off screen speaker audio separation algorithm developed by Owens et al. [28]. The outputs of our pipeline are spatial tactile effects that consist of the spatial distribution and the intensities of the tactile effects. In order to obtain the spatial distribution, we calculate the location of the sound source in the video using an audio-guided visual attention mechanism [36]. This method creates a heatmap representing the probability distribution of the sound source within the scene. Only the diegetic audio obtained from the audio separation algorithm is then used as the input data for this process since non-diegetic audio is not visually

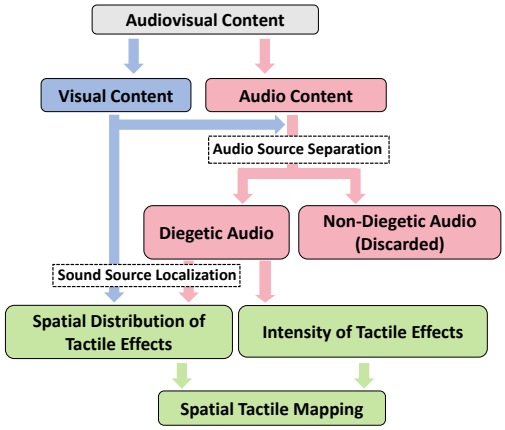

Figure 1: Automatic tactile effects generation pipeline uses both visual and audio features to separate diegetic audio signal and determine the location of the tactile stimuli. The intensity of the generated haptic effects is only decided by the diegetic audio signals.

present in the scene. The amplitudes from the diegetic audio signal are converted to the intensities of the tactile effects while satisfying the frequency response of the actuator. Finally, we combined the two components (i.e., spatial distribution and intensities) to generate tactile effects at the computed locations with the computed intensities.

3.1 Audio Source Separation

Audio source separation is a necessary step before translating audiovisual content to meaningful haptic mappings. There are mainly two types of audio in a video: diegetic and non-diegetic sounds. Diegetic sounds originate from the objects on the screen, like car engine, chainsaw or music instrument while non-diegetic sounds include background music and narrator's voice. Since haptic effects are mostly used to enhance physical events happening within the scenes [8], it is necessary to filter out non-diegetic sounds from the mixed audio channel.

Among state-of-the-art audio source separation techniques, the model developed by Owens et al. [28] is the most relevant for our application since it allows us to separate diegetic and non-diegetic audio signals. Therefore, we adopted their open-sourced multisensory net and u-net models [29] as building blocks to separate non-diegetic audio signals and only use the diegetic sound in our framework (Fig. 1). Although Owens et al. [28] only demonstrated a speech separation use case with an on-screen speaker and an off-screen human speaker, we also found it effective in separating the audio stream from on-screen non-human sounding objects and an off-screen narrator. Next in Sec. 3.2, we will translate the obtained diegetic audio signals into spatial tactile mappings by locating the sounding object.

3.2 Sound Source Localization

Using the extracted audiovisual content without the non-diegetic audio component, we generate the spatial tactile channel. As mentioned in Sec. 3, our framework obtains the distribution of the spatial

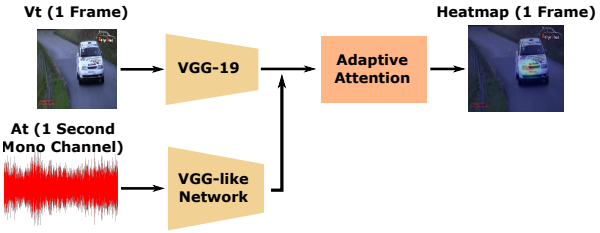

Figure 2: The pipeline for sound source localization. The visual features are extracted through a VGG-19 network [34] while the audio segment is processed by a VGG-like network [10]. The audio-guided visual attention is then used to generate a heatmap showing the location of the sounding object.

haptic channel by localizing the sounding object in the scene. In this section, we describe the methods to achieve the spatial and temporal sound event localization.

3.2.1 Our Implementation based on the State-of-the-Art. There are many existing approaches to the sound event localization problem. Among all the sound event localization approaches, we follow the method developed by Tian et al. [36] for our automatic haptic effects generation pipeline. Their model [36] is both open-sourced and is capable of recognizing various sounding objects in the audiovisual scene. This model was originally developed for audio-visual event localization so it consists of five major modules including visual and audio features extraction, audio-guided visual attention, temporal modeling, multimodal fusion, and temporal labeling. Among the five major modules of the pipeline developed by Tian et al., the visual and audio feature extraction module preprocesses the audiovisual content and further feeds the intermediate results to the other modules. The audio-guided visual attention module generates the sound localization heatmap using the extracted visual and audio features. Thus, we modified and used these two modules to locate the sounding object within a video.

As shown in Fig. 2, three neural networks were utilized in this step. A VGG-19 network [34] pre-trained on ImageNet was used to extract visual features. The original model by Tian et al. [36] sampled 16 RGB video frames from a 1 second video clip and did a global average pooling over the 16 extracted *pool*5 features to generate one $512 \times 7 \times 7$-D feature map. This will result in the same generated heatmap for the entire 1 second video clip which does not meet our requirement for dynamic response of haptic effects. Thus, we modified the visual feature extraction process so that each frame of a video clip is sent to the VGG-19 network to extract its *pool*5 feature without averaging the other frames. This will generate a sound source localization map for each frame in a video clip. Then, a VGG-like network [10] was used to extract a 128-D audio representation for each 1 second audio segment. Since the location of the sounding source is more sensitive to the visual features than the audio features, we used each individual frame as the visual input and 1 second segment as the audio input to reduce noise. A video clip longer than 1 second was separated into multiple 1 second chunks and the sound source localization was conducted on each of the chunk.

The resulting sound localization heatmap is shown in Fig. 2. The audio-guided visual attention method will only highlight the sound source location (e.g., engine of the car) instead of the entire sounding object (e.g., car). Several random background regions will be attended by this model if there are no audio-visual events happening. However, the performance of our method will not degrade due to this phenomenon because the intensities of our tactile effects translated from audio signal intensities will also be very weak, which will not distract users.

Since sound source localization is also a well-established research topic in the computer vision community with many working models, we mainly trimmed and modified the most relevant model [36] as we discussed above to utilize it as a building block for our automatic tactile effects generation pipeline. This sound localizing module will be fed with diegetic audio that we get from the audio separation module along with the visual frames to generate a localization heatmap of the sounding object at each frame.

3.2.2 Limitations. There are some limitations for the current audio-guided visual attention model [36]. Currently, it only works for 28 events (e.g., Racing car, Dog barking, Chainsaw, etc) due to its relatively small training dataset (4143 videos). However, the model itself is generalizable and can be extended to a larger number of event categories given a larger training dataset. Another limitation is that, when there are multiple sounding objects from the same category in the scene, the current model will have difficulties finding the correct sound source. For example, a race car in the street generates much more sound than an electric car. However, since any types of cars will be classified as the car object by this method, both cars may be highlighted in the sound source localization heatmap. This will lead to a mismatch between spatial haptic effects converted from the generated heatmap and the audiovisual content. However, selecting the correct sounding object among other objects from the same category is also a difficult task for humans if there is only a mono audio channel. The potential effects of this limitation will also be discussed in Sec. 4.

3.3 Spatial Tactile Effects

In this section, we describe our spatial tactile stimuli generation pipeline based on the diegetic audio stream obtained from the audio source separation method described in Sec. 3.1 and the sound source localization heatmap generated through a set of neural networks (Sec. 3.2). To compute the intensities of the tactile effects, we use the amplitudes of the diegetic audio signal whereas the heatmap from the sound source localization is used to determine the spatial distribution of the tactile stimuli.

3.3.1 Intensity. As Fig. 3 shows, the intensities of the haptic signals are calculated from the amplitudes of the diegetic audio stream. Since haptic actuators have a limited frequency response range, the audio signals are first downsampled with a sampling rate slightly higher than $2F_s$ according to the Nyquist Theorem, where F_s is the resonant frequency of the haptic actuator. A commonly used downsampling method is to sample one data point in every n data points in an audio stream s_1. However, this downsampling method overemphasizes the low-frequency components in the audio stream s_1. In some cases, there can be a significant difference between the downsampled waveform and the original audio waveform leading to a mismatch between the haptic and audio signals that compromises the user experience. Thus, this method is not adopted in our pipeline. Our system downsamples using a "running sum" method [24] which can retain the characteristics of the audio signals. Assuming the original audio stream s_1 has a sampling rate of r_1 and the target downsampled data stream s_2 has a sampling rate of r_2, we need to convert every $n = r_1/r_2$ signals from s_1 into one data point in s_2. We first calculate the sum of the absolute amplitude of every n signals in s_1, thus obtaining a data stream $s_1{}'$ that has a sampling rate of r_2. We further add a minus sign to every other data point in $s_1{}'$ so that we get a data stream s_2 with a waveform similar to s_1 as Fig. 3 depicts. We can observe that although there are minor differences between the original audio stream and downsampled data stream, the overall waveform shape is maintained.

3.3.2 Spatial Distribution Mapping. The distribution of the spatial haptic effects is determined by the sound source localization heatmap. Since this heatmap is a probability distribution of the sound sources rather than an audio intensity map, darker areas in the heatmap do not indicate weaker audio signals in the visual scene. Thus, instead of providing weaker tactile vibrations proportional to the heatmap values in the darker areas, we filter out these dark areas and do not render any haptic signals at those regions. To do this, we first carry out a percentile contrast stretching on the sound source localization heatmap (Fig. 3) with 1st and 99th percentile in the histogram converted into 0 and 255 while other pixels remapped to 0 to 255. This filtering step is commonly used in the image processing pipeline to boost the image contrast [32]. Next, we downsample the contrast stretched sound source localization heatmap to a 3×3 tactile map which reflects the number of actuators in our vibrotactile device. A moving average calculation is used to remove the random noise in the tactile map, which is caused by inaccuracy of the sound localization model. Specifically, $M_n = 0.95 * M_{n-1} + 0.05 * M_{new}$ where M_n and M_{n-1} is the filtered tactile map at time step n and $n-1$, M_{new} is the incoming tactile map with noise. Another contrast stretching is further carried out for the created tactile mapping for the same reason as mentioned above. To reduce the effects of the low probability regions, we apply a threshold similar to the prior work [18]. Tactile map pixels with a value lower than a certain threshold (e.g., 150 out of 255) were removed. A higher threshold can reduce noise although sacrificing the tactile mappings' ability to show minor tactile effects. The threshold was decided by our empirical tests.

The heatmap from the sound source localization can often output noise. As shown in Fig. 3, the bright region on the left of the heatmap is due to noise and the bright region on the right is the actual sounding object. However, this noise can be detected and removed by comparing with the heatmaps from the neighbouring time steps. As shown in Fig. 3, the distribution of the bright region in the heatmap is maintained after the contrast stretching and downsampling steps. After the moving average step, the bright pixel values on the left side of the spatial tactile mapping is significantly reduced. This is because other sound localization heatmaps (not shown in Fig. 3) generated at the neighbouring time steps do not contain such bright regions on the left side of the mapping. Since we apply a weight of 0.05 to the new tactile map as mentioned

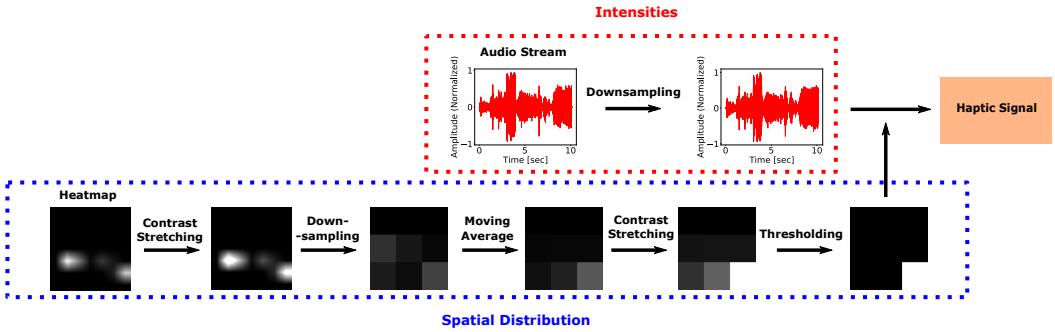

Figure 3: The generation pipeline builds the spatial tactile mapping by calculating both the spatial distribution and the intensities of the tactile stimuli. Amplitudes of audio stream after downsampling is translated to the intensities of the tactile stimuli. The sound source localization heatmap is used to determine the spatial distribution of the haptic effects.

Figure 4: Spatial tactile mappings generated for video examples downloaded from YouTube. Whiter regions indicate stronger tactile amplitudes.

above, the noise on a single frame has minor effect on the filtered tactile map. Thus, these pixels are construed as random noise generated by the sound source localization algorithm and removed by the moving average step.

Although the moving average mitigates some harmful noise, it also reduces its capability to capture fast moving objects. If the pipeline is used to generate spatial tactile effects for body parts with lower spatial tactile perception resolution (e.g., torso or arm), a longer time window can be used to remove the random noise more effectively. Even if the objects slightly change their positions in the video, they can still be presented by the same pixel in the spatial tactile mapping due to its lower resolution. On the other hand, if the spatial tactile stimuli is used for body parts with a higher spatial tactile sensitivity (e.g., hand or fingertip), the moving average step should assign more weights to the current video frame so that the generated spatial tactile mapping can better match the sounding object. Since the current vibrotactile display outputs tactile effects on user's back which has a low spatial tactile sensitivity, we use a longer time window in the moving average step.

After both the audio and visual components are processed, the amplitudes of the audio signals are mapped to the control voltages of each haptic actuator according to the spatial tactile map. Since the intensities of haptic effects are designed to be proportional to the amplitudes of the audio signals, the last step of the pipeline is to normalize the root mean square (RMS) of tactile stimuli intensities to the intensities of the audio signals. Each individual haptic actuator's intensity I_i, proportional to the control voltage, is decided by $T_i * Amp / \sum_{i=1}^{k} T_i$ where T_i is the value of the corresponding tactile map pixel, Amp is the amplitude of the audio signal and k is the number of the tactile pixels.

Examples of the final spatial tactile mappings created by our framework are shown in Fig. 4. More details are available in the supplemental materials. Videos examples were downloaded from YouTube. Note that there are no tactile signals in one of the screenshots due to its low audio signal intensity.

3.4 Tactile Display

To evaluate the performance of our spatial tactile effects generation pipeline, we developed a 3×3 vibrotactile array to render the tactile stimuli.

3.4.1 Location. For 4D movies or other immersive media with a haptic channel, a commonly used configuration is to apply the vibrotactile display to people's backs, since users are often seated and do not need to dawn and doff a device if it is integrated into a chair. The back also provides a relatively flat, large area surface for spatial tactile feedback, which is needed given the relatively low tactile spatial acuity outside the hands, face and tongue [14]. Thus, we adopt a similar setup [18, 23] by developing a 2-D vibrotactile array inside a chair backrest cushion (as shown in Fig. 5(a).) resting against user's back to demonstrate our automatic spatial tactile effects generation results.

3.4.2 Hardware Design. The design requirements of such a vibrotactile display include sufficient tactile stimuli, minimal interference between different haptic actuators and low distracting audio noise. To generate sufficient vibration intensities, we utilize 30W haptic actuators as Fig. 5(a) depicts (Dayton Audio TT25-16 PUCK Tactile Transducer Mini Bass Shaker) in our setup. This haptic actuator can provide sufficient stimulation to the user's back which has lower tactile sensitivity than other areas commonly used for haptic effects (e.g., hand or arm). To reduce the interference among haptic actuators, our haptic actuators are attached to a lumbar support back cushion with high-density memory foam which provides a good balance between firmness and softness. While too much softness will absorb actuator vibrations and weaken the haptic effects, too much firmness will result in a cross-talk between the different

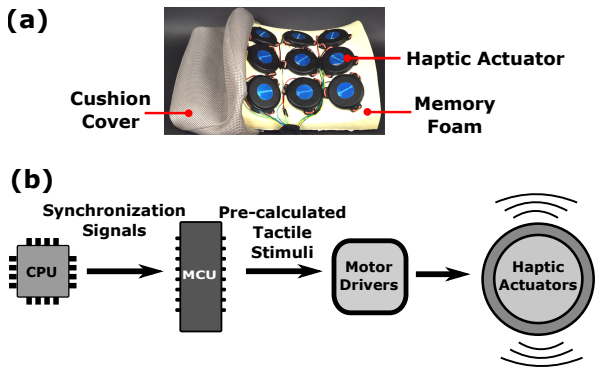

Figure 5: (a) Chair cushion with haptic actuators inside as a vibrotactile display. (b) Diagram of the tactile rendering hardware system.

haptic actuators due to longer vibration propagation distance, thus jeopardizing the localization of spatial haptic effects. Furthermore, while this actuator is essentially a voice coil, it is custom designed as a tactile transducer to minimize any distracting audio effects accompanied with the vibrations.

The number of haptic actuators and array size are also important design considerations for a vibrotactile display. Prior work [15] demonstrated that a user's back has a high tactile localization accuracy rate with a 3×3 array of vibrotactile actuators placed 60 mm apart. Same configuration has been used to enhance the video viewing experience [17]. Thus, we also arrange our haptic actuators in a 3×3 array. Since our haptic actuators have a diameter of 70 mm, we placed them with a 10.5 cm inter-tactor spacing to avoid cross talk.

Response time is another consideration for the vibrotactile display. According to Kim [16], latencies shorter than the time to play a single visual frame is in an acceptable range. Since our actuator has a frequency response range of 20 - 80 Hz which corresponds to a rise time lower than 6 ms, it is sufficient to synchronize with videos with 30 frames per second (fps) frame rate (33 ms per frame).

To control the vibrotactile display, we designed and fabricated a driver circuit. A medium-power motor driver (Pololu MAX14870) with a peak current of 1.7 A was used for each tactile transducer. A microcontroller (Teensy 3.6) was used to connect the computer and the motor drivers for tactile effects processing. Since our vibrotactile display is installed on a chair cushion rather than used as a wearable device on body (e.g., hand or arm), we did not strictly optimize the form factor of our electronic components. The primary goal of this prototype is to demonstrate the feasibility and performance of our automatic haptic effects generation pipeline.

3.4.3 Tactile Rendering Process. We developed a C# program to render the spatial tactile mappings that we obtained in Sec. 3.3 with our vibrotactile display. Since our pipeline takes approximately 5 minutes to process a 10-second video clip, we computed and saved the required tactile signals in the microcontroller offline. Up to 1000 seconds of tactile signals can be pre-loaded to our microcontroller (Teensy 3.6). When the user begins watching a video, a start signal with the current video identifier is sent to the microcontroller (Fig. 5(b)) which then triggers the haptic actuators with the pre-computed signals. A signal is sent from the CPU to the microcontroller every 100 ms to re-synchronize (Fig. 5(b)) these two subsystems. Our haptic actuators are low-noise voice coils which have a much shorter rising time of about 4 ms, considering a frequency response of 20 - 80 Hz. Since the haptic latency is within an acceptable range [16], we did not intentionally send driving signals ahead of time to compensate for the latency. This is in contrast to some prior work [18, 20] which used eccentric rotating mass (ERM) vibration motors for tactile rendering. Due to the large latency time of these actuators, the tactile stimuli needs to be sent ahead of time to keep a good synchronization with the audiovisual content.

4 USER EVALUATION OF TACTILE EFFECTS BASED ON CROSS-MODAL FEATURES

In this section, we describe our user study on evaluating the performance of the tactile effects that were automatically generated from cross-modal information as described in the previous sections. To test this, we compare our cross-modality method with the state-of-the-art method that utilizes only the visual channel [18]. We also compare videos from these methods to a baseline condition without any tactile effects. Thus, three conditions were evaluated in this user study: plain videos without any tactile effects, videos from the modified version of saliency-driven visual-based method [17] and videos from our method based on cross-modal features.

4.1 Hypotheses

H1: Our cross-modal method and the saliency-driven method will have similar performance to spatially synchronize the tactile stimuli with the audiovisual content

The saliency-driven method [18] detects motions of objects since the visual saliency is closely related to the spatial and temporal changes of the visual features. In comparison, our method computes the location of the sounding object in the scene using the method developed by Tian et al. [36]. Therefore, if both of the methods are implemented correctly, they should have comparable spatial synchronization performance for most scenes.

H2: Our cross-modal method will yield better temporal synchronization between the tactile stimuli and the audiovisual content than the saliency-driven method.

By taking both audio and visual features into consideration, our cross-modal method should be better in detecting the start of an event while the visual saliency sometimes can be misleading.

4.2 Method

The main objective of this study is to compare the different types of tactile effects: no tactile effects, tactile effects based on visual saliency-driven method, and tactile effects from our cross-modal method. In order to see how these compare across different types of videos, we presented to the participants eight videos with different scene complexity such as the number of audiovisual events and the number of objects in the scene.

4.2.1 Independent Variables. Three conditions were presented for each video: (1) plain video without any tactile effects, (2) tactile

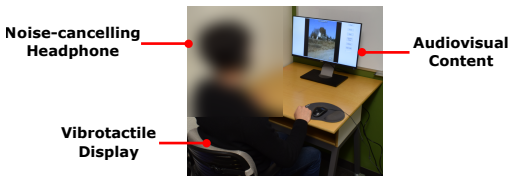

Figure 6: Participants sat against a chair cushion with vibrotactile actuators during the study. The vibrotactile device provided tactile stimuli that were synchronized to the shown videos.

effects generated by the modified saliency-driven method [18], and (3) our cross-modal method. Plain videos were used as the baseline condition to evaluate how much the spatial tactile effects augment the user experience. To provide a comparison with the state-of-the-art visual-based automatic tactile effects generation pipeline, the modified saliency-driven method [18] was chosen as the comparison condition.

The saliency-driven method [18] firstly builds a spatial saliency map for each frame by finding the visually apparent features. It also constructs a temporal saliency map by looking at frame to frame difference to emphasize the dynamic motion of the objects. However, the saliency-driven method [18] did not consider foreground motion versus background motion. If there are any abrupt camera movements in the video, it cannot avoid assigning tactile effects to the moving background even though the foreground objects (e.g. car in a racing game) should be the main focal point of the tactile rendering. This limitation will compromise the performance of the visual-only method in many videos. Thus. in our implementation of the saliency-driven tactile generation pipeline, we modified the method used by Kim et al. by adopting a neural-network-based method [13] for detecting visual saliency. The RMS of the tactile stimuli generated by the modified saliency-driven method and our method are also normalized for a fairer comparison.

4.2.2 Video Content. We downloaded and used eight video clips from YouTube for the experiment. See Table 1 for a brief summary of each video. The complexity of the scene such as the number of audiovisual events and the parameters of the objects in the scene (e.g., number, size, and motion of objects) are major factors that can influence the quality of the generated tactile stimuli. Thus, we selected example videos for the user study which covered various combinations of these factors to more thoroughly evaluate different tactile effects generation methods. All the video clips are included in the supplemental materials.

4.2.3 Study Setup. As shown in Fig. 6, a 24 inch LCD display was used to present the videos while a noise cancelling headphone (Audio-Technica ATH-ANC7b) was used to deliver the audio channel as well as preventing participants from hearing the tactile transducer noises. Participants were seated approximately 65 cm from the display. A chair cushion was equipped with the 3×3 tactile transducers inside as shown in Fig. 5(a), which was used to provide the tactile effects to the participant's back. Participants were asked to refrain from wearing thick or multiple layers of clothing to ensure adequate delivery of the tactile effects.

4.2.4 Participants. We recruited 20 participants (12 M, 8 F) from our institution. Ages ranged from 21 to 44 (M = 25.8, SD = 5.1). Participants were compensated 15 USD for their participation, and the experiment generally lasted for an average of 40 minutes.

4.2.5 Design and Procedure. We used a one-factor within-subject design where the independent variable was the tactile effects presented throughout a video. Every participant experienced 3 (tactile effects) ×8 (videos) = 24 trials in a random order. A practice session demonstrating the capabilities of the system was carried out before the formal user study in order to reduce the "novelty effect" for the participants. After every eight trials, participants were given two minute long breaks.

After each trial, participants filled out a modified version of the quality of experience questionnaire (See Table 2). We designed this around the concept of **Presence**. Witmer and Singer [38] determined four factors for presence including (1) Control, (2) Sensory, (3) Realism, and (4) Distraction. Since the participants passively received the tactile effects, the "Control" category is not relevant for this study. We also excluded the "Realism" item as our tactile effects are designed to enhance the audiovisual experience which does not necessarily have to replicate the real world. The "Sensory" factor assesses how each modality is solicited in the experience and the "Distraction" section was chosen to help to evaluate if our tactile effects are distracting to the participants. Thus, these two items were included in the questionnaire. We also added two additional items to the QoE questionnaire including the "Novelty" and "Immersion" questions which are frequently used for evaluation of tactile effects accompanied with multimedia [18, 20]. We developed and asked a single question for each category as shown in Table 2.

For sessions with tactile effects, three additional questions were added to the questionnaire to better understand whether the haptic content matched the video spatially and temporally (Table 2). All question use a 0-10 Likert Scale, where 10 represents a strong agreement with the statement and 0 indicates strong disagreement.

4.2.6 Analysis. To examine the effects of the three conditions (No tactile effect, tactile effects from the saliency-driven method, and tactile effects based on cross-modal method), a Mauchly's Test of Sphericity followed by a one-way repeated measures ANOVA were performed for Q1-Q4 in Table 2. If Mauchly's Test of Sphericity is violated, a Greenhouse-Geisser correction was used to calculate the F and p values from ANOVA. Bonferroni-corrected post-hoc tests were used to determine which pairs of means are significantly different. For Q5-Q7, a paired t-test was conducted to compare the tactile effects from the saliency-driven and cross-modality methods.

4.3 Results and Discussion

4.3.1 Overall Results and Trends. As shown in Fig. 7,our cross-modal method significantly outperformed the saliency-driven method for all the items in the QoE (Q1 - Q4, $p < 0.001$). The plain videos were not statistically different from the videos based on saliency-driven condition in terms of Sensory, Novelty and Immersion, while the videos from saliency-driven method were significantly worse than the plain videos in terms of Distraction ($p < 0.001$). Although the videos from saliency-driven method were not rated higher than

Table 1: Summary of the videos used for the user study

Movie	M1	M2	M3	M4	M5	M6	M7	M8
Summary	flying airplane	moving truck	train collision	wall clock	barking dog	serene forest	log sawing	busy street
Length	10s	10s	10s	10s	20s	10s	10s	20s
Excerpt								

Table 2: Questionnaire used during the user study

NO.	Factor	Question	Scale
Q1	Sensory	The multi-sensory interaction of the system helped with its content delivery.	0-10
Q2	Distraction	I can focus on the content without being distracted by the delivery methods (e.g. display, headphone and haptic actuators).	0-10
Q3	Novelty	I found this system interesting to use.	0-10
Q4	Immersion	I was immersed in the movie.	0-10
Q5		The vibrations were spatially matched with the movie.	0-10
Q6		The vibrations were temporarily matched with the movie.	0-10
Q7		It was straightforward to understand why there were vibrations.	0-10

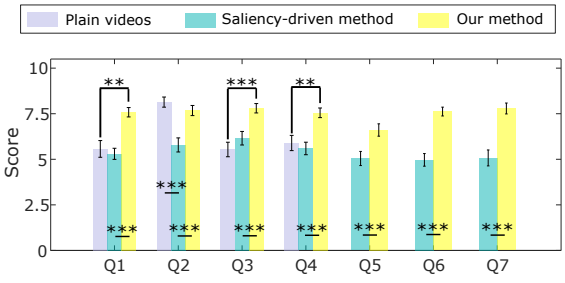

Figure 7: The aggregated results for all of the eight examples are shown. * p <0.05, ** p<0.01, *** p<0.001. The results compare performance of the baseline condition (i.e., plain videos without tactile effects), the modified saliency-driven method [18] and our method for the seven questionnaire items (Q1-Q7, See Table 2.)

the plain videos in aggregate, it was more effective in specific scenarios, which we will discuss in the following paragraphs.

Between the videos with tactile effects from our cross-modal method and plain videos, there were significant differences ($p < 0.01$) in terms of the Sensory, Novelty and Immersion categories. Our method was rated with a lower mean score than plain videos in the Distraction item (Q2), albeit in a statistically insignificant manner. The results demonstrate that for various examples we tested, our method helped to improve user experience.

Tactile questions (Q5 - Q7) in Fig. 7 were aimed to evaluate how well the tactile effects match the audiovisual content. The overall results for eight videos in Fig. 7 showed that our cross-modal method performs significantly better ($p < 0.001$) than the modified saliency-driven method in terms of synchronizing with the videos both spatially (Q5 in Table 2) and temporally (Q6 in Table 2). Participants also were able to better understand the rationale behind the tactile effects (Q7 in Table 2).

4.3.2 In-depth Analysis of Each Video. For each video, we analyze the performance of the three conditions. We apply the same analysis as described in Sec. 4.3.1.

As we described in Sec. 1, M1 depicts a flying airplane which is a challenging task for an algorithm as they need to track a highly dynamic object. Although our method performed better than the saliency-driven method in the aggregated results, Fig. 8(b) M1 shows that our method was not significantly better than the saliency-driven method. This is within our expectation since the saliency-driven method is more specialized in tracking moving objects. This result confirms that our method based on cross-modality information is as good as the visual-only method for localization.

M2 is an example with more temporally complex audiovisual events (e.g. a truck honking). Results in Fig. 8(b) M2 show that our method is significantly better ($p < 0.01$) than the saliency-driven method for temporal synchronization (Q6), indicating that cross-modality information is more effective than pure visual features in temporal synchronization with the events in the scene. Although our method obtained a higher mean score for spatial synchronization (Q5), there was no statistical significant difference, which suggest that both methods have comparable performance for spatially tracking a large moving object.

M3 displayed a train running into a truck, a more complex event than M2. It is important to generate the spatial tactile effects synchronized with the collision event. Although the collision between the two moving objects caused a crowd of dust which may potentially be detected by a visual-only method, it is more accurate to locate the collision with the audio information. This is demonstrated in Fig. 8(b) M3 in which our method was significantly better ($p < 0.001$) in all of the tactile-related questions (Q5-Q7). In addition, our method is rated significantly higher than the plain video condition in terms of Sensory and Novelty($p < 0.05$), and higher than the saliency-driven method for Immersion ($p < 0.05$) as shown in Fig. 8(a) M3. Results of M3 show that our cross-modal method is able to temporally track an event with a higher precision than the visual-based method.

M4 mainly tested the algorithms' ability to generate spatial tactile effects for discrete audiovisual events (i.e., clock ticking) in a relatively tranquil scene. For all the items in QoE, our method was rated significantly higher ($p < 0.01$) than the saliency-driven method. In addition, our method significantly outperformed ($p < 0.001$) the saliency-driven method for all tactile-related questions (Fig. 8(b)). This is mainly because our method leveraged the audio features, thus tracking the ticks more accurately.

M5 evaluated the algorithms' performance to both spatially and temporally track multiple audiovisual events. As Fig. 8(a) depicts, for the Sensory, Novelty and Immersion items, our method was

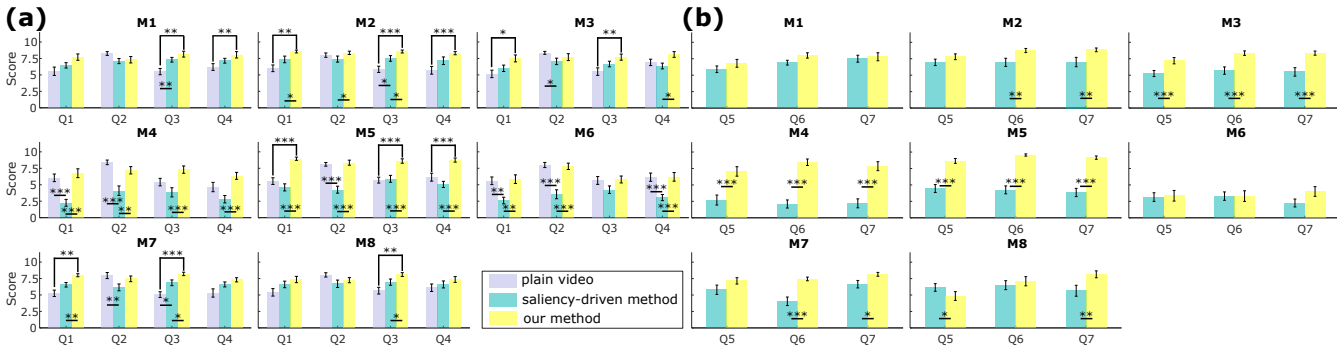

Figure 8: (a) Subjective ratings for QoE. (b) Subjective ratings for tactile related questions. Standard error bar is shown for each measurement. * p <0.05, ** p<0.01, *** p<0.001. The results compare performance of plain videos, the modified saliency-driven method [18] and our method for the seven questionnaire items (Q1-Q7, See Table 2.)

significantly better ($p < 0.001$) than both the plain video condition and the saliency-driven method. As shown in Fig. 8(b), our method significantly outperformed ($p < 0.001$) the saliency-driven method both for spatially and temporally matching tactile effects with the audiovisual content. While the saliency-driven method generated a more continuous tactile stimuli due to the movement of the dog in the scene, our method took the audio features into account, thus was able to emphasize the tactile effects on the more semantically salient audiovisual events (e.g. dog barking).

M6 is a tranquil scene with a person walking in a serene winter forest. It mainly tested whether the algorithms can generate appropriate level of tactile effects when there are no significant events in the scene. As shown in Fig. 8(a), both the plain video condition and our method were significantly better ($p < 0.01$) than the saliency-driven method for the Sensory, Distraction and Immersion items. Our method generated tactile stimuli with intensities proportional to the intensities of the audio. Thus, the amplitudes of the tactile effects were minimal in a quiet scene. In contrast, the saliency-driven method produced pronounced tactile effects.

M7 was used to compare the algorithms' ability to analyze the activities in the scene and render tactile stimuli for the proper scene semantics. From Fig. 8(a) we can observe that in terms of the Sensory and Novelty items, our method was significantly better than the plain video condition ($p < 0.01$) and the saliency-driven method ($p < 0.05$). Our method was also significantly better than the saliency-driven method ($p < 0.001$) in terms of the temporal synchronization. Since our method took the audio features into account when generating tactile effects, it is able to stop right when person stops sawing in the scene

M8 is an example with a much higher spatial complexity. From Fig. 8(b), we can observe that the saliency-driven method significantly outperformed ($p < 0.05$) our method when spatially matching the tactile stimuli to the audiovisual content. Our method performed less well in this example due to one of the limitations of our sound source localization algorithm. If there are multiple sounding objects from the same category (e.g. a fleet of cars) in the scene, our sound source localization model is not smart enough to determine which of the objects is the main sounding source. Thus, tactile effects can be assigned to any one of the objects in the same category.

From the examples above, we can conclude that our cross-modal method outperformed the saliency-driven method in terms of both the QoE and the synchronization of tactile effects. We infer that the cross-modality features that our method used helped to improve synchronization of the spatial tactile stimuli with the audiovisual content. Although the saliency-driven method and our method obtained comparable results for spatially tracking moving objects and adding tactile effects to the desired location in some video examples (Fig. 8(b) M1, M2, M6, M7, M8) as we hypothesized in *H1*, our method is more effective in temporally synchronizing tactile effects to the examples (Fig. 8(b) M2, M3, M4, M5, M7) confirming our hypothesis *H2*. Thus, cross-modal features allow algorithm to provide a more spatiotemorally synchronized tactile effects and improve the overall user experience compared to methods based only on the visual features.

5 LIMITATIONS AND FUTURE WORK

One of the limitations of our current pipeline is that the sound source localization model we used [36] can only identify 28 distinct event categories. If none of the 28 events are present in a video, our model will generate a random heatmap for sound localization which substantially diminishes the user experience. Thus, we only used 10 - 20 second long video clips with events previously trained for our model. This constraint allows a more meaningful comparison for our user study. However, we can overcome this limitation in the future by collecting and training the model with larger, more diverse video datasets.The sound source localization model also has a limitation dealing with videos with multiple sounding source in the same scene, especially when they are from the same category (e.g. a fleet of cars). Although the model will be able to extract features from the audio signals and assign one or several cars as the sounding objects, the real sounding source may be a different car within the scene. This limitation can potentially be resolved in the future by training the model on videos with stereo audio channels to help understand the locations of the sounding objects.

We used a 3×3 array of haptic actuators in this paper as it has been shown to be the appropriate resolution for back of a person [15]. However, if the vibrotactile display is to be mounted on different body parts of a user with a higher tactile spatial resolution, a vibrotactile display with a higher resolution will be necessary.

Ultimately, the goal of automatic tactile content authoring method is to achieve a similar level of performance as a haptic designer. In the future, we plan to compare the tactile effects by our cross-modal method and those manually authored by a haptic designer.

6 CONCLUSIONS

We developed a framework to automatically generate spatial tactile effects based on cross-modality features from a video. First, neural networks are used to separate diegetic sound information which is then used to locate the sounding object in the scene. The intensity of the diegetic audio is translated to intensity of the tactile stimuli while the probability heatmap of the sounding object is mapped to spatial distribution of the tactile effects. Using an array of 3×3 haptic actuators on the back of the users, we conducted a human subject experiment to evaluate and compare videos with tactile effects from our cross-modal method, a modified version of the saliency-driven visual-based method [18], and videos without any tactile effects. The study results demonstrate that the spatial tactile effects generated by our cross-modal framework are more promising in providing spatiotemorally synchronized and immersive content than those generated based on visual features only.

REFERENCES

[1] Andy Brady, Brian MacDonald, Ian Oakley, Stephen Hughes, and Sile O'Modhrain. 2002. Relay: a futuristic interface for remote driving. In *proceedings of EuroHaptics*. 8–10.

[2] Jongeun Cha, Mohamad Eid, and Abdulmotaleb El Saddik. 2009. Touchable 3D video system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 5, 4 (2009), 29.

[3] Jongeun Cha, Yo-Sung Ho, Yeongmi Kim, Jeha Ryu, and Ian Oakley. 2009. A Framework for Haptic Broadcasting. *IEEE MultiMedia* 16, 3 (2009), 16–27.

[4] Jongeun Cha, Yongwon Seo, Yeongmi Kim, and Jeha Ryu. 2007. An authoring/editing framework for haptic broadcasting: passive haptic interactions using MPEG-4 BIFS. In *Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'07)*. IEEE, 274–279.

[5] Angela Chang and Conor O'Sullivan. 2005. Audio-haptic feedback in mobile phones. In *CHI'05 extended abstracts on Human factors in computing systems*. ACM, 1264–1267.

[6] Seungmoon Choi and Katherine J Kuchenbecker. 2012. Vibrotactile display: Perception, technology, and applications. *Proc. IEEE* 101, 9 (2012), 2093–2104.

[7] Fabien Danieau, Julien Fleureau, Audrey Cabec, Paul Kerbiriou, Philippe Guillotel, Nicolas Mollet, Marc Christie, and Anatole Lécuyer. 2012. Framework for enhancing video viewing experience with haptic effects of motion. In *2012 IEEE Haptics Symposium (HAPTICS)*. IEEE, 541–546.

[8] Fabien Danieau, Anatole Lécuyer, Philippe Guillotel, Julien Fleureau, Nicolas Mollet, and Marc Christie. 2012. Enhancing audiovisual experience with haptic feedback: a survey on HAV. *IEEE transactions on haptics* 6, 2 (2012), 193–205.

[9] Hend M Hamed and Rehab A Hema. 2009. The Application of the 4-D Movie Theater System in Egyptian Museums for Enhancing Cultural Tourism. *Journal of Tourism* 10, 1 (2009).

[10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[11] Ali Israr and Ivan Poupyrev. 2011. Tactile brush: drawing on skin with a tactile grid display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2019–2028.

[12] Sungjune Jang, Lawrence H Kim, Kesler Tanner, Hiroshi Ishii, and Sean Follmer. 2016. Haptic edge display for mobile tactile interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3706–3716.

[13] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. 2018. Deepvs: A deep learning based video saliency prediction approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 602–617.

[14] Lynette A Jones, Brett Lockyer, and Erin Piateski. 2006. Tactile display and vibrotactile pattern recognition on the torso. *Advanced Robotics* 20, 12 (2006), 1359–1374.

[15] Lynette A Jones and Kathryn Ray. 2008. Localization and pattern recognition with tactile displays. In *2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. IEEE, 33–39.

[16] Myongchan Kim. [n.d.]. Saliency-driven Real-time Tactile Effects Authoring. ([n. d.]).

[17] Myongchan Kim, Sungkil Lee, and Seungmoon Choi. 2012. Saliency-driven tactile effect authoring for real-time visuotactile feedback. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 258–269.

[18] Myongchan Kim, Sungkil Lee, and Seungmoon Choi. 2013. Saliency-driven real-time video-to-tactile translation. *IEEE transactions on haptics* 7, 3 (2013), 394–404.

[19] Yeongmi Kim, Jongeun Cha, Ian Oakley, and Jeha Ryu. 2009. Exploring tactile movies: An initial tactile glove design and concept evaluation. *Ieee Multimedia* (2009).

[20] Yeongmi Kim, Jongeun Cha, Jeha Ryu, and Ian Oakley. 2010. A tactile glove design and authoring system for immersive multimedia. *IEEE MultiMedia* 17, 3 (2010), 34–45.

[21] Jaebong Lee and Seungmoon Choi. 2012. Evaluation of vibrotactile pattern design using vibrotactile score. In *2012 IEEE Haptics Symposium (HAPTICS)*. IEEE, 231–238.

[22] Jaebong Lee and Seungmoon Choi. 2013. Real-time perception-level translation from audio signals to vibrotactile effects. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2567–2576.

[23] Paul Lemmens, Floris Crompvoets, Dirk Brokken, Jack Van Den Eerenbeemd, and Gert-Jan de Vries. 2009. A body-conforming tactile jacket to enrich movie viewing. In *World Haptics 2009-Third Joint EuroHaptics conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. IEEE, 7–12.

[24] Kevin A Li, Timothy Y Sohn, Steven Huang, and William G Griswold. 2008. Peopletones: a system for the detection and notification of buddy proximity on mobile phones. In *Proceedings of the 6th international conference on Mobile systems, applications, and services*. ACM, 160–173.

[25] Nadia Magnenat-Thalmann and Ugo Bonanni. 2006. Haptics in virtual reality and multimedia. *IEEE MultiMedia* 13, 3 (2006), 6–11.

[26] Eunji Oh, Minkyoung Lee, and Sujin Lee. 2011. How 4D Effects cause different types of Presence experience?. In *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*. ACM, 375–378.

[27] Sile O'Modhrain and Ian Oakley. 2004. Adding interactivity: active touch in broadcast media. In *12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2004. HAPTICS'04. Proceedings*. IEEE, 293–294.

[28] Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 631–648.

[29] Andrew Owens and Alexei A. Efros. 2018. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. https://github.com/andrewowens/multisensory.

[30] Sile O'Modhrain and Ian Oakley. 2003. Touch TV: Adding feeling to broadcast media. In *European Conference on Interactive Television: from Viewers to Actors*. 41–17.

[31] Abdur Rahman, Abdulmajeed Alkhaldi, Jongeun Cha, Abdulmotaleb El Saddik, et al. 2010. Adding haptic feature to YouTube. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1643–1646.

[32] Robert J Schalkoff. 1989. *Digital image processing and computer vision*. Vol. 286. Wiley New York.

[33] Jongman Seo, Reza Haghighi Osgouei, Soon-Cheol Chung, and Seungmoon Choi. 2016. Vibrotactile Rendering of Gunshot Events for 4D Films. In *IEEE Haptics Symposium*, Vol. 2.

[34] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[35] Colin Swindells, Seppo Pietarinen, and Arto Viitanen. 2014. Medium fidelity rapid prototyping of vibrotactile haptic, audio and video effects. In *2014 IEEE Haptics Symposium (HAPTICS)*. IEEE, 515–521.

[36] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 247–263.

[37] Markus Waltl. 2010. *Enriching multimedia with sensory effects: annotation and simulation tools for the representation of sensory effects*. VDM Verlag.

[38] Bob G Witmer and Michael J Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence* 7, 3 (1998), 225–240.