Towards Context-aware Automatic Multimodal Haptic Effect Generation for Home Theatre Environments

Yaxuan Li

Department of Electrical & Computer Engineering
McGill University
Montréal, Québec, Canada

May 2022

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

2022/05/18

Abstract

The integration of haptic technology into entertainment systems, such as in Virtual Reality and 4D cinema, enables novel immersive experiences for users, and its increasing prevalence drives the demand for efficient haptic authoring systems. Over 80,000 4D cinema chairs are now in use in 731 auditoriums across 65 countries, and the market is still growing. Taking one 4D movie company as an example, 4DX had its best year ever at the worldwide box office in 2019, surpassing 2018's record of $286 million by 12 percent. However, designing haptic effects manually is very expensive and time-consuming. In this thesis, we propose an automatic, mutlimodal pipeline for vibrotactile content creation that substantially improves user hapto-audiovisual (HAV) experience based on contextual audio and visual content from movies. The algorithm extracts significant features from video files to generate corresponding haptic stimuli, which are rendered on a low-cost system of nine actuators attached to a viewing chair. Using this system, we conducted a user study ($n = 16$) and quantified user experience according to the sense of immersion, preference, harmony, and discomfort. The results indicate that the haptic patterns generated by our algorithm complement the movie content, and provide an immersive and enjoyable HAV user experience. We also investigated the incorporation of other haptic modalities by manipulating temperature and airflow. To create thermal effects, we examined the scene's warmth and coolness levels using extracted colours and mapped them to temperature. To create airflow effects, we used optical flow analysis to identify frames with high-intensity movement. For frames with rapid changes, we employed depth estimation to determine the action of the scene's characters and author airflow effects for them to create motion feeling for users while watching movies. In the end, our study suggests that the pipeline can facilitate the efficient creation of 4D effects and could therefore be applied to improve the viewing experience in home theatre environments.

Résumé Scientifique

L'intégration de la technologie haptique dans les systèmes de divertissement, comme dans la réalité virtuelle et le cinéma 4D, permet de nouvelles expériences immersives pour les utilisateurs, et sa prévalence croissante stimule la demande de systèmes de création haptiques efficaces. Plus de 80 000 fauteuils de cinéma 4D sont désormais utilisés dans 731 auditoriums dans 65 pays, et le marché continue de croître. Prenant l'exemple d'une société de cinéma 4D, 4DX a connu sa meilleure année au box-office mondial en 2019, dépassant le record en 2018 (286 millions de dollars) de 12%. Cependant, la conception manuelle d'effets haptiques est très coûteuse et prend beaucoup de temps. Dans cette thèse, nous proposons un pipeline automatique et multimodal pour la création de contenu ressenti tactile qui améliore considérablement l'expérience hapto-audiovisuelle (HAV) de l'utilisateur basée sur le contenu audio et visuel des films. L'algorithme extrait les caractéristiques significatives des fichiers vidéo pour générer des stimuli haptiques correspondants, qui sont créés en utilisant un système à faible coût composé de neuf actionneurs attachés à une chaise de visualisation. À l'aide de ce système, nous avons mené une étude utilisateur (n = 16) et quantifié l'expérience utilisateur en fonction du sentiment d'immersion, de préférence, d'harmonie et d'inconfort. Les résultats indiquent que les modèles haptiques générés par notre algorithme complètent le contenu du film et offrent une expérience utilisateur VHA immersive et agréable. Nous avons également étudié l'incorporation d'autres modalités haptiques en manipulant la température et le flux d'air. Pour créer des effets thermiques, nous avons examiné les niveaux de chaleur et de fraîcheur de la scène à l'aide de couleurs extraites et les avons mappés à la température. Pour créer des effets de flux d'air, nous avons utilisé l'analyse de flux optique pour identifier les cadres avec un mouvement de haute intensité. Pour les images avec des changements rapides, nous avons utilisé l'estimation de la profondeur pour déterminer l'action des personnages de la scène et créer des effets de flux d'air pour qu'ils créent une sensation de mouvement pour les utilisateurs tout en regardant des films. En fin de compte, notre étude suggère que le pipeline peut faciliter la création efficace d'effets 4D et pourrait donc être appliqué pour améliorer l'expérience de visionnage dans les environnements de cinéma maison.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Jeremy Cooperstock, who always supported me and guided me throughout my graduate life. He gave me precious insights, thoughtful guidance, patience in academic writing, and freedom in pursuing my research ideas. The meetings and discussions with him always inspired me to look outside the box from various viewpoints to develop a complete criticism.

Moreover, I want to thank all the laboratory members of Shared Reality Lab who gave me feedback, comments, and constructive suggestions. Particularly, I would like to thank Yongjae Yoo and Antoine Weill-Duflos, who actively helped with the projects. My special thank you to David Marino, Clara Ducher, Hyejin Lee, Preeti Vyas, Marc Demers, and Yassine Bouanane, who warmly welcomed me when I first came to Canada. They are my best support all the time. They gave me so much help in every small aspect of life, and we have established a deep friendship. A special thank you to Max Henry, who helped me proofread and edit the thesis. Moreover, many thanks to the examiner of my thesis, Prof. Narges Armanfard.

From the bottom of my heart, I would like to express my most profound appreciation to my family and friends who support me. Without my family, I could not have had the opportunity to pursue my dream and come to McGill. More importantly, a special thank you to Prof. Cheng Zhang, Prof. David Lindlbauer, and Dr. Shibo Zhang, who gave me so many insightful suggestions and guided me in growing from a naive green hand to an independent researcher. Finally, I also want to thank my friends, Hang Zhang, Jiajie Li, Yuanzhe Gong, Chen He, Huiyuan Yang, Tuochao Chen, Songyun Tao and many others, who were supportive and brought so much happiness to my life. Thanks to my treasure, Guangyi Zhang, who gave me so much support and care that kept me on the right track, especially facing big challenges. I want to give a big hug to all my cheerful companions who shared me ideas and motivated me to go beyond. You gave me light and warmth in the most challenging time of my life. I always hope to make some cool technology. I want to give my deepest gratitude to all the friends who support me in this long journey to achieving my dream.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advent of four-dimensional (4D) theatre technology has had a profound effect on viewer experience. Vestibular (motion) feedback, bursts of air, mists, thermal stimuli, and vibrotactile effects are all now commonly employed in immersive movie-watching environments. Such components enrich the movie watching experience through enhanced realism and greater immersion [1]. There is a growing need for content that can take advantage of the 4D capabilities of movie theatres, without necessitating undue production costs.

This demand has likely grown as a result of the COVID-19 pandemic, which has further pushed consumption of cinema content to the home theatre environment, supported by a variety of streaming services such as Netflix, Amazon Prime, and Disney+ [2]. For home theatre setups, an inexpensive, but highly limited option is the use of haptic chairs that directly convert the audio track into vibrotactile feedback through haptic actuators installed in the chair. While more sophisticated, semi-automatic methods of haptic authoring have been explored in both academia and industry, the effort remains mostly manual, dependent on skilled industry experts, time-consuming, and costly [3]. The typical process takes three designers approximately 16 days to produce the effects for a feature-length 4D film [3]. This cost is an obvious disincentive to the creation and distribution of 4d movies, and motivates exploration of more efficient authoring methods.

The main contribution of this work is an automatic haptic effect generation algorithm, designed primarily for the needs of the 4D home theatre content experience, along with a ($n = 16$) user study to evaluate its performance on six movie clips across a variety of genres including sci-fi, action, cartoon, family-comedy, and horror/thriller. The proposed algorithm, illustrated in Figure 1.1, generates haptic effects based on high-level features extracted from the audio-visual stream. The algorithm determines information about the high-level context of the scene such as tonal properties (a composite of the mood, narrative, and specific genre features); location of events and characters; dynamics of the atmosphere; and intensity of action. We used metrics that reflect salient perceptual characteristics of the audio and visual modalities. These include four psychoacoustic parameters, used in the audio analysis to determine the intensity of vibration (Figure 1.1(b)) and estimates of visual saliency for every frame (Figure 1.1(c)), which are used to determine the spatiotemporal distribution of vibrotactile effects (Figure 1.1(d)). The resulting haptic effects are delivered through a vibrotactile chair with nine actuators (Figure 1.1(e)). Besides the vibrotactile effects generation algorithm, we further explored the possibilities of incorporating additional haptic modalities including thermal and airflow effects.

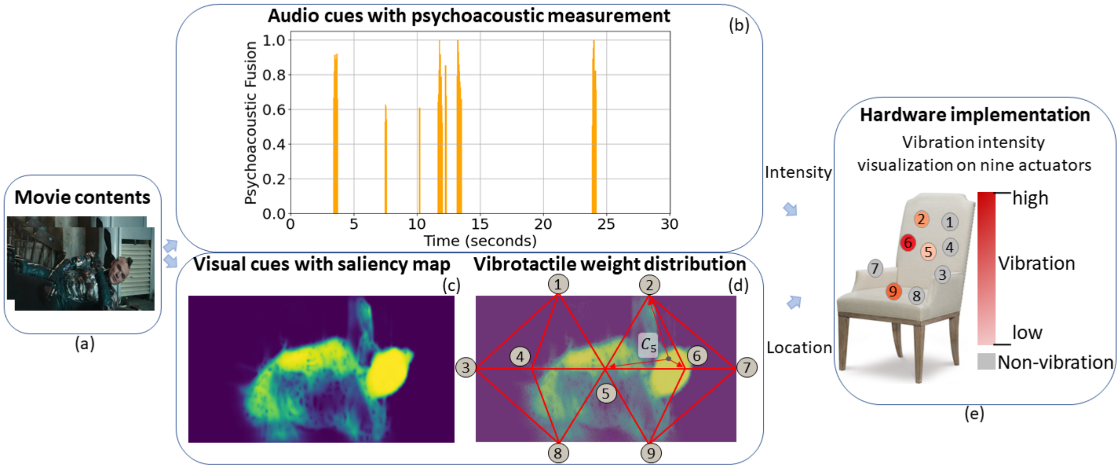


Fig. 1.1: Overview of the haptic effects generation algorithm. With movie content as input, the fusion of psychoacoustic measurements indicating human perception of sound determines vibration intensity. In parallel, the saliency map suggests the event location in the scene, and assigns weights to actuators for the generation of the spatiotemporal vibrotactile effects.

This thesis is organized in six chapters. The present chapter introduces the problem, background, objectives and contributions of this work. Chapter 2 gives an overview of the literature and the background information necessary to comprehend the subsequent chapters. Chapter 3 introduces the analysis for extracting features from audiovisual modalities and information fusion for generating vibrotactile patterns. Chapter 4 presents algorithm deployment on hardware and the user study for evaluation. Chapter 5 investigates creating multi haptic modalities home theatre with thermal and airflow effects. Chapter 5 investigates incorporating thermal and airflow effects into multimodal home theatre chair. Finally, the thesis concludes with Chapter 6, which summarizes the findings of this work and discusses future work possibilities for this research.

Chapter 2

Related Work

2.1 Haptics for Enhancing the Viewing Experience

Inclusion of multisensory information has been demonstrated to greatly enhance user experience in terms of immersion, flow, absorption, and engagement in a range of entertainment activities, including virtual and augmented reality, and gaming. Incorporating multisensory information into audio-visual entertainment has been demonstrated to greatly enhance user experience in terms of immersion, flow, absorption, and engagement. The effect has been shown in gaming and virtual and augmented reality settings [4, 5, 6, 7, 8, 9]. There is also a growing interest in applying similar effects in 4D film, enhancing the cinematic viewing experience beyond stereoscopic (3D) cinema.

Numerous companies manufacture 4D movie "platforms," which provide leg tickling, vestibular (motion) effects, thermal stimulation and other haptic effects for use in movie theatres and theme parks, for example, D-Box Technologies, MediaMation (MX4D), and CJ 4D Plex (4DX). With respect to the inclusion of haptic effects, Danieu *et al.* presented a perceptually organized review of HAV devices [1], describing the three types of haptic feedback: 1) tactile feedback with temperature stimuli [10, 11], vibration [12, 13, 14, 15, 16, 11, 17], and pressure [18, 19, 20, 21]; 2) kinesthetic feedback resulting from limb position movement [22, 23] and force [24, 23, 25, 26, 27]; and 3) proprioception or vestibular feedback arising from body motion [24, 28]. Gaw *et al.* and Delazio *et al.* demonstrated the feasibility of using force feedback to increase viewers' ability to understand the presented media content [22, 29] and Dionisio *et al.* highlighted the effectiveness of thermal stimuli in VR [10].

2.2 Authoring of Vibrotactile Stimuli

To achieve a high-quality overall experience, it is essential to create haptic effects that match the audiovisual content of the scene. At present, this relies predominantly on the intuition and experience of expert designers, working manually with effects authoring platforms; for example, Macaron [30], a web-based vibrotactile effects editor, Vibrotactile Score [31], based on composition of patterns of musical scores, H-Studio [32], Immersion's Haptic Studio,[1] HFX Studio [33], posVibEditor [34] and SMURF [35].

As noted previously, use of such haptic authoring tools tends to be costly and time-consuming. To address this challenge, automatic haptic effects generation methods have been developed, which typically produce haptics from audio properties. An early effort by Chafe tried to combine the vibrotactile sensation with musical instruments [36]. Similarly, Chang and O'Sullivan used low-frequency components, extracted from audio files, to provide vibration on a mobile phone using a multi-function transducer as an actuator [37]. For the creation of vibrotactile effects in games, Chi *et al.* found that simple mapping of the audio according to frequency bands resulted in excessive ill-timed vibrations, causing feelings of haptic numbness and annoyance [38]. To avoid this problem, they established key moments when vibrations should occur by identifying target sounds in real-time using an acoustic similarity measurement. Contrary to other researchers, who employed frequency bands and filters to extract suitable audio features for haptic conversion [39, 40], Lee and Choi [41] translated auditory signals to vibrotactile feedback by bridging the perceptual characteristics of loudness and roughness. Their algorithm calculates the level of two perceptual characteristics and converts them into the intensity of vibrotactile effects.

Another direction for haptic authoring employs cues from the visual stream to drive the generation of haptic effects. For example, Rehman *et al.* took the graphical display of a soccer game, divided the field into five areas, and mapped each to an area-based vibration pattern to indicate the position of the ball [13]. Similarly, Lee *et al.* mapped the trace of a soccer ball to vibration effects on an array of 7 × 10 actuators mounted on the forearm, enabling a spatiotemporal vibrotactile mapping of the location of the ball in the game [17]. Moreover, the use of computer vision and physical motion modeling techniques has been investigated to extract key elements of the scene or estimate camera motion from first-person videos. These cues are applied to create corresponding haptic effects, including vestibular (motion) [3], force [24], and vibrotactile feedback [28, 42, 43].

[1] `https://www.immersion.com/technology/#hapticdesign-tools`

More recent research has demonstrated impressive performance using Generative Adversarial Networks to produce tactile signals from images of different materials [44, 45, 46, 47]. Although much research has been conducted on the generation of haptics from either audio or visual data, we are the first to consider audiovisual information on a perceptual level simultaneously and to automatically develop context-aware automatic vibrotactile effects.

Chapter 3

Context-aware Multimodal Analysis for Vibrotactile Effect Generation and Information Fusion

©2021 ACM Reprinted, with permission, from

Yaxuan Li, Yongjae Yoo, Antoine Weill-Duflos, and Jeremy Cooperstock. 2021. Towards Context-aware Automatic Haptic Effect Generation for Home Theatre Environments. In Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology (VRST '21). Association for Computing Machinery, New York, NY, USA, Article 13, 1–11. DOI: https://doi.org/10.1145/3489849.3489887

Preface

This chapter presents work that was published in the 27th ACM Symposium on Virtual Reality Software and Technology. In this chapter, we introduce the details of the algorithm used for feature extraction and analysis of audio and visual information. Our pipeline makes use of four auditory parameters as described in Section 3.1.1 and a visual saliency map mentioned in Section 3.1.2.

In Section 3.1.1, we assess the perceived quality of the film's audio track using the psychoacoustic criteria of sharpness, booming, low-frequency energy, and loudness. We employ these auditory qualities to calculate the vibrotactile actuation's time and amplitude.

In Section 3.1.2, we compute the saliency map for each frame using the visual stimuli, which defines the area of predicted visual interest. The saliency map is used to effectively communicate the position and direction of vibrotactile rendering. Using the valuable features from both auditory and visual modalities, we combine these data to create vibrotactile effects, as described in Section 3.2.

Author's Contribution

This work is the result of collaboration. Section 3.1.1 describes code written by Dr. Yongjae Yoo, which provides two functions used for the calculation of psychoacoustic variables, as described in Zwicker and Fastl [48]. The work described in Section 3.2 relates to the ideas proposed by the author, who also implemented the associated functions and conducted the experiments. Dr. Antoine Weill-Duflos provided valuable feedback in the process of algorithm design. Prof. Cooperstock supervised the research, gave critical feedback and edited the paper.

3.1 Automatic Haptic Effect Authoring Algorithm

While previous research in either academia or industry endeavors to generate automatic vibrotactile effects based on a mapping from audio or visual contents, we propose instead a multimodal algorithm that integrates the contextual information from *both* audio and visual streams. We also take into account the perceptual psychology of the expected audience to generate reasonable patterns that align automatically with the movie contents.

Our pipeline employs four acoustic measures (Section 3.1.1) and the visual saliency map (Section 3.1.2). For the former, we use the psychoacoustic parameters of sharpness, booming, low-frequency energy and loudness to quantify the perceptual quality of the movie's audio track. The timing and amplitude of the vibrotactile actuation are determined from these auditory features. For example, gunshot sounds in a fight scene or booming sounds in racing or piloting scenes create an atmosphere of intense emotions and, ideally, increase the immersion of the audience. For the latter, we estimate the saliency map from the visual stimuli for each frame, which identifies the region of expected visual interest [49]. As shown in a previous study [42], the map can provide effective location and direction information for vibrotactile rendering; however, it cannot automatically determine the magnitude of the haptic parameters. Finally, we integrate the information from both audio and visual features to generate the vibrotactile effects. These are presented through an array of nine vibration motors, installed in a chair. Although designed for spatiotemporal vibrotactile feedback in this study, our pipeline for automated haptic effects authoring could likewise be applied to other haptic modalities such as temperature, airflow or impact.

3.1.1 Audio Modality Analysis

The sounds corresponding to an event are especially significant since the audio and haptic modalities share several common characteristics, such as the physical propagation principle and perceptual properties (roughness, consonance, etc.). In movie clips, sound often accompanies important events, such as the one depicted in Figure 3.1. The importance of sound cues to human attention motivated the work of Evangelopoulos *et al.* who formulated measures of audiovisual saliency by modelling perceptual and computational attention to such cues [50]. Huang and Elhilali further demonstrated that auditory salience is also context dependent [51], since the salient components vary according to specific characteristics of the scene. Accordingly, to achieve an effective multimedia user experience, haptic effects created from the auditory modality should consider characteristics of the source content, such as the genre of material [52, 28] The generation of haptic effects that are congruent to the given audiovisual contents has been studied for decades. A common approach is the extraction of bass and treble components from the low-frequency component of a music signal [37] to map to different frequencies [52]. Such techniques based on audio attributes have been employed to create meaningful and enjoyable haptic effects in many apps and games, using tools such as Apple Core Haptics[1] and Lofelt Studio.[2]

In this study, we emphasized achieving a high-level match between haptic effects and movie events to improve the experience of media content consumption. We attempted to achieve this through the use of four psychoacoustic measures: low-frequency energy, loudness, sharpness, and booming.

In terms of vibrotactile perception, humans are the most sensitive between 160 and 250 Hz [53]. Following the approach of Hwang *et al.* [52], we consider frequencies up to 215 Hz as borderline bass-band, as this corresponds approximately to the maximum frequency of the motor we use. We calculate the low-frequency energy value as the integral of the amplitudes of the frequency components up to 215 Hz.

Loudness is a well-known perceptual measure, indicating the perceived perception of sound pressure. It is calculated in the decibel scale as a weighted integration of spectral loudness, $N(f)$, for frequency f,

$$\begin{aligned} L &= \int_{0.3Bark}^{24Bark} N(f)df, \\ N(f) &= \frac{1}{W_k} 20log_{10}I(f). \end{aligned} \tag{3.1}$$

W_k is a weighting factor, which is extracted from the equal-loudness contour of the International Standards Organization on loudness (ISO 532) [54]. $I(f)$ is the intensity of the sound component at frequency f.

Sharpness and booming are also weighted summations of spectral energy in the frequency domain, emphasizing high- and low-frequency components, respectively. For example, a glass-breaking sound would exhibit high sharpness and low booming while the engine sound of a sports

[1] https://developer.apple.com/documentation/corehaptics

[2] https://lofelt.com/

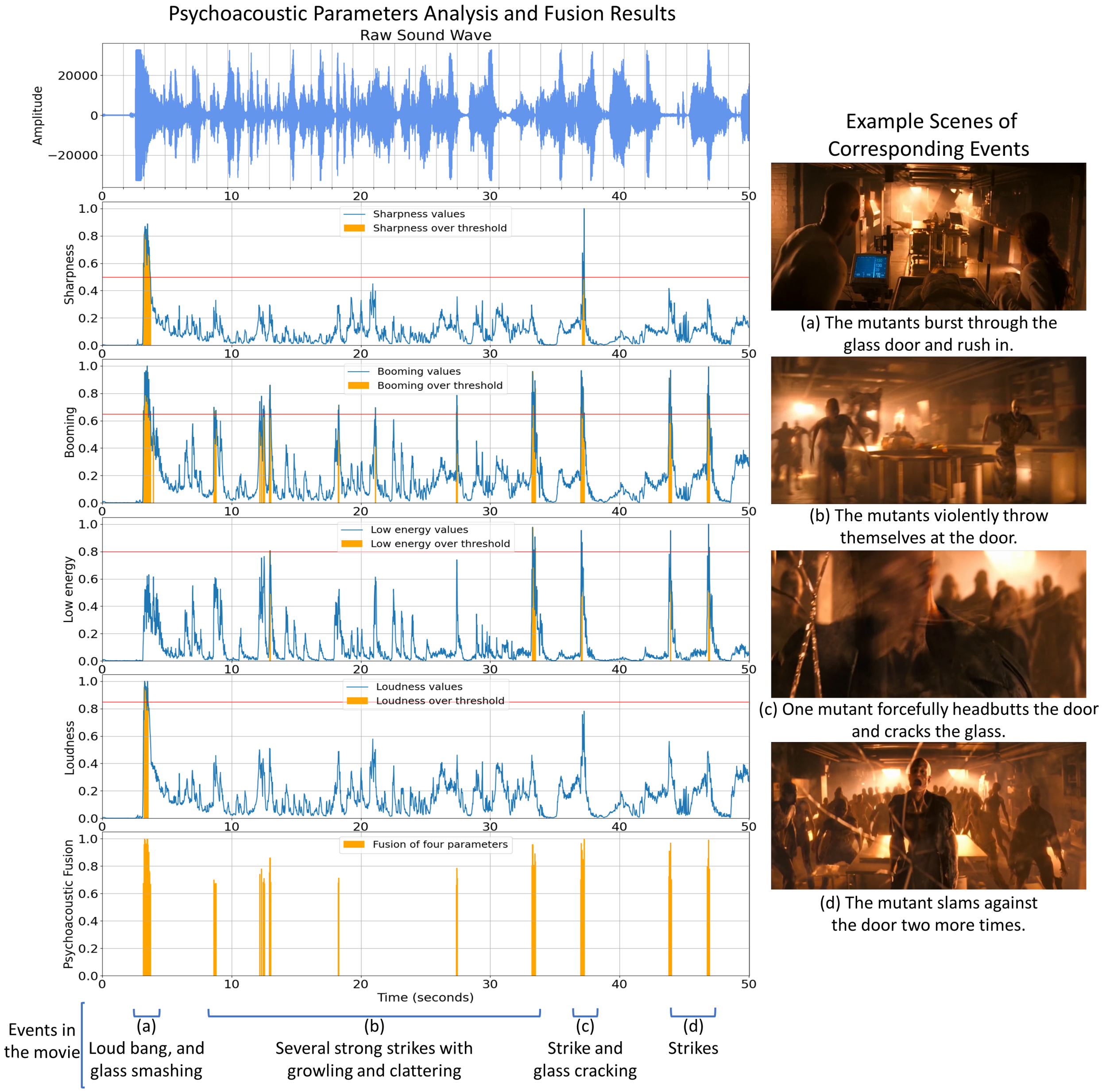


Fig. 3.1: An example of psychoacoustic analysis on a movie scene for sharpness, booming, low energy, and loudness with thresholds values of 0.5, 0.65, 0.8, 0.85, respectively. (Note that parameter values are normalized to 0 to 1.) The fusion of four parameters provides the values for vibrotactile rendering of the events presented in the movie scene.

car would be the opposite. These values can be derived by the weighted integral of spectral loudness $N(f)$ divided by loudness L. Sharpness S is calculated as follows:

$$S = 0.11 \frac{\int_{0.3Bark}^{24Bark} N(f)g(f)f\,df}{L} \tag{3.2}$$

where $g(f)$ is the weighting function. Booming B is calculated as

$$B = \frac{\int_{0.3Bark}^{24Bark} N(f)f\,df}{(0.49529Bark(f) + 0.14176)L} \tag{3.3}$$

The reader is referred to the reference of Zwicker and Fastl [48] for further details regarding these psychoacoustic variables and the associated calculations.

We used these metrics to analyze the attributes of sound effects in movie scenes, inferred the characteristics of the sound played, and then decided on the intensity of the haptic effects. Previous efforts on the haptic effects generation from sound often utilized low-frequency components since haptic sensation lies in the lower end of the frequency spectrum. Such approaches help to make the movie scenes more immersive, especially highly dynamic and violent ones including elements like explosions, gunfights, races, and chase sequences. However, these approaches often cannot filter out human voices, propagating incongruent vibration effects alongside the intended effects. Moreover, effects with high-frequency components, such as sci-fi (e.g., laser weapons or aircraft sounds) or magical effects, are often omitted since these scenes do not have enough low-frequency components. To overcome such problems, we sought another psychoacoustic parameter. From our observations, we picked sharpness and booming to represent effects with high and low frequency components, respectively. Loudness and low-frequency energy were selected to embody the absolute magnitude of sound intensity and attributes of the sound effects. In general, the portion and magnitude of low-frequency components increase when the source object associated with the sound effect is large and located near the camera. Examples include the sound spectrum of guns firing, cannons, bombs or other explosions.

For data processing, we used the 44.1 kHz recorded stereo sound of the movie clip. Calculation of the four psychoacoustic parameters was performed using the sound data stream sampled every 20 ms. We then normalized each parameter in the range of $[0, 1]$ by dividing by its maximum.

To determine the haptic output, we first select parameter-specific thresholds for use during the extraction of the target sounds from the movie scenes. For our study, we used post-normalization values of (sharpness, booming, low-energy, loudness) = (0.5, 0.65, 0.8, 0.85). If a calculated acoustic parameter exceeds any of these thresholds, vibrotactile effects will be presented. The threshold values were determined empirically as suitable to reduce undesirable haptic effects caused from sounds that are not related to the event context, such as narration or background music.

Users can adjust these thresholds according to the movie contents with our haptic-effects-generation pipeline. The intensity of vibration is determined by the highest parameter value among the four parameters. For example, if the audio of the scene results in a parameter vector of (0.6, 0.4, 0.5, 0.9), the vibration intensity would be 0.9 times the maximum. The intensity of vibration was spatially mapped using the vibrotactile actuator array in the fusion phase, described in Section 3.2.

3.1.2 Visual Modality Analysis

In parallel with sound processing, we also analyzed the visual content of the movie source. As demonstrated by Kim *et al.* [42], the salient region is an effective indicator of the location of the events that occur on screen and can be associated with the spatial distribution of the actuators. Accordingly, we first predicted the saliency area of the frames using a convolutional neural network (CNN) model, choosing this framework because of its superior performance on 2D image tasks such as object detection [55, 56], recognition [57, 58, 59, 60], and segmentation [61, 62, 63] compared to other machine learning algorithms.

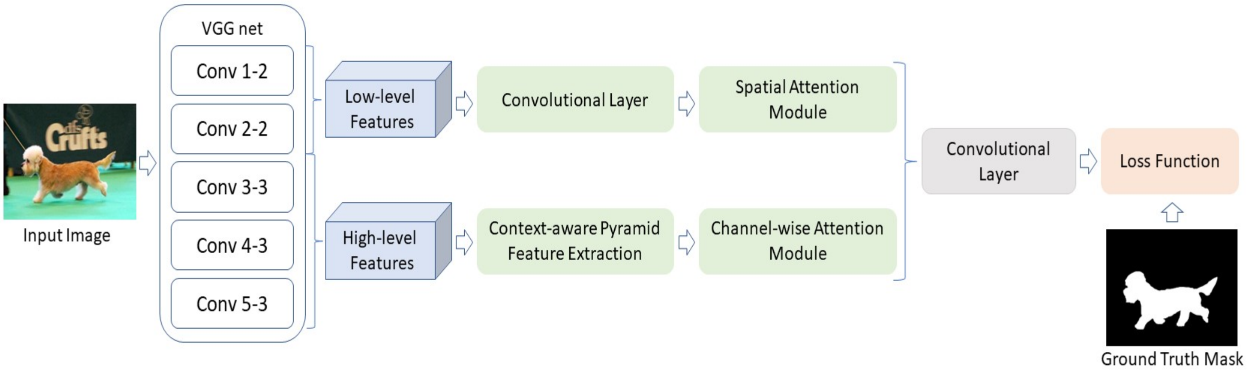


Fig. 3.2: Structure of the saliency detection model.

To implement the model for saliency detection, we followed the pyramid feature selective network presented in [64].[3] The overall architecture of the model is illustrated in Figure 3.2. The input image is first fed into convolutional layers based on VGG net [57] as CNN has shown a strong ability to learn features of the images. Several modules with attention mechanisms are then deployed as attention is a perfect fit to select features and detect areas of interest. More specifically, the branch with Conv 3-3, Conv 4-3, Conv 5-3 blocks based on VGGnet [57], context-aware pyramid feature extraction (CPFE) module, and channel-wise attention (CA) module captures high-level features with the rich context of the images. The CPFE module adopts atrous convolution with dilation rates of 3, 5, and 7 to obtain context information from the multi-receptive field. The CA module assigns different weights to channels according to their response to salient objects, while another branch with Conv 1-2, Conv 2-2 taken from VGGnet [57] extracts low-level features of the

[3] https://github.com/CaitinZhao/cvpr2019_Pyramid-Feature-Attention\ -Network-for-Saliency-detection

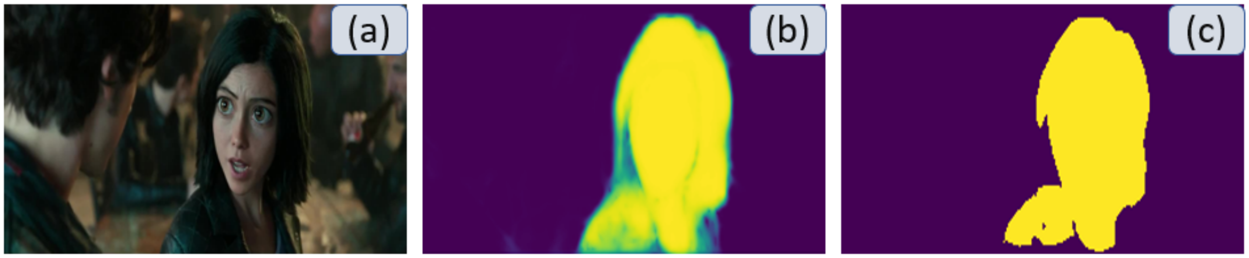


Fig. 3.3: Saliency map generated by CNN model: (a) raw input image, (b) predicted saliency mask from the model, and (c) result after applying a filter on the saliency mask to remove noisy salient edges.

images. A spatial attention module follows and aims to refine detailed information of the salient area and filter out the noises in the background that may distract saliency mask generation.

We pre-trained the model on the DUTS-train dataset [65] with 10,553 training images. We then developed an interface for generating saliency maps for our dataset composed of movie frames. As shown in Figure 3.3, we ultimately obtained the salient area (shown in yellow) and background area (shown in dark purple) for each frame.

3.2 Information Fusion for Vibrotactile Effect Generation

After analyzing the audio and visual modalities, we obtained the targeted frames with psychoacoustic measurements and the area of interest for each frame. In the following, we consider how to fuse psychoacoustic and visual features in our haptic rendering system. For our haptic rendering platform, the system uses a chair with nine vibrotactile actuators, integrated into the back, seat, and arms as illustrated in Figure 3.4. The saliency map must be down-sampled, typically to a small grid, for example, of 3 × 3 [42] or 2 × 3 actuators [66], making contact with the user's back and hips for rendering moving tactile sensations. Amemiya *et al.* simulated vibration in the seat pan, presented in conjunction with a visual display of optical flow to intensify the perception of forward velocity [67]. Inspired by phantom tactile sensations [68], we developed an algorithm that provides vibrotactile patterns with reasonably high spatiotemporal resolution. To provide an opportunity for improved immersion, we also placed actuators on the armrests in order to extend the rendering of effects beyond the user's back and seat.

We determined the timing and intensity of vibration rendered by each actuator as follows. First,

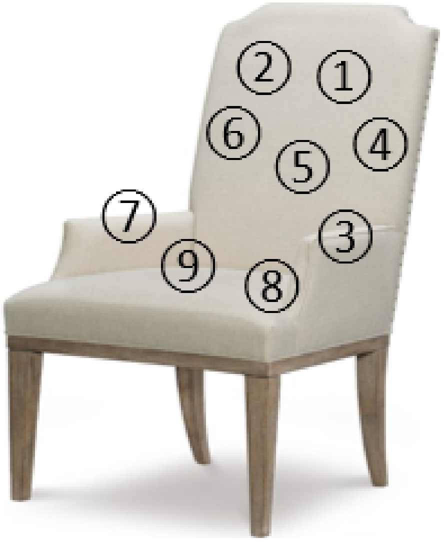


Fig. 3.4: Image division by actuators.

we mapped the locations of the nine actuators on the chair to on-screen vertices $A_1, A_2, ..., A_9$ as shown in Figure 3.5. Considering the shape of the chair and the aim of covering the user's back, arms, and bottom, we distributed A_3, A_7 on the armrest and A_8, A_9 on the seat. We arranged A_{1-2} and A_{4-6} in a manner that provided coverage over the user's back and obtained a configuration of actuators of 14 split triangles of similar size, as needed for later steps of our rendering algorithm. This arrangement resulted from several iterations of self-tests and pilot tests conducted by the authors. We arranged the actuators so that $\overline{A_1A_8}$, $\overline{A_2A_9}$ divide the frame into three equal horizontal ranges; $\overline{A_3A_4}$, $\overline{A_4A_5}$, $\overline{A_5A_6}$, $\overline{A_6A_7}$ divide the frame into four equal horizontal ranges, and $\overline{A_3A_7}$ bisects it vertically. We overlaid this actuator position division mask on each frame, which is generated by the CNN model with the predicted saliency region. With connections drawn between the nearest neighbors, this actuator distribution divides the frame into fourteen triangles, as shown in Figure 3.6(b). In each triangle, we calculated the centroid C_i of the saliency shape, $i \in \{1, ..., 14\}$ and the distance from centroid C_i to its three adjacent vertices. In summary, figure 3.6 provides an overview and an example of C_{12} in $\triangle A_5A_6A_9$.

Second, we assigned weights that are inversely proportional to the distance between the centroid and the three adjacent vertices. More specifically, we defined $\alpha_{\triangle i}^{A_n}$ as the reciprocal of the

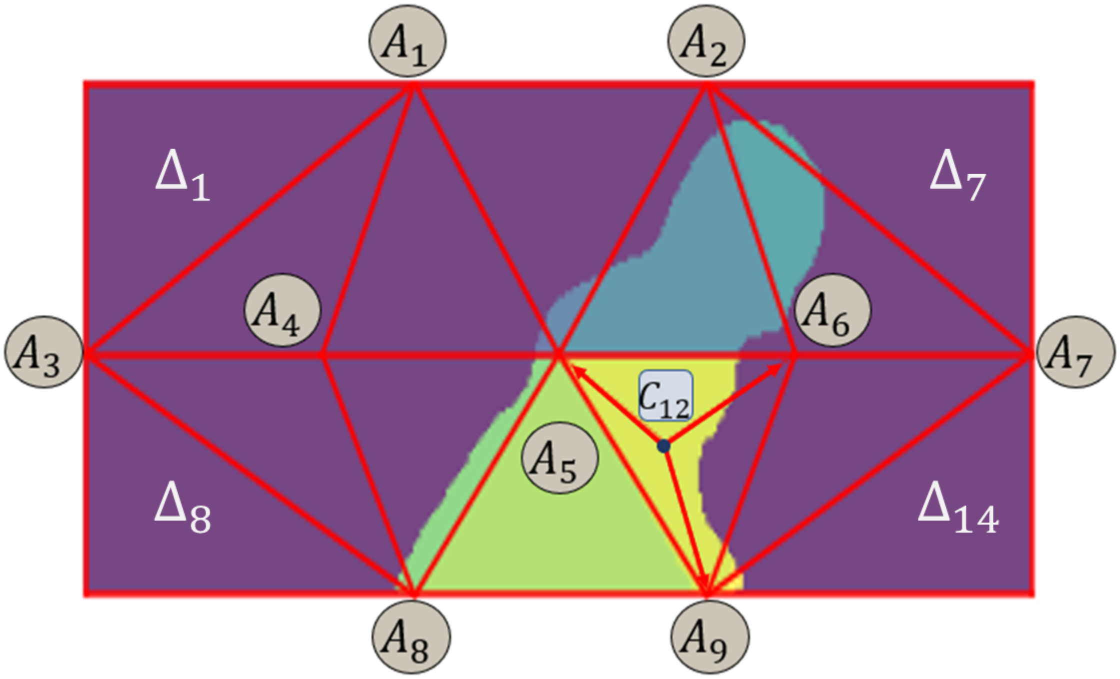


Fig. 3.5: Nine-actuators layout dividing the scene and distributing the vibrotactile intensity for rendering. The example of △12 with centroid C_{12} distributing weight to actuator A_5, A_6, and A_9 is shown.

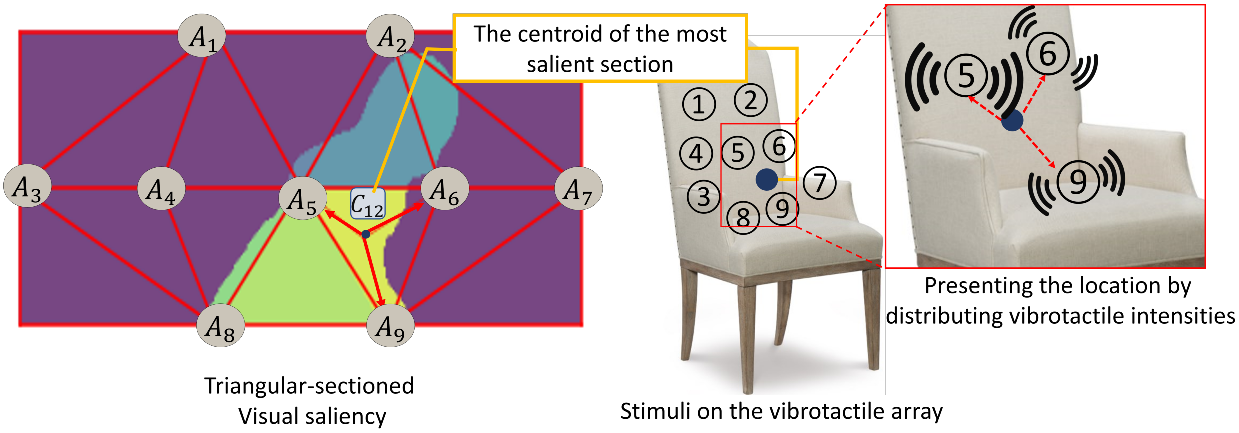

The centroid of the most salient section
A_1
A_2
A_3
A_4
A_5
C_{12}
A_6
A_7
A_8
A_9
Triangular-sectioned Visual saliency
1
2
4
5
6
3
7
8
9
Stimuli on the vibrotactile array
5
6
9
Presenting the location by distributing vibrotactile intensities

Fig. 3.6: Nine-actuators layout dividing the scene and distributing the vibrotactile intensity for rendering. The example of △12 with centroid C_{12} distributing weight to actuator A_5, A_6, and A_9 is shown.

distance between C_i and the adjacent vertex A_n, $C_iA_n^{-1}$. An example of $\triangle 12$ with actuator A_5, A_6, and A_9 is shown in Figure 3.7. Next, we assigned a weight of $w_{\triangle i}^{A_{n_k}}$ to A_{n_k} by

$$w_{\triangle i}^{A_{n_k}} = \frac{\alpha_{\triangle i}^{A_{n_k}}}{\alpha_{\triangle i}^{A_{n_1}} + \alpha_{\triangle i}^{A_{n_2}} + \alpha_{\triangle i}^{A_{n_3}}}, \quad k \in \{1,2,3\} \tag{3.4}$$

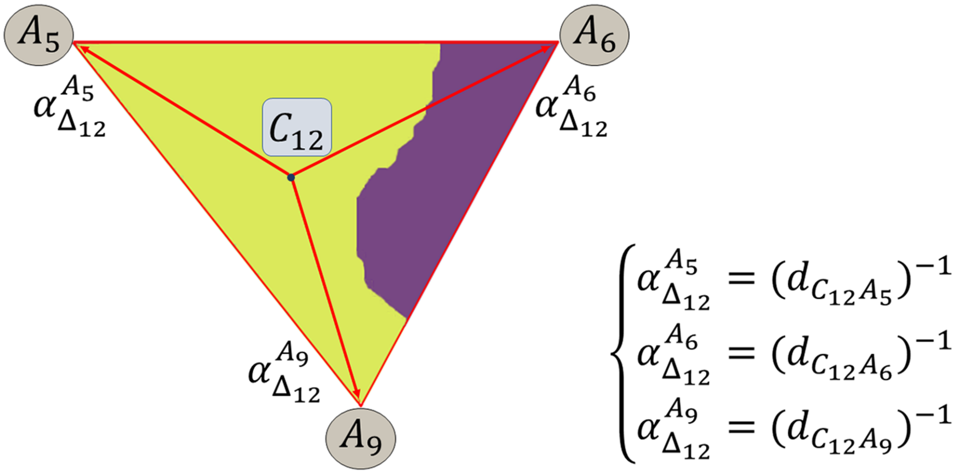
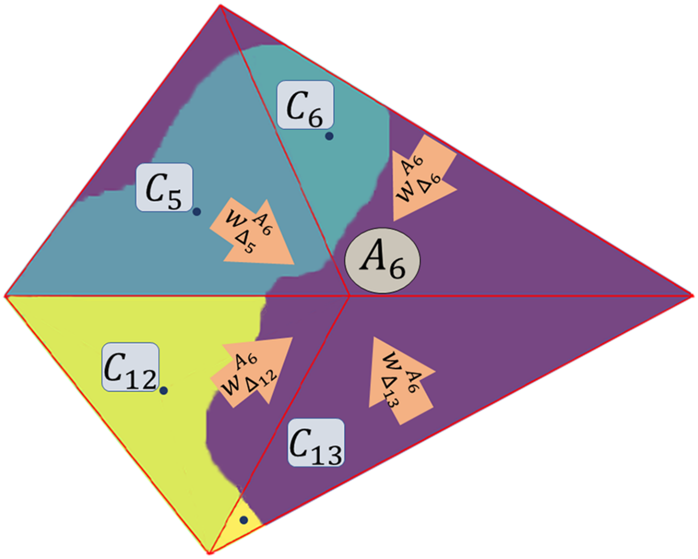

A_5
A_6
C_{12}
$\alpha_{\Delta_{12}}^{A_5}$
$\alpha_{\Delta_{12}}^{A_6}$
$\alpha_{\Delta_{12}}^{A_9}$
A_9
$\begin{cases} \alpha_{\Delta_{12}}^{A_5} = (d_{C_{12}A_5})^{-1} \\ \alpha_{\Delta_{12}}^{A_6} = (d_{C_{12}A_6})^{-1} \\ \alpha_{\Delta_{12}}^{A_9} = (d_{C_{12}A_9})^{-1} \end{cases}$

Fig. 3.7: Example of $\triangle 12$ with actuator A_5, A_6, and A_9 as vertexes and $\alpha_{\Delta_{12}}^{A_5}$, $\alpha_{\Delta_{12}}^{A_6}$, $\alpha_{\Delta_{12}}^{A_9}$ distributing to three vertexes respectively.

where $A_{n_k}, k \in \{1,2,3\}$ are three adjacent vertexes of $\triangle i$. Then, for each actuator A, we summed the weights $w_{\triangle i}^{A_{n_k}}$ adding from its adjacent triangles to obtain W_{A_i}. As shown in Figure 3.8, we have four weights contributing to W_{A_6} of actuator A_6: $w_{\Delta_5}^{A_6}$, $w_{\Delta_6}^{A_6}$, $w_{\Delta_{12}}^{A_6}$, $w_{\Delta_{13}}^{A_6}$, which are obtained from Equation 3.4. We then normalized all the weights of each actuator A and generated rendering weights β_n, $n \in \{1,2,...,9\}$.

Finally, we obtained the overall vibration intensity for the current frame by linearly mapping the psychoacoustic measurement values as described in Section 3.1.1 to the actuators' amplitudes, multiplied by the rendering weight β_n of each actuator. As a result, the location of vibration follows the salient region; when the user's visual attention rests on the left side of the movie scene, the corresponding vibrotactile effects with intensity derived from audio analysis, generated using the above mapping,will also be presented on the left side of the body, as seen in Figure 3.6.



$$W_{A_6} = w^{A_6}_{\Delta_5} + w^{A_6}_{\Delta_6} + w^{A_6}_{\Delta_{12}} + w^{A_6}_{\Delta_{13}}$$

$$\begin{cases} w^{A_6}_{\Delta_5} = \dfrac{\alpha^{A_6}_{\Delta_5}}{\alpha^{A_2}_{\Delta_5} + \alpha^{A_5}_{\Delta_5} + \alpha^{A_6}_{\Delta_5}} \\ w^{A_6}_{\Delta_6} = \dfrac{\alpha^{A_6}_{\Delta_{12}}}{\alpha^{A_2}_{\Delta_6} + \alpha^{A_6}_{\Delta_6} + \alpha^{A_7}_{\Delta_6}} \\ w^{A_6}_{\Delta_{12}} = \dfrac{\alpha^{A_6}_{\Delta_{12}}}{\alpha^{A_5}_{\Delta_{12}} + \alpha^{A_6}_{\Delta_{12}} + \alpha^{A_9}_{\Delta_{12}}} \\ w^{A_6}_{\Delta_{13}} = \dfrac{\alpha^{A_6}_{\Delta_{13}}}{\alpha^{A_6}_{\Delta_{13}} + \alpha^{A_7}_{\Delta_{13}} + \alpha^{A_9}_{\Delta_{13}}} \end{cases}$$

Fig. 3.8: Example of four weights $w^{A_6}_{\Delta_5}$, $w^{A_6}_{\Delta_6}$, $w^{A_6}_{\Delta_{12}}$, and $w^{A_6}_{\Delta_{13}}$ contributing to W_{A_6} of actuator A_6

Chapter 4

Hardware Implementation for Vibrotactile Effect Generation and User Study

Yaxuan Li, Yongjae Yoo, Antoine Weill-Duflos, and Jeremy Cooperstock. 2021. Towards Context-aware Automatic Haptic Effect Generation for Home Theatre Environments. In Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology (VRST '21). Association for Computing Machinery, New York, NY, USA, Article 13, 1–11. DOI: https://doi.org/10.1145/3489849.3489887

Preface

This chapter presents work that was published in the 27th ACM Symposium on Virtual Reality Software and Technology. As described in the previous chapter, we first obtained psychoacoustic measures to determine the targeted frames for haptic effects generation and each frame's region of visual interest after extracting features from audio and visual modalities. Then we combined the two streams of information and converted them into vibrotactile sensations that can be rendered on the hardware. In this chapter, we deployed the algorithm on hardware as described in Section 4.1. In order to verify the performance of our algorithm, Section 4.2 introduces the user study conducted on 16 participants. The results of this study are presented in Section 4.3. Concluding the chapter, Section 4.4 discusses the limitations of the algorithm in its current implementation, and how the algorithm might be extended in the future.

All experiments were approved by McGill University's Research Ethics Board Office[1], REB file #21-02-023. To take part in the experiment, participants were required to acknowledge and accept a consent form.

Author's Contribution

This work is a result of collaboration. Yaxuan Li designed the user study, applied for REB, conducted the user study, and analyzed the results. Dr. Yongjae Yoo prvided valuable feedback and guidance throughout the research and edited the paper. Dr. Antoine Weill-Duflos gave suggestions for designing the study. Prof. Cooperstock supervised the research, gave critical feedback, and edited the paper.

[1] https://www.mcgill.ca/research/research/compliance/human

4.1 Hardware Implementation

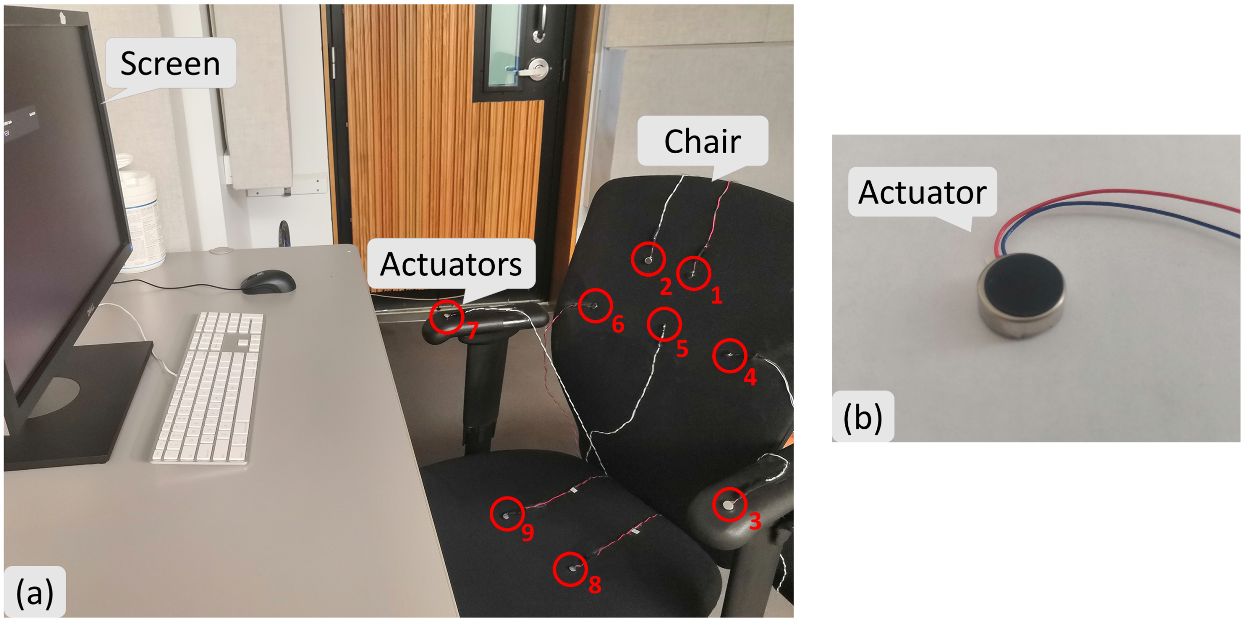


Fig. 4.1: Hardware implementation and experiment setup.

We built the hardware system with nine eccentric rotating mass (ERM) vibrotactile actuators (Seeed Studio; RB-See-403, $\phi = 10\,\text{mm}$) controlled by a Teensy 3.2 microcontroller, connected to the media player computer. The actuators are embedded in the chair as illustrated in Figure 4.1. Vibration amplitude ranges of 1.77–5.31 g on the armrests of the chair, 2.18–5.63 g on the seat, and 2.41–7.99 g on the backrest were measured by a triaxial accelerometer (PCB electronics; Model 356A01) and a data acquisition card (National Instruments; Model USB-4431). The frequency of vibration ranged from 110 to 230 Hz, and the rise time of the ERM motors was approximately 50 ms. Although this response time is non-negligible, it is also not excessive for our purposes. First, as reported by Lee and Choi [41], in terms of the delay of driving actuation, real-time haptic applications on mobile devices incur approximately 20-30 ms delay, but participants perceive such haptic effects as "simultaneous". Second, the movie industry standardized on 24 frames per second, whereas television productions typically use 25 (PAL) or 30 (NTSC) frames per second. Although higher frame-rate recording is commonly used in digital productions, television broadcasts are nevertheless mostly confined to the lower rates, for which each frame therefore represents

33-42 ms. Accordingly, the 50 ms rise time of the actuator constitutes only a minor difference, which is difficult to perceive as asynchronous. As a movie starts playing, the microcontroller receives vibration amplitudes synchronously from the computer and triggers the ERMs to present vibration effects every 100 ms.

4.2 User Study

This section describes the user study for evaluating the effectiveness of our haptic effects generation algorithm, employing the hardware implementation described in Section 3.2. In the user study, we compared participants' subjective ratings regarding haptic effects generated from different conditions with several movie clips.

4.2.1 Participants

We recruited 16 participants for the experiment (20-34 years old, 9 male, 7 female). No participants reported any known sensory disorders that could affect their auditory, vision, or haptic perception. They were also asked to wear thin T-shirts to allow for better perception of vibrotactile stimuli on their torso. Each participant was compensated approximately 8 USD for their participation.

4.2.2 Stimuli

We tested four different haptic rendering conditions, described in Table 4.1. Six short (approximately one-minute) movie clips were selected, including scenes with fighting, shooting, chasing, or exploding. They were trimmed from four different movies: *Alita: Battle Angel* (2019), *The Boss Baby* (2017), *I Am Legend* (2007), and *Escape Room* (2019). Table 4.2 summarizes the movie clips, their length, and the number of haptic effects rendered in each scene. In total, 24 combinations from six movie clips and four rendering conditions were provided to each participant. Each of the six movie clips was considered as a block, and within the block, the presentation order of haptic effects was randomized.

Table 4.1: Four conditions of haptic effects generations.

Condition	Description
Random	Vibration effects were generated **randomly** and presented through the video clips.
Audio	Vibration effects were generated based on **the audio parameters** described in Section 3.1.1 only, and presented actuator A_1, A_2, and A_5.
Visual	Vibration effects were presented in the **salient area** as mentioned in Section 3.1.2, but not considered audio variables. All vibrotactile actuators corresponding to the salient area in the screen vibrated.
Multimodal	The **full pipeline** described in Section 3.1 and 3.2, were used for generating vibration effects.

Table 4.2: Six movie clips for user study with their time length and components in the scene

Number	Movie clip	Time length (second)	Components in the scene
1	*Alita: Battle Angel* (2019) [Clip A]	58	Grappling, Melee fighting
2	*Alita: Battle Angel* (2019) [Clip B]	62	Brawling, Crashing, Tasering
3	*The Boss Baby* (2017)	60	Chasing, Fighting, Laughing, Giggling, Crunching
4	*I Am Legend* (2007) [Clip A]	62	Gunshots, Slamming, Crashing, Screaming
5	*I Am Legend* (2007) [Clip B]	73	Gunshots, Barking, Snarling, Growling
6	*Escape Room* (2019)	74	Whooshing, Exploding, Crawling

4.2.3 Methods and Procedure

The user study was conducted in a recording studio at the author's institution, depicted in Figure 4.1. Participants sat approximately 45-55 cm in front of a 21 inch monitor, and wore headphones (Sony, WH-H900N) to hear the movie clip sounds, and prevent them from hearing the faint noises produced by the vibration motors. The participants were first asked to read through the consent form and listen to the explanation of the experiment. All the participants agreed and signed the consent form, and then they were asked to sit on the vibrotactile chair. Next, the experimenter confirmed that the participant could easily detect vibrations from each of the actuators. The experimenter also confirmed that the intensity of the vibration was not excessive, as this might cause discomfort throughout the experiment.

A short training session was given to the participants. They experienced a 20-second-long movie clip cropped from *Alita: Battle Angel*. The clip was different from the movie clips used in the main session introduced in Table 4.2. Four different haptic conditions, as described in Table 4.1, were presented along with the movie.

In every trial in the main session, the participants watched the movie while perceiving the haptic effects. They were then given a questionnaire (Table 4.3) to evaluate their experience with the haptic rendering in terms of *Immersion, Preference, Harmony,* and *Discomfort*, using Likert scales ranging from 0 (strongly disagree) to 10 (strongly agree). After completing the questionnaire, the participants took a short break with at least one minute to avoid fatigue and adaptation.

Table 4.3: The given statements for the four items on the questionnaire for assessing the haptic-audio-visual experience.

Immersion	The haptic effects make me more immersed in the movie.
Preference	I liked the experience of vibration sensations while watching this movie clip.
Harmony	The haptic effects are consistent with the content of the scene.
Discomfort	The haptic effects are uncomfortable for me.

The entire procedure took approximately one hour per participant, and was carried out under approval of the McGill University Research Ethics Board, REB # 21-02-023, in compliance with regional COVID-19 restrictions. This included paying attention to sanitary measures, wiping down of the apparatus after each use, and participants and the experimenter wearing face masks, while maintaining a 2 m distance.

4.2.4 Data Analysis

We initially averaged the 16 participants' ratings per each of the 24 combinations for a simple comparison by plots (Figure 4.2). For the statistical analysis of the results, we conducted a two-way ANOVA, one-way ANOVA and Tukey's HSD post-hoc tests for each of the four subjective ratings to see the effect's significance.

4.3 Results

4.3.1 ANOVA analysis

We performed a two-way analysis of variance (ANOVA) test for immersion, preference, harmony, and discomfort to understand the effects of haptic rendering methods and movie content. Table 4.4 summarizes the statistical analyses. For all the variables, the haptic rendering method largely affected the user's evaluation (effect size η^2 near 0.5). The effects of the movie clip's content were also significant in immersion, preference, and harmony scores, despite having small effect sizes. Interestingly, for discomfort rating, the effect size of the haptic rendering factor drastically decreased, and the movie's effects turned out to be insignificant. The interaction terms were not significant, meaning the effects of different haptic rendering methods were not affected by different movie contents.

Table 4.4: Two-way ANOVA results for the four subjective rating.

Measure	Factor	Statistics	Effect size (η^2)
Immersion	Rendering	$F(3,45) = 138.52$, $p < .001$	0.512
	Movie	$F(5,75) = 3.47$, $p = .004$	0.021
	Interaction	$F(15,225) = 1.25$, $p = .232$	0.023
Preference	Rendering	$F(3,45) = 120.51$, $p < .001$	0.480
	Movie	$F(5,75) = 2.74$, $p = .018$	0.018
	Interaction	$F(15,225) = 1.21$, $p = .258$	0.024
Harmony	Rendering	$F(3,45) = 158.97$, $p < .001$	0.556
	Movie	$F(5,75) = 2.32$, $p = .043$	0.013
	Interaction	$F(15,225) = 0.61$, $p = .864$	0.011
Discomfort	Rendering	$F(3,45) = 16.25$, $p < .001$	0.113
	Movie	$F(5,75) = 2.20$, $p = .054$	0.025
	Interaction	$F(15,225) = 0.75$, $p = .735$	0.026

4.3.2 In-depth Analysis

To understand the results better, we conducted a more in-depth analysis on the rendering methods in each of the movie clips. Figure 4.2 shows the average scores of the four subjective scores. Note that the scale was inverted in discomfort ratings, i.e., a lower score indicates better performance (less discomfort). We performed a one-way analysis of variance (ANOVA) test to assess the performance of the four haptic conditions for each movie clip. In terms of the subjective evaluation metrics immersion, preference, and harmony, all four haptic rendering conditions had a statistically significant effect ($p < 0.01$ or $p < 0.05$) on user experience while viewing all six movie clips. However, in terms of discomfort, the haptic rendering conditions only had a statistically significant effect on user experience while viewing *Alita: Battle Angel* [Clip A], *The Boss Baby*, and *Escape Room*.

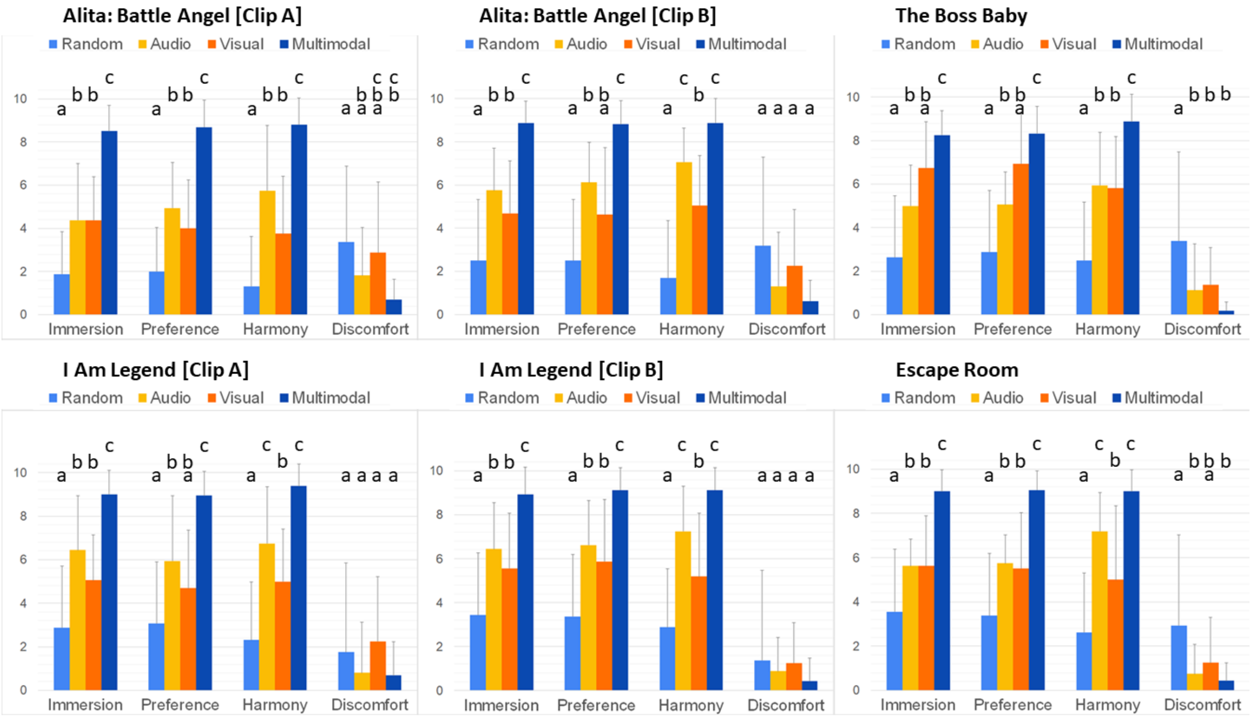


Fig. 4.2: Results of experiment with six movie clips to evaluate the users' haptic-audio-visual experience by a questionnaire in terms of immersion, preference, harmony, and discomfort. The error bars represent standard errors. The conditions with the same letters above the bar indicate there are no significant differences among them.

The detailed statistical results are illustrated in Table 4.5. For each of the one-way ANOVAs, we conducted a posthoc analysis of Tukey's HSD test to understand which components are statistically meaningful. Posthoc grouping labels are presented above the bar; different alphabet coding indicates a significant difference between items.

Table 4.5: Results of one-way ANOVA for each movie clip in terms of immersion, preference, harmony, and discomfort. The first row of each clip represents $F(3, 45)$ value and the second row with $p < 0.01$ or $p < 0.05$ indicates there is a significant effect.

Measure / Movie	Immersion	Preference	Harmony	Discomfort
Alita: Battle Angel [Clip A]	29.495 $p < .001$	33.021 $p < .001$	27.962 $p < .001$	3.141 $p = .032$
Alita: Battle Angel [Clip B]	23.932 $p < .001$	20.126 $p < .001$	37.443 $p < .001$	2.586 $p = .061$
The Boss Baby	26.263 $p < .001$	21.293 $p < .001$	21.36 $p < .001$	6.400 $p = .001$
I Am Legend[Clip A]	25.405 $p < .001$	17.010 $p < .001$	30.233 $p < .001$	1.762 $p = .164$
I Am Legend[Clip B]	18.370 $p < .001$	18.115 $p < .001$	24.377 $p < .001$	1.152 $p = .336$
Escape Room	21.986 $p < .001$	20.562 $p < .001$	22.922 $p < .001$	5.072 $p = .003$

Here, the summaries of observations are:

- The effect of the haptic rendering method has a clear significant effect for all the subjective ratings, showing that the multimodal haptic rendering condition performed best.
- A general trend of Random < Visual < Audio < Multimodal can be observed throughout the plots. Though there is only a slight difference between audio and visual rendering scores, we can infer that time-synchronicity would be more important than presenting the location of the event.
- For discomfort rating, the content of the movie seems to affect the user's evaluation. For example, discomfort ratings show a significant difference in *Alita* [Clip A], *Boss Baby*, and *Escape Room* clips, but the others show no differences.

4.4 Discussion

4.4.1 General Discussion

We can observe that the multimodal haptic effect condition delivers the best user experience with regards to immersion, preference, and harmony, with the lowest level of discomfort. The results solidly present the effectiveness of our algorithm. Furthermore, using both visual and audio features in the haptic effects generation increased the subjective rating scores.

In all the experimental conditions, multimodal rendering received ratings over 8 out of 10. We can conclude the rendering algorithm effectively communicated the context of the scene. The two features, location and amplitude, are well mapped and easily distinguishable by the participants and helped them in their understanding of the scene. Moreover, this was achieved without the use of high-performance actuators such as voice coils.

A point for further discussion is user discomfort, which includes the effects of fatigue caused by an immoderate number of vibrations, or excessive cognitive load due to the rapidly varying stimuli. As illustrated in Figure 4.2, we observed that multimodal rendering achieves the lowest level of discomfort. Furthermore, the large standard deviation indicates a relatively substantial individual variance on the "comfortable range" on vibration perception.

We found that the score of psychoacoustic parameters-based rendering is slightly better than saliency-based rendering; however, the difference was not statistically significant. This implies that time synchronicity is more crucial than location matching for vibrotactile rendering. However, we picked mostly dynamic scenes for this study, which are accompanied by clear, distinctive sound effects. Scenes that are not dynamic and violent (e.g., slow camera motion, speaking, romance scenes, etc.) need further investigation.

4.4.2 Extensibility of the Algorithm

The algorithm's flexibility lends itself to scaling and expansion to other applications such as Virtual Reality (VR), Augmented Reality (AR), and games. The pipeline also allows for retroactively introducing 4D effects to previous movies made without 4D in mind. We provide a UI for our program that allows users to adjust thresholds for each of the psychoacoustic parameters of sharpness, booming, low-energy, and loudness. Once a video is selected to be processed, users can choose and tune psychoacoustic measurements and their thresholds. The program will generate a text file for driving actuators and a video for haptic effects visualization. Haptics practitioners can use our tool to create vibrotactile effects for their videos regardless of VR or gaming content.

Moreover, since the feature extraction part is detached from the calculation of the rendering parameters, we can separately add or remove audiovisual parameters in that stage. For example, colour map extraction [69] and camera motion estimation [70, 71] could be easily integrated with the current system. Higher-level contextual information can be obtained by using machine learning techniques, such as violent scene detection [72, 73, 74, 75] or semantic segmentation [76, 77] to make haptics more comprehensively fit the storyline.

In terms of haptic actuation mechanisms, different modalities such as thermal or airflow sensations could be easily attached. For example, in a related experimental multi-haptic armrest design [78], there are two different types of vibrations, thermal sensation, airflow, and poking mechanisms. These mechanisms would be mapped to multiple different audiovisual features extracted from the scene and would be expected to improve the watching experience. The challenge for such conversion is to determine the appropriate mapping between the movie content or detected events and the associated type of mechanism or actuator to drive in response. Of course, there remains the challenge of parameter selection and tuning to ensure the appropriate mapping between the content or detected event and the associated type of actuator to drive in response. We are currently investigating these issues. Regarding application scenarios, we initially considered a home theatre watching experience, where the algorithm is easily attached to the streaming services to provide 4D haptic effects. The commercialization of this and similar work could have an impact on the future of at-home content consumption.

4.4.3 Limitations

Despite offering automatic haptic authoring and promising application possibilities, our pipeline suffers from a number of limitations. In Section 3.1.2, we presented our visual saliency detection algorithm for distributing vibrotactile effects to produce an immersive experience. However, we observed a small number of scenes with strong sound, but no visual objects on screen, such as ambush attacks from a ghost or monster in the dark, an electricity blackout, or heavy wind blowing out the fire in the scene. In such cases where sound components exist in the absence of corresponding visual objects, we fixed the saliency area in the middle of the frame and assigned A_5 to take responsibility for rendering vibrotactile actuation. Furthermore, movies contain many blurred frames to guarantee the fluidity of characters' actions, which poses a challenge for the saliency detection model and results in imprecise edge detection of objects.

These limitations may be overcome by more powerful, accurate, or movie-specific detection or segmentation machine learning models, incorporating, for example, methods of stereo sound localization [79], attention [80], and depth extraction [81], and equally, taking advantage of large-scale movie datasets such as IMDB and Movielens. We emphasize that the proposed technique does not aim to serve as a replacement for manual design by haptics experts. The quality of the haptic patterns from our pipeline is inferior to those created by laborious manual authoring. However, we believe that our pipeline provides useful insights, and offers a preliminary version of high-quality effects that can be implemented in an early stage in the content authoring.

4.5 Conclusion

In this study, we proposed an automatic multimodal haptic rendering algorithm for movie content, which extracts audiovisual features to infer basic contextual information of the scene and renders haptic effects. We especially targeted ways to improve the home theatre experience. Using a haptic chair equipped with an array of vibration motors, we conducted a user study to evaluate the effectiveness of the algorithm in terms of user experience by measuring subjective ratings of immersion, preference, harmony, and discomfort. The results demonstrate that the proposed multimodal rendering noticeably improved the viewing experience. Our future work will expand the algorithm and include additional types of haptic modalities.

Chapter 5

Automatic Thermal and Airflow Effect Generation for Immersive Viewing Experience

Preface

In the preceding chapters, we introduced an algorithm for authoring vibrotactile stimuli based on multimedia input. In this chapter, following the idea of enhancing experience for the 4D movie in the home theatre, we describe an immersive system that generates thermal and airflow-based stimuli. Section 5.2 presents our approach to generating thermal effects. Section 5.3 covers our approach to airflow generation. The emphasis in this chapter is on algorithm feasibility rather than hardware implementation; therefore, rather than carrying out a user study to evaluate the hardware, we present algorithm output as a series of visualizations.

5.1 Introduction

Tactile perception, which includes sensation of temperature, is mediated by a variety of receptors in humans, including mechanoreceptors and thermoreceptors. Haptic sensation is seen as a result of these receptors cooperating. For instance, when examining an object's tactile properties, one senses both texture and temperature simultaneously. With the vibrotactile effects generation algorithm, the majority of existing haptic technology is geared towards simulating mechanical interactions between humans and objects.

However, it is well established that temperature sensitivity enhances haptic experience [82, 83, 84]. People are capable of detecting the temperature of the exhibited virtual items or the surrounding environment. Given this, we anticipate that thermal stimuli, if delivered in a sufficient manner, would create a greater sense of immersion in the haptic experience (such as the one presented in preceding chapters). It is well documented that one's impression of a material, in terms of haptic perception, is highly dependant on its inherent heat conductivity [85]. Thus, temperature sensitivity provides an additional information channel that thermoreceptors are capable of processing. In this chapter, we describe image analysis techniques that convert colour into temperature. This information can be used to render thermal stimuli alongside a video, to further enhance the immersive experience.

In addition, the use of airflow has been widely explored for improving immersion in various applications, including in either physical environment or virtual reality [86, 87, 88, 89]. In this chapter we analyze video-frame motion, using optical flow and depth estimation techniques, to generate airflow effects that improve viewer immersion.

5.2 Colour Feature Extraction and Thermal Effect Generation

Although colour perception is commonly regarded as a feature related to vision, studies have documented that it can also influence other sensory modalities, such as thermal, olfactory, and gustatory perceptions as well [90, 91, 92, 93, 94]. For example, surfaces that are predominantly red are often described as warm, while those that are blue tend to be described as cold [95]. There are commonly three methods for calculating the crossmodal correlation between colour and temperature: (1) subjective measurements such as asking participant to rate the warmth or coolness of coloured patches and stimuli [96, 97]; (2) objective measures of reaction time, e.g., Implicit Association

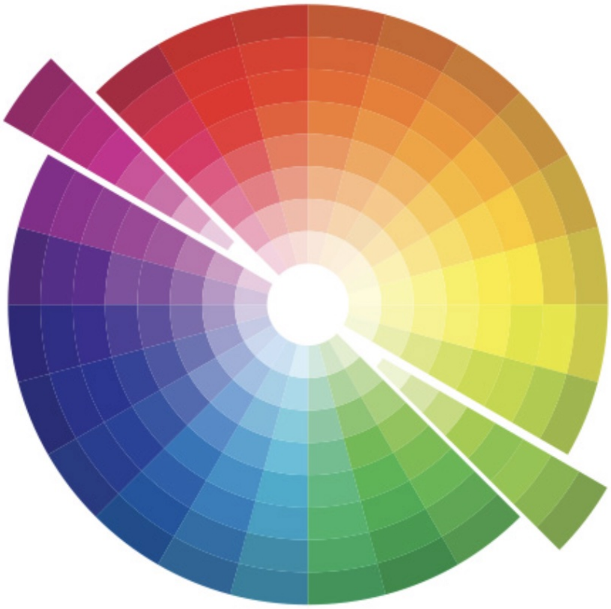


Fig. 5.1: Warm and cold colour wheel.

Test [98]; (3) a task to investigate whether priming by colour, or by thermal information, could invoke a cross-modal association Outside of perceptual research, the connection between colour and temperature has had a widespread impact in the disciplines of architectural and industrial design [99]. In the following, we introduce an approach to colour extraction and analysis, and then describe how the analysis is mapped to a thermal effect.

To begin, the algorithm calculates the average value for each of the three RGB channels, across all pixels in a frame. An example of this procedure can be seen in Figure 5.2. Then we calculate the warmth or coldness of each frame. Based on the colour theory [1], colours can be observed as cold or warm. The colour wheel has a set of hues associated with the sun, warmth, and fire. These hues are referred to be warm colours because they elicit warm emotions in individuals. Yellow, red, orange and various variations of these colours are considered warm. These colours have long wavelengths, making them immediately noticeable and vibrant. On the other hand, another set on the colour wheel is regarded to be cold. These hues have a chilling impact on humans, such

[1] `https://justpaint.org/defining-warm-and-cool-colors-its-all-relative/#:~:text=A%20dividing%20line%20splits%20the,and%20Magenta%20(Figure%202).`

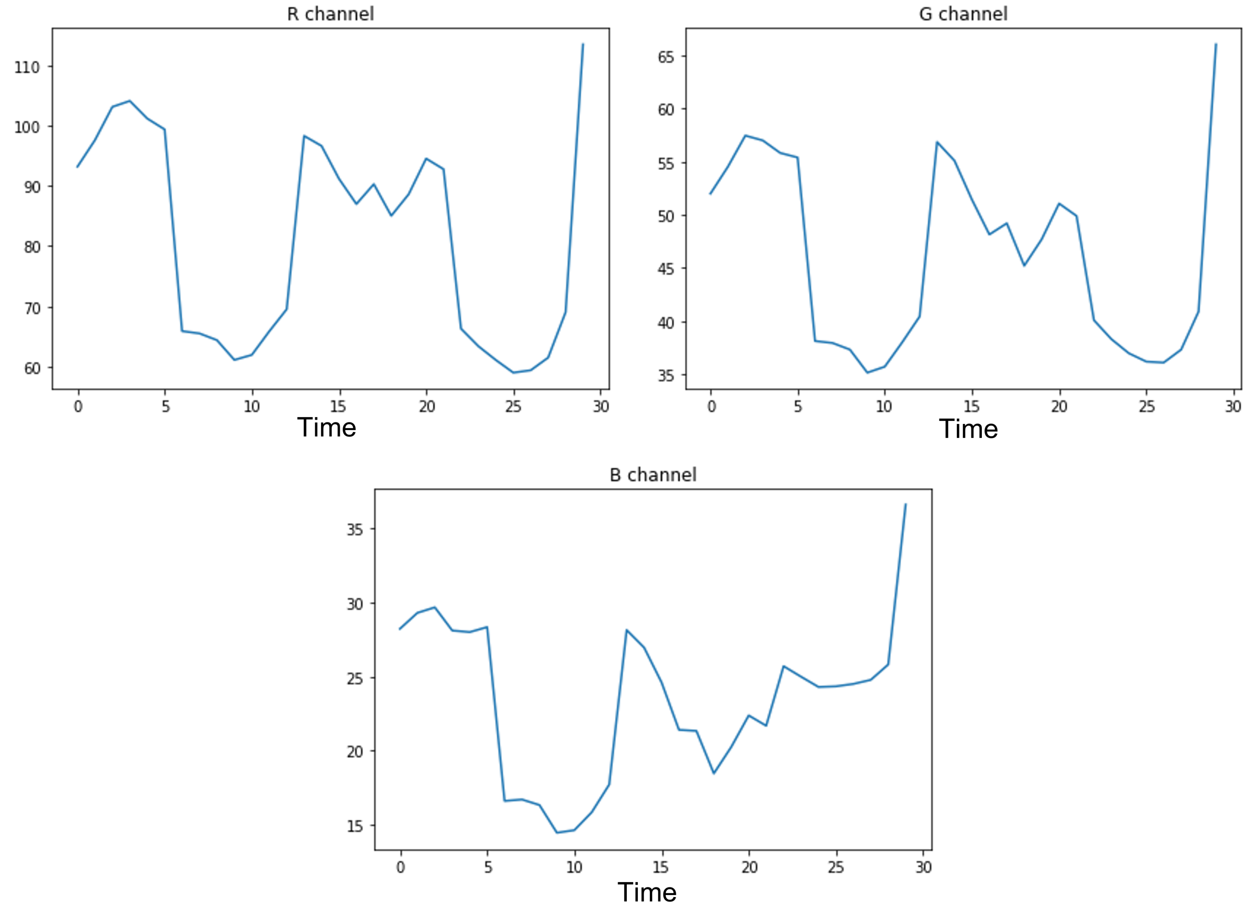


Fig. 5.2: An example clip for RGB changes. The values of red, green, and blue channels change over time as the movie clip plays.

as green, blue, purple, and shades of these hues [100, 101]. Neutral hues like white, black, and grey, for example, are neither warm nor cool. Depending on the undertones, a colour may lean towards one side of the spectrum. For example, a cream colour with a yellow undertone would look warm, whilst a grey with green undertones would appear cool. Brown is considered warm when it includes yellow or orange undertones. We choose red and green, two extremes on the colour wheel. Subsequently, warmth is calculated as the distance of the average colour from red, and coldness is calculated as the distance from green. Distance is calculated according to the colour wheel in Figure 5.1 using Equation 5.1. However, one thing worth pointing out is that our algorithm, which strictly computes based on the definition, can have limitations for some scenarios that neglect some context and background implications. For example, some people may associate a field of grass with warmth, while green is considered a cold colour in the background of colour theory.

$$\begin{aligned} Dis_R &= \sqrt{(Col_R - R_0)^2 + (Col_G - R_1)^2 + (Col_B - R_2)^2} \\ Dis_G &= \sqrt{(Col_R - G_0)^2 + (Col_G - G_1)^2 + (Col_B - G_2)^2} \\ Dis_B &= \sqrt{(Col_R - B_0)^2 + (Col_G - B_1)^2 + (Col_B - B_2)^2} \end{aligned} \tag{5.1}$$

After obtaining the absolute difference between the observed colour, and the anchor colour in each frame, we then compute the relative contribution of green and red with:

$$ratio = \frac{Dis_R}{Dis_G}. \tag{5.2}$$

With this value, we can compute the mapping between colour and temperature by setting the pure red (RGB=[255,0,0]) as the high temperature threshold, and the pure green (RGB=[0,255,0]) as the low temperature threshold.

$$Temperature = Temp_h + \frac{Temp_l - Temp_h}{Thred_l - Thred_h} \times ratio \tag{5.3}$$

where h means high, l means low, *Temp* is the temperature, and *Thred* is the threshold.

From this equation, we generated an example temperature prediction For a better understanding of the results of the above calculations, we produce an example using a movie clip, as shown in Figure 5.3. The threshold values we set are $Thred_l = 0, Thred_h = 1, Temp_l = 30,$ and $Temp_h = 55$. These values were determined empirically. As can be seen in Figure 5.3, the temperature goes up when there is an explosion event, and goes down when the explosion dissipates.

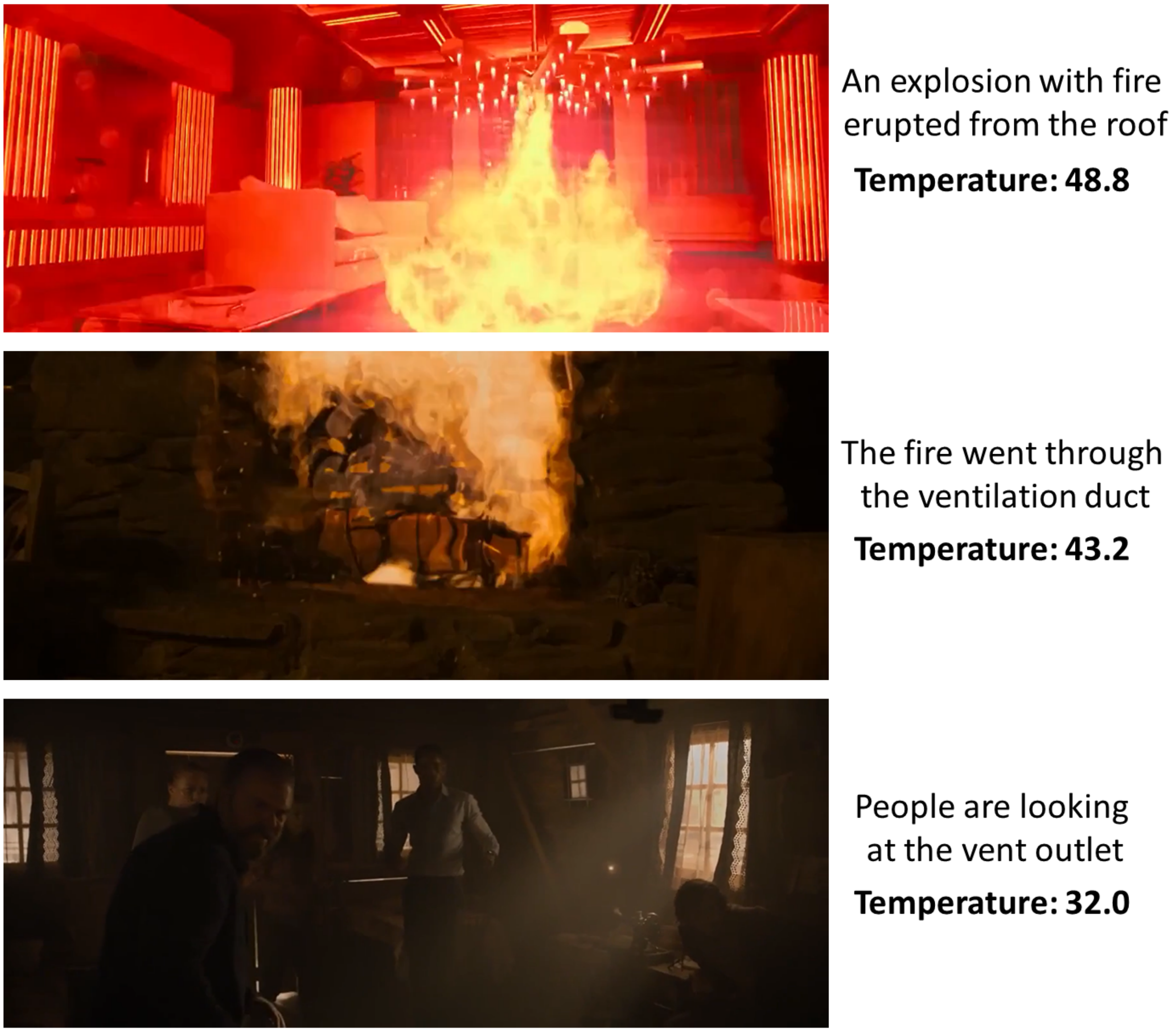

An explosion with fire erupted from the roof
Temperature: 48.8
The fire went through the ventilation duct
Temperature: 43.2
People are looking at the vent outlet
Temperature: 32.0

Fig. 5.3: Frame examples with changes in temperature.

5.3 Airflow Effect Generation

We illustrate our proposed algorithmic method to generate airflow effects in Figure 5.4, using the example of a flying butterfly. Broadly speaking, we use optical flow analysis to select the scenes with high intensity of motion (as shown in Figure 5.4 (a), (b), and (c)). For these frames, we then use depth information to produce airflow that suits the motion portrayed on screen.

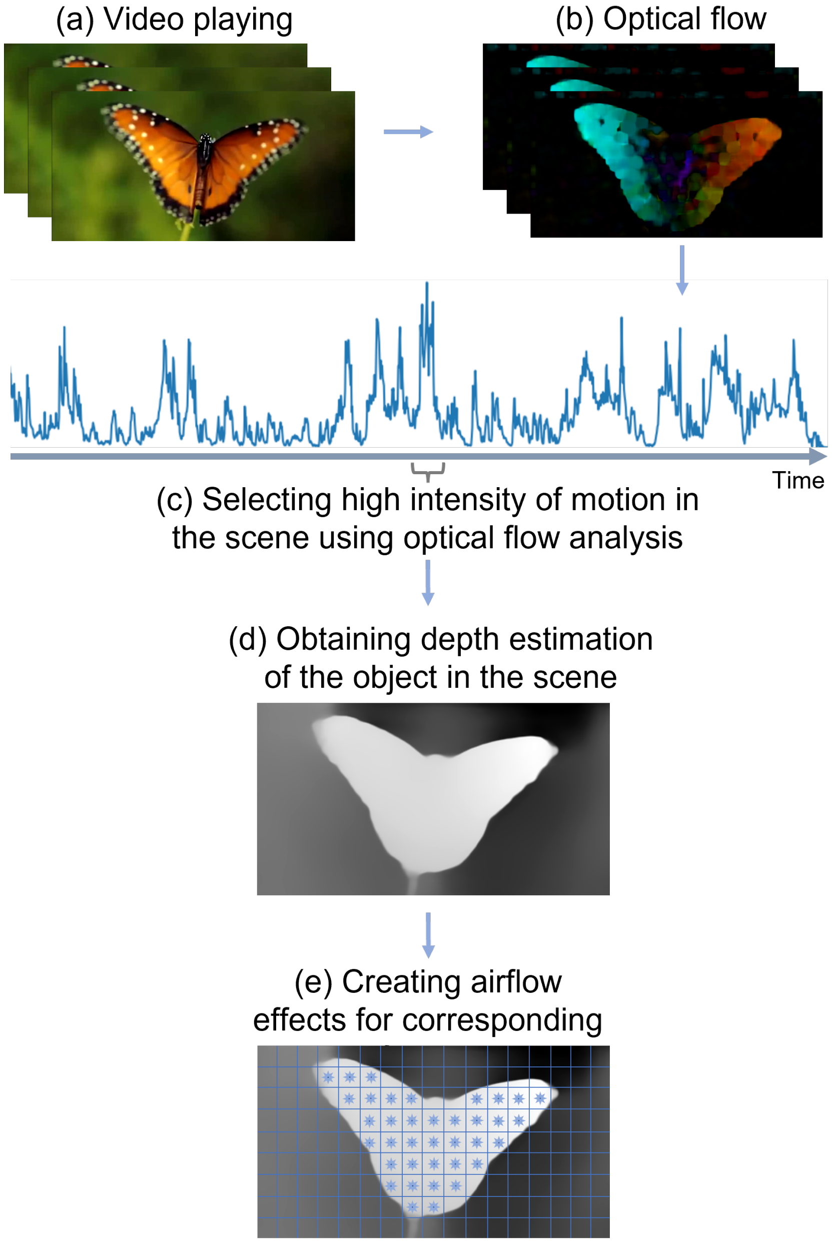

(a) Video playing
(b) Optical flow
Time
(c) Selecting high intensity of motion in the scene using optical flow analysis
(d) Obtaining depth estimation of the object in the scene
(e) Creating airflow effects for corresponding

Fig. 5.4: Frame examples with optical flow visualization.

5.3.1 Optical Flow estimation

Optical flow captures the movement of objects between successive frames by considering the relative motion of the object to the camera. Optical flow or optic flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and a scene. As shown in Figure 5.5, the optical flow formulation may be defined as follows:

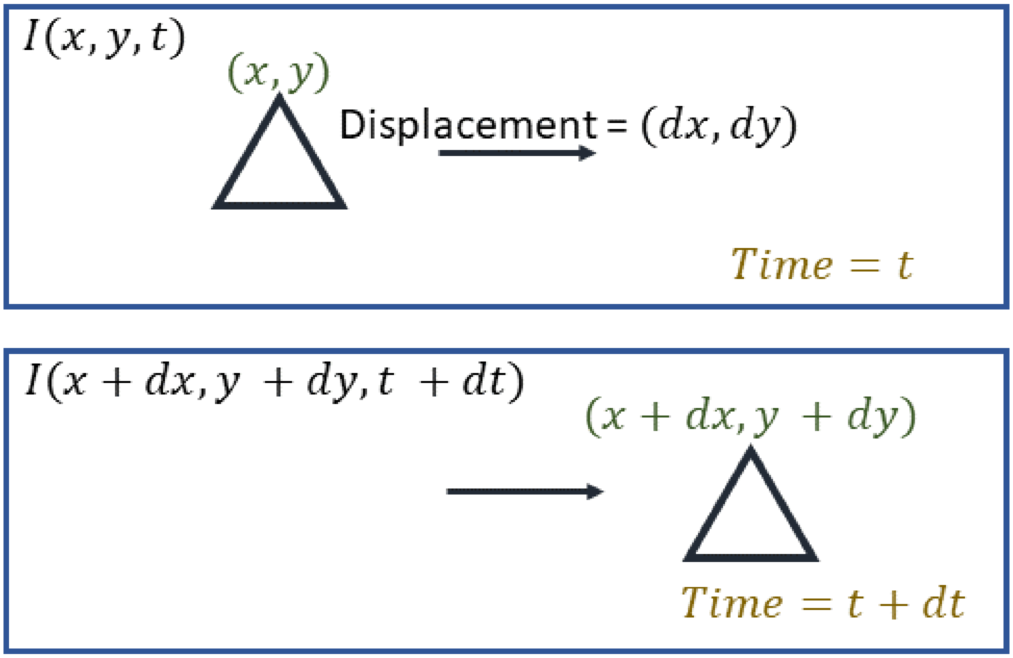


Fig. 5.5: The diagram of optical flow.

where we define the image intensity between consecutive frames as $I(\cdot)$, for an object at location (x,y), at time t. The pixels representing the object of interest are displaced by $(\mathrm{d}x,\mathrm{d}y)$ over time $\mathrm{d}t$; while the intensity of consecutive frames changes from $I(x,y,t)$ to $I(x+\mathrm{d}x,y+\mathrm{d}y,t+\mathrm{d}t)$. To simplify the model, we assume that pixel intensities associated with the object stay constant throughout subsequent frames:

$$I(x,y,t) = I(x+\delta x, y+\delta y, t+\delta t). \tag{5.4}$$

Then we eliminate common terms using a Taylor Series Approximation of the RHS, with the following formula:

$$\begin{aligned} I(x+\delta x, y+\delta y, t+\delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + \ldots \\ \Rightarrow \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = 0. \end{aligned} \tag{5.5}$$

By dividing by dt, we obtain Equation 5.6 from Equation 5.5:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0, \tag{5.6}$$

where $u = \mathrm{d}x/\mathrm{d}t$ and $v = \mathrm{d}y/\mathrm{d}t$. Therefore, we find that $\mathrm{d}I/x$, $\mathrm{d}I/\mathrm{d}y$, and $\mathrm{d}I/\mathrm{d}t$ are the image gradients along the horizontal axis, the vertical axis, and time, respectively.

Ultimately, the problem of optical flow comes down to solving $u(\mathrm{d}x/\mathrm{d}t)$ and $v(\mathrm{d}y/\mathrm{d}t)$ to determine movement over time. The calculation for optical flow can be regarded as the way to compute velocity in image space. However, we can not directly solve Equation 5.6, as there are two unknown variables (u and v) in only one equation. To handle this problem, we usually use an algorithmic approach such as Lucas-Kanade to compute the solution.

Techniques for computation of optical flow broadly fall into two groups: sparse optical flow and dense optical flow. Sparse optical flow provides flow vectors for a few points of interests, e.g., a few pixels displaying the edges or corners of an object, points that may have been extracted in advance using techniques such as Shi-Tomashi [102], Harris [103], and others. On the other hand, dense optical flow generates the flow vectors for the full frame, generating as many as one flow vector per pixel. As one might expect, dense optical flow is much more computationally expensive. We choose dense optical flow in this project as the length of the movies in our experiment can be handled feasibly. As shown in Figure 5.6, we visualize the optical flow with iridescence when the main characters move in the scene.

Airflow effects may not be appropriate for every frame in a movie. As such, we need a method of determining whether a frame has sufficient movement to justify airflow effects. To this end,

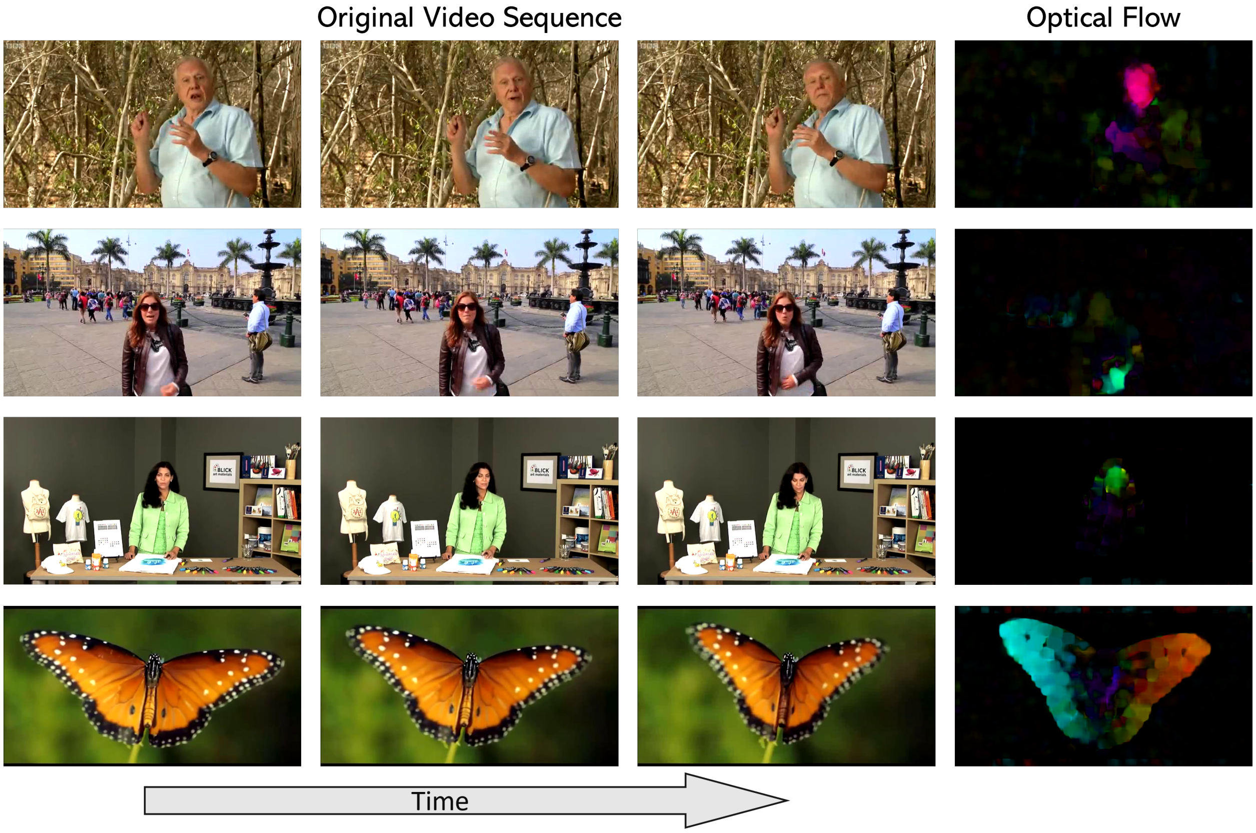


Fig. 5.6: Frame examples with optical flow visualization.

we calculate a summary value by summing all optical flow values. If this value is above a given threshold, we consider that this frame has enough motion to warrant an airflow effect, and the frame is passed on to later stages of the algorithm for depth-based processing. By thresholding this value, we can select the moments with the most intense movement; and it is in these moments that we generate airflow. This threshold is set manually and experimentally according to the movie content. In our experience, there is no one threshold value that is appropriate to all content. High intensity movements, such as crashes or explosions, require airflow that complements the motion on screen. Therefore, we employ depth estimation to obtain the distance of the objects from the viewer and generate the airflow effects to imitate the movements of the objects and characters in the movie. Such airflow effects can imitate the intensive movements in the movie and contribute towards providing users with an immersive experience.

5.3.2 Depth Estimation

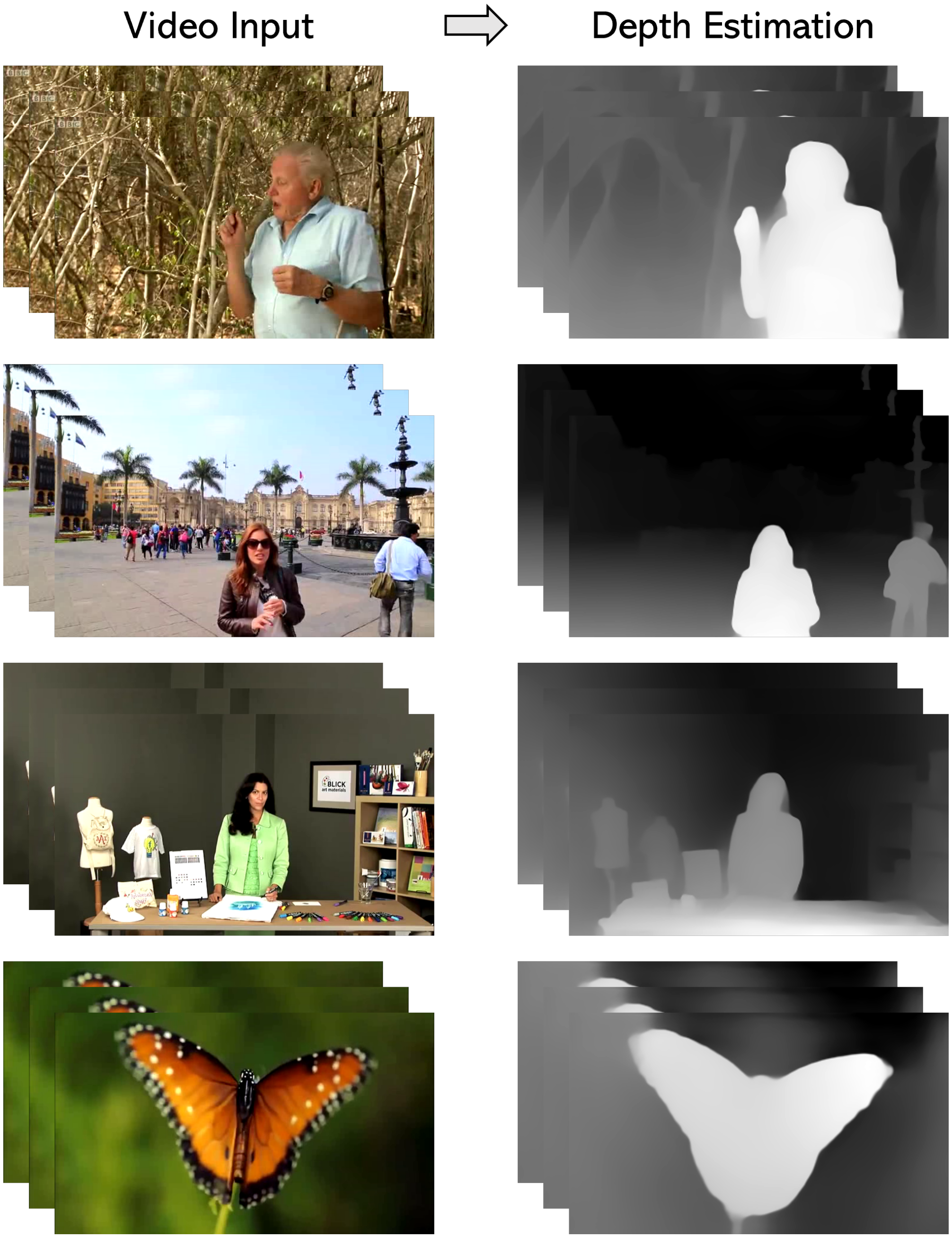
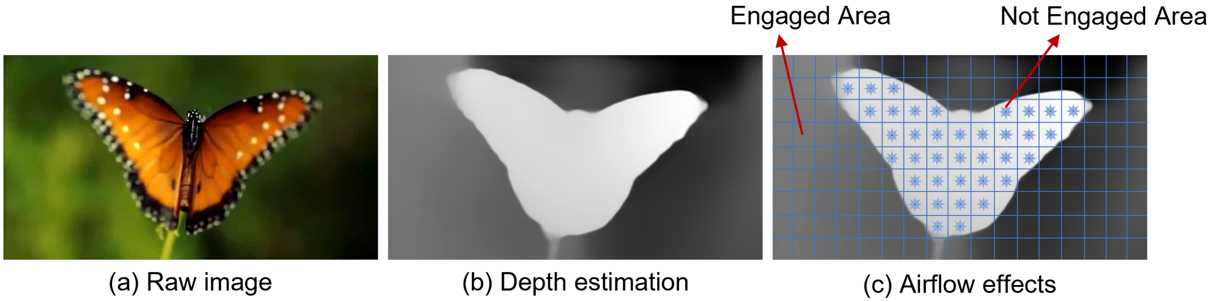


Fig. 5.7: Frame examples for depth estimation.

Depth information comprises distance between a given object and camera. Traditional approaches to depth estimation include Structure from Motion (SFM) (pipeline with calculation for correspondence estimation, pose estimation, triangulation, and bundle adjustment) [104], and Multi-view Stereo (MVS) [105]; however both techniques have limitations in terms of accuracy, reliability and robustness in obtaining correspondence. If either algorithm fails to calculate correspondence, the rest of the procedure is halted. Even when it succeeds, Multi-view Stereo [105] tends to generate a noisy depth map, with spurious holes within whole objects. Compared to traditional methods that match points across frames and geometric triangulation, algorithms that are based on machine learning are much more flexible in dealing with difficult circumstances [106, 107, 108]. However, it is possible that results produced by machine learning algorithms may lack geometric consistency at times. In recent years, hybrid algorithms have been introduced, which combine the best features of both methodologies [109, 110, 111, 112]. These algorithms often take supplemental inputs in the form of accurate per-frame camera postures, which are approximated using SFM. For the work presented here, we choose to use the state-of-art depth estimation work proposed by Kopf et al. [113]. The authors use a depth prior, learned from single image depth estimation using a convolutional neural network and improve the alignment of the depth maps. They also employed a geometry-aware depth filter to bring out fine depth details and eliminate residual jitter, rather than blurring them as a result of the prior stage's imperfect alignment. Our algorithm is capable of producing depth estimation for each individual frame in a video. We show several example images and their depth estimations in Figure 5.7.

5.3.3 Airflow effects Generation



(a) Raw image (b) Depth estimation (c) Airflow effects

Fig. 5.8: Frame examples for depth estimation.

Depth estimation is used in the next step of airflow effect rendering. It is not enough to know which frames merit an airflow effect; we also need to know where to generate airflow so that it reflects the motion on the screen. Importantly, we also do not want to generate airflow for motion that appears to be too far in the distance, as this would not reflect real-life experience. We can use the spatial information in the depth estimation to map airflow instructions to a spatial array of fans. In regions of the frame where high-motion objects are sufficiently close to the camera, we turn on the corresponding fan in the fan array, giving an illusion of spatially appropriate airflow. One thing to point out is that we only design the algorithm and visualize the expected effects instead of deploying them on a real hardware system. It may require a complex apparatus with multiple airflow ejection sites and sophisticated equipment in real scenarios. In addition, fans tend to have fairly wide cones of airflow distribution while we need relatively small and precise airflow shoot. The performance may be poorer than we expected with real hardware.

The implementation of such a system is beyond the scope of this thesis, which is primarily focused on translating multimedia streams into haptic control signals. However, we provide an example frame as shown in Figure 5.8(c), which give a schematic representation of how this system might work. As the workflow shown in Figure 5.4, after obtaining a depth estimation for the high intensity frames, an airflow system can generate an appropriate corresponding airflow effect. We visualize the virtual airflow effects that would be generated from a video, as a proof of concept. For an example, see Figure 5.8. Here, the algorithm correctly produces an airflow signal that matches the butterfly's location, which is cued by the flapping of its wings. In this example, one might install an array of 17 × 9 fans as represented by the grid on the right. As shown in Figure 5.8(c), we may allocate the working and non-working areas for the array of fans by superimposing the depth information and the grid of fans on top of each other. More specifically, the 17 × 9 fan array divides the scene into grids. We overlap the grid with depth estimation. The cells with small depth values indicate that there are objects that are relatively close to the observer. As a result, we turn on those fans that correspond to small depth values. As illustrated in Figure 5.8(c), only the fan within the outline of the butterfly is engaged, whereas the other fans corresponding to the background green tree remain inactive.

Chapter 6

Conclusion

In this thesis, we introduced an automatic, multimodal haptic rendering system for movie content. The program uses audiovisual features to infer a fundamental scene context, and generate haptic effects. Our algorithms are valuable as they can significantly reduce the workload of human labour for authoring haptic effects for movies, which is expensive and time-consuming with manual design.

In the first part of this work as presented in Chapter 3 and Chapter 4, we introduce an automatic vibrotactile effects generation algorithm that analyzes audio and visual cues from the movie. Its audio analysis is based on four psychoacoustic parameters: sharpness, booming, low-frequency energy and loudness. These acoustic features are used to calculate the time and amplitude of vibrotactile actuation. We choose these parameters as they usually contain significant amount of detailed information about the story and atmosphere in the movie. For instance, gunshots in a fight scene or booming noises in racing or flying sequences create an environment of high emotions. The system also uses a saliency map, which defines the area of predicted visual interest, to generate spatial effects of those vibrations produced according to audio analysis. As demonstrated in previous works, a saliency map can yield effective position and direction parameters for vibrotactile rendering, however it cannot automatically calculate an appropriate vibrotactile magnitude. Finally, we combine the results from both the audio and visual analyses elements to create vibrotactile effects.

To study its effectiveness, this system was implemented with an array of nine vibration motors embedded in a chair. The setup was designed to reflect a potential home theatre experience. We performed a user study with 16 participants to investigate the algorithm's efficacy in terms of user experience. Participants rated the device in terms of subjective immersion, preference, harmony, and discomfort. Participant ratings indicate that our multimodal rendering technique offers a significantly enhanced viewing experience.

Moreover, we further explored two other types of haptic modalities as presented in Chapter 5, using thermal and airflow effects. For thermal effects, we analyzed the warmth and coldness level of the visual scene by analyzing its average colour. For airflow effects, we isolated frames with high intensity movements using optical flow analysis, and then used depth analysis to estimate an appropriate strength for the airflow. As a result, people may immerse themselves in the scene's rapid motions.

In the scope of this thesis, we present algorithms to design automatic haptic effects based on audiovisual information analysis, which can dramatically reduce the workload of manual authoring. In the future, even for the actual cinema business, where film lengths can be extended and genres can be broadened, the algorithm can be still applied. For high-quality production, we could employ the outputs of our algorithm as a starting point and fine-tune the specific details with manual design. At the current stage as a research prototype, we only implement the hardware with nine eccentric rotating mass (ERM) vibrotactile actuators for evaluation. However, we believe that the algorithm can be deployed on more delicate hardware systems to further improve the user experience. Although the current setup is designed for home theatre environments, we believe that our system may also provide a valuable starting point for haptic effect designers in professional contexts.

References

[1] F. Danieau, A. Lécuyer, P. Guillotel, J. Fleureau, N. Mollet, and M. Christie, "Enhancing Audiovisual Experience with Haptic Feedback: A Survey on HAV," *IEEE Transactions on Haptics*, vol. 6, no. 2, pp. 193–205, 2012.

[2] "TV Watching and Online Streaming Surge during Lockdown," 2020, August 05. [Online]. Available: https://www.bbc.com/news/entertainment-arts-53637305

[3] J. Lee, B. Han, and S. Choi, "Motion Effects Synthesis for 4D Films," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 10, pp. 2300–2314, 2015.

[4] Zhiying Zhou, A. D. Cheok, Wei Liu, Xiangdong Chen, Farzam Farbiz, Xubo Yang, and M. Haller, "Multisensory Musical Entertainment Systems," *IEEE MultiMedia*, vol. 11, no. 3, pp. 88–101, 2004.

[5] D. Shin, "How Does Immersion Work in Augmented Reality Games? A User-Centric View of Immersion and Engagement," *Information, Communication & Society*, vol. 22, no. 9, pp. 1212–1229, 2019.

[6] L. Zou, I. Tal, A. Covaci, E. Ibarrola, G. Ghinea, and G.-M. Muntean, "Can Multisensorial Media Improve Learner Experience?" in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys'17. New York, NY, USA: Association for Computing Machinery, 2017, p. 315–320. [Online]. Available: https://doi.org/10.1145/3083187.3084014

[7] E. Kruijff, A. Marquardt, C. Trepkowski, J. Schild, and A. Hinkenjann, "Designed Emotions: Challenges and Potential Methodologies for Improving Multisensory Cues to Enhance User Engagement in Immersive Systems," *The Visual Computer*, vol. 33, no. 4, pp. 471–488, 2017.

[8] A. Covaci, L. Zou, I. Tal, G.-M. Muntean, and G. Ghinea, "Is Multimedia Multisensorial? - A Review of Mulsemedia Systems," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, Sep. 2018. [Online]. Available: https://doi.org/10.1145/3233774

[9] G. Ghinea, C. Timmerer, W. Lin, and S. R. Gulliver, “Mulsemedia: State of the Art, Perspectives, and Challenges,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 1s, Oct. 2014. [Online]. Available: https://doi.org/10.1145/2617994

[10] J. Dionisio, “Virtual Hell: A Trip Through the Flames,” *IEEE Computer Graphics and Applications*, vol. 17, no. 3, pp. 11–14, 1997.

[11] S. Palan, R. Wang, N. Naukam, L. Edward, and K. J. Kuchenbecker, “Tactile Gaming Vest (TGV),” 2010.

[12] M. Abdur Rahman, A. Alkhaldi, J. Cha, and A. El Saddik, “Adding Haptic Feature to YouTube,” in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM ’10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1643–1646. [Online]. Available: https://doi.org/10.1145/1873951.1874310

[13] S. u. Rehman, J. Sun, L. Liu, and H. Li, “Turn Your Mobile Into the Ball: Rendering Live Football Game Using Vibration,” *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 1022–1033, 2008.

[14] P. Lemmens, F. Crompvoets, D. Brokken, J. van den Eerenbeemd, and G. de Vries, “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*, 2009, pp. 7–12.

[15] M. Waltl, *Enriching Multimedia with Sensory Effects: Annotation and Simulation Tools for The Representation of Sensory Effects*. VDM Verlag, 2010.

[16] Y. Kim, J. Cha, J. Ryu, and I. Oakley, “A Tactile Glove Design and Authoring System for Immersive Multimedia,” *IEEE MultiMedia*, vol. 17, no. 3, pp. 34–45, 2010.

[17] B.-C. Lee, J. Lee, J. Cha, C. Seo, and J. Ryu, “Immersive Live Sports Experience with Vibrotactile Sensation,” in *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction*, ser. INTERACT’05. Berlin, Heidelberg: Springer-Verlag, 2005, p. 1042–1045. [Online]. Available: https://doi.org/10.1007/11555261_100

[18] D. Ablart, C. Velasco, and M. Obrist, “Integrating Mid-air Haptics into Movie Experiences,” in *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, 2017, pp. 77–84.

[19] T. Hoshi, M. Takahashi, T. Iwamoto, and H. Shinoda, “Noncontact Tactile Display Based on Radiation Pressure of Airborne Ultrasound,” *IEEE Transactions on Haptics*, vol. 3, no. 3, pp. 155–165, 2010.

[20] Y. Suzuki and M. Kobayashi, "Air Jet Driven Force Feedback in Virtual Reality," *IEEE Computer Graphics and Applications*, vol. 25, no. 1, pp. 44–47, 2005.

[21] J. Alexander, M. T. Marshall, and S. Subramanian, "Adding Haptic Feedback to Mobile TV," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 1975–1980. [Online]. Available: https://doi.org/10.1145/1979742.1979899

[22] D. Gaw, D. Morris, and K. Salisbury, "Haptically Annotated Movies: Reaching Out and Touching the Silver Screen," in *The 14th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS'06)*, 2006, pp. 287–288.

[23] Y. Kim, Sunyoung Park, Hyungon Kim, Hyerin Jeong, and J. Ryu, "Effects of Different Haptic Modalities on Students' Understanding of Physical Phenomena," in *2011 IEEE World Haptics Conference*, 2011, pp. 379–384.

[24] F. Danieau, J. Fleureau, P. Guillotel, N. Mollet, A. Lécuyer, and M. Christie, "HapSeat: Producing Motion Sensation with Multiple Force-Feedback Devices Embedded in a Seat," in *Proceedings of the 18th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 69–76. [Online]. Available: https://doi.org/10.1145/2407336.2407350

[25] S. O'Modhrain and I. Oakley, "Adding Interactivity: Active Touch in Broadcast Media," in *Proceedings of 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS '04).*, 2004, pp. 293–294.

[26] F. Danieau, J. Fleureau, A. Cabec, P. Kerbiriou, P. Guillotel, N. Mollet, M. Christie, and A. Lécuyer, "Framework for Enhancing Video Viewing Experience with Haptic Effects of Motion," in *2012 IEEE Haptics Symposium (HAPTICS)*, 2012, pp. 541–546.

[27] J. Cha, M. Eid, and A. El Saddik, "Touchable 3D Video System," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 5, no. 4, pp. 1–25, 2009.

[28] J. Seo, S. Mun, J. Lee, and S. Choi, "Substituting Motion Effects with Vibrotactile Effects for 4D Experiences," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–6. [Online]. Available: https://doi.org/10.1145/3173574.3174002

[29] A. Delazio, K. Nakagaki, R. L. Klatzky, S. E. Hudson, J. F. Lehman, and A. P. Sample, "Force Jacket: Pneumatically-actuated Jacket for Embodied Haptic Experiences," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[30] O. S. Schneider and K. E. MacLean, “Studying Design Process and Example Use with Macaron, A Web-based Vibrotactile Effect Editor,” in *2016 IEEE Haptics Symposium (HAPTICS)*. IEEE, 2016, pp. 52–58.

[31] J. Lee, J. Ryu, and S. Choi, “Vibrotactile Score: A Score Metaphor for Designing Vibrotactile Patterns,” in *World Haptics 2009 - Third Joint EuroHaptics conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, 2009, pp. 302–307.

[32] F. Danieau, J. Bernon, J. Fleureau, P. Guillotel, N. Mollet, M. Christie, and A. Lécuyer, “H-Studio: An Authoring Tool for Adding Haptic and Motion Effects to Audiovisual Content,” in *Proceedings of the Adjunct Publication of the 26th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST ’13 Adjunct. New York, NY, USA: Association for Computing Machinery, 2013, p. 83–84. [Online]. Available: https://doi.org/10.1145/2508468.2514721

[33] F. Danieau, P. Guillotel, O. Dumas, T. Lopez, B. Leroy, and N. Mollet, “HFX Studio: Haptic Editor for Full-Body Immersive Experiences,” in *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST ’18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3281505.3281518

[34] J. Ryu and S. Choi, “posVibEditor: Graphical Authoring Tool of Vibrotactile Patterns,” in *2008 IEEE International Workshop on Haptic Audio visual Environments and Games (HAVE’08)*, 2008, pp. 120–125.

[35] S.-K. Kim, “Authoring Multisensorial Content,” *Signal Processing: Image Communication*, vol. 28, no. 2, pp. 162–167, 2013.

[36] C. Chafe, “Tactile Audio Feedback,” in *Proceedings of the International Computer Music Conference*. International Computer Music Association, 1993, pp. 76–76.

[37] A. Chang and C. O’Sullivan, “Audio-Haptic Feedback in Mobile Phones,” in *CHI ’05 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA ’05. New York, NY, USA: Association for Computing Machinery, 2005, p. 1264–1267. [Online]. Available: https://doi.org/10.1145/1056808.1056892

[38] D. Chi, D. Cho, S. Oh, K. Jun, Y. You, H. Lee, and M. Sung, “Sound-Specific Vibration Interface using Digital Signal Processing,” in *2008 International Conference on Computer Science and Software Engineering*, vol. 4, 2008, pp. 114–117.

[39] M. Karam, F. A. Russo, and D. I. Fels, “Designing the Model Human Cochlea: An Ambient Crossmodal Audio-Tactile Display,” *IEEE Transactions on Haptics*, vol. 2, no. 3, pp. 160–169, 2009.

[40] S. Nanayakkara, E. Taylor, L. Wyse, and S. H. Ong, “An Enhanced Musical Experience for the Deaf: Design and Evaluation of a Music Display and a Haptic Chair,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’09. New York, NY, USA: Association for Computing Machinery, 2009, p. 337–346. [Online]. Available: https://doi.org/10.1145/1518701.1518756

[41] J. Lee and S. Choi, “Real-time Perception-level Translation from Audio Signals to Vibrotactile Effects,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2567–2576.

[42] M. Kim, S. Lee, and S. Choi, “Saliency-Driven Tactile Effect Authoring for Real-Time Visuotactile Feedback,” in *Proceedings of the 2012 International Conference on Haptics: Perception, Devices, Mobility, and Communication - Volume Part I*, ser. EuroHaptics’12. Berlin, Heidelberg: Springer-Verlag, 2012, p. 258–269. [Online]. Available: https://doi.org/10.1007/978-3-642-31401-8_24

[43] ——, “Saliency-Driven Real-Time Video-to-Tactile Translation,” *IEEE Transactions on Haptics*, vol. 7, no. 3, pp. 394–404, 2014.

[44] Y. Li, H. Zhao, H. Liu, S. Lu, and Y. Hou, “Research on Visual-tactile Cross-modality Based on Generative Adversarial Network,” *Cognitive Computation and Systems*, 2021.

[45] H. Liu, D. Guo, X. Zhang, W. Zhu, B. Fang, and F. Sun, “Toward Image-to-tactile Cross-modal Perception for Visually Impaired People,” *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 521–529, 2020.

[46] Y. Ujitoko, Y. Ban, and K. Hirota, “GAN-based Fine-tuning of Vibrotactile Signals to Render Material Surfaces,” *IEEE Access*, vol. 8, pp. 16 656–16 661, 2020.

[47] Y. Ujitoko and Y. Ban, “Vibrotactile Signal Generation from Texture Images or Attributes using Generative Adversarial Network,” in *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2018, pp. 25–36.

[48] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer Science & Business Media, 2013, vol. 22.

[49] E. Niebur, “Saliency Map,” *Scholarpedia*, vol. 2, no. 8, p. 2675, 2007.

[50] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, and Y. Avrithis, “Movie Summarization Based on Audiovisual Saliency Detection,” in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 2528–2531.

[51] N. Huang and M. Elhilali, "Auditory Salience using Natural Soundscapes," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2163–2176, 2017.

[52] I. Hwang, H. Lee, and S. Choi, "Real-Time Dual-Band Haptic Music Player for Mobile Devices," *IEEE Transactions on Haptics*, vol. 6, no. 3, pp. 340–351, 2013.

[53] R. T. Verrillo, "Vibrotactile Sensitivity and the Frequency Response of the Pacinian Corpuscle," *Psychonomic Science*, vol. 4, no. 1, pp. 135–136, 1966.

[54] "Iso 532-1:2017," Nov 2017. [Online]. Available: https://www.iso.org/standard/63077.html

[55] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[56] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep Learning for Generic Object Detection: A Survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020.

[57] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[58] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, 2020.

[59] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based Human Motion Recognition with Deep Learning: A Survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

[60] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 476–483.

[61] F. Sultana, A. Sufian, and P. Dutta, "Evolution of Image Segmentation Using Deep Convolutional Neural Network: A Survey," *Knowledge-Based Systems*, vol. 201, p. 106062, 2020.

[62] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[63] Q. Geng, Z. Zhou, and X. Cao, “Survey of Recent Progress in Semantic Image Segmentation with CNNs,” *Science China Information Sciences*, vol. 61, no. 5, p. 051101, 2018.

[64] T. Zhao and X. Wu, “Pyramid Feature Selective Network for Saliency detection,” *CoRR*, vol. abs/1903.00179, 2019. [Online]. Available: http://arxiv.org/abs/1903.00179

[65] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, “Learning to Detect Salient Objects with Image-level Supervision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 136–145.

[66] A. Israr and I. Poupyrev, “Tactile Brush: Drawing on Skin with a Tactile Grid Display,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 2019–2028.

[67] T. Amemiya, K. Hirota, and Y. Ikei, “Tactile Apparent Motion on The Torso Modulates Perceived Forward Self-motion Velocity,” *IEEE Transactions on Haptics*, vol. 9, no. 4, pp. 474–482, 2016.

[68] O. S. Schneider, A. Israr, and K. E. MacLean, “Tactile Animation by Direct Manipulation of Grid Displays,” in *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 21–30. [Online]. Available: https://doi.org/10.1145/2807442.2807470

[69] L.-P. Yuan, W. Zeng, S. Fu, Z. Zeng, H. Li, C.-W. Fu, and H. Qu, “Deep Colormap Extraction from Visualizations,” *arXiv preprint arXiv:2103.00741*, 2021.

[70] T. Zhang and C. Tomasi, “Fast, Robust, and Consistent Camera Motion Estimation,” in *Proceedings of 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1. IEEE, 1999, pp. 164–170.

[71] J. Tang, J. Folkesson, and P. Jensfelt, “Geometric Correspondence Network for Camera Motion Estimation,” *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 1010–1017, 2018.

[72] G. Mu, H. Cao, and Q. Jin, “Violent Scene Detection Using Convolutional Neural Networks and Deep Audio Features,” in *Chinese Conference on Pattern Recognition*. Springer, 2016, pp. 451–463.

[73] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su, “Violence Detection in Movies,” in *2011 Eighth International Conference Computer Graphics, Imaging and Visualization*. IEEE, 2011, pp. 119–124.

[74] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, "The Mediaeval 2011 Affect Task: Violent Scene Detection in Hollywood Movies," in *MediaEval 2011 Workshop*, 2011.

[75] J. Yu, W. Song, G. Zhou, and J.-j. Hou, "Violent Scene Detection Algorithm Based on Kernel Extreme Learning Machine and Three-dimensional Histograms of Gradient Orientation," *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 8497–8512, 2019.

[76] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context Encoding for Semantic Segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[77] H. Noh, S. Hong, and B. Han, "Learning Deconvolution Network for Semantic Segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[78] N. J. A. Pollet, E. Uzan, P. B. Ruivo, T. Abravanel, A. Talhan, Y. Yoo, and J. R. Cooperstock, "Multimodal Haptic Armrest for Immersive 4D Experiences," in *IEEE World Haptics Conference, Work in Progress*, 2021.

[79] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal Audiovisual Saliency Network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4766–4776.

[80] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-Supervised Salient Object Detection via Scribble Annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 546–12 555.

[81] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "RGB-D Salient Object Detection: A Survey," *Computational Visual Media*, pp. 1–33, 2021.

[82] G.-H. Yang, K.-U. Kyung, M. A. Srinivasan, and D.-S. Kwon, "Quantitative Tactile Display Device With Pin-Array Type Tactile Feedback and Thermal Feedback," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* IEEE, 2006, pp. 3917–3922.

[83] R. Wettach, C. Behrens, A. Danielsson, and T. Ness, "A Thermal Information Display for Mobile Applications," in *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services*, 2007, pp. 182–185.

[84] J. C. Stevens and B. G. Green, "Temperature–Touch Interaction: Weber's Phenomenon Revisited," *Sensory processes*, 1978.

[85] R. L. Klatzky, D. Pawluk, and A. Peer, “Haptic Perception of Material Properties and Implications for Applications,” *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2081–2092, 2013.

[86] S. Cardin, D. Thalmann, and F. Vexo, “Head Mounted Wind,” in *proceeding of the 20th annual conference on Computer Animation and Social Agents (CASA2007)*, no. CONF, 2007, pp. 101–108.

[87] F. Hülsmann, J. Fröhlich, N. Mattar, and I. Wachsmuth, “Wind and Warmth in Virtual Reality: Implementation and Evaluation,” in *Proceedings of the 2014 Virtual Reality International Conference*, 2014, pp. 1–8.

[88] T. Moon and G. J. Kim, “Design and Evaluation of a Wind Display for Virtual Reality,” in *Proceedings of the ACM symposium on Virtual reality software and technology*, 2004, pp. 122–128.

[89] T. Nakano, S. Saji, and Y. Yanagida, “Indicating Wind Direction Using a Fan-Based Wind Display,” in *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2012, pp. 97–102.

[90] J. Winzen, F. Albers, and C. Marggraf-Micheel, “The Influence of Coloured Light in the Aircraft Cabin on Passenger Thermal Comfort,” *Lighting Research & Technology*, vol. 46, no. 4, pp. 465–475, 2014.

[91] P. C. Berry, “Effect of Colored Illumination Upon Perceived Temperature,” *Journal of Applied Psychology*, vol. 45, no. 4, p. 248, 1961.

[92] F. H. Durgin, L. Evans, N. Dunphy, S. Klostermann, and K. Simmons, “Rubber Hands Feel the Touch of Light,” *Psychological Science*, vol. 18, no. 2, pp. 152–157, 2007.

[93] D. A. Zellner and M. A. Kautz, “Color Affects Perceived Odor Intensity,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 16, no. 2, p. 391, 1990.

[94] S. Bayarri, C. Calvo, E. Costell, and L. Durán, “Influence of Color on Perception of Sweetness and Fruit Flavor of Fruit Drinks,” *Food Science and Technology International*, vol. 7, no. 5, pp. 399–404, 2001.

[95] C. A. Bennett and P. Rey, “What’s So Hot About Red?” *Human Factors*, vol. 14, no. 2, pp. 149–154, 1972.

[96] B. Wright, “The Influence of Hue, Lightness, and Saturation on Apparent Warmth and Weight,” *The American Journal of Psychology*, vol. 75, no. 2, pp. 232–241, 1962.

[97] G. A. Michael and P. Rolhion, “Cool Colors: Color-Induced Nasal Thermal Sensations,” *Neuroscience Letters*, vol. 436, no. 2, pp. 141–144, 2008.

[98] N. Sriram and A. G. Greenwald, “The Brief Implicit Association Test,” *Experimental psychology*, vol. 56, no. 4, pp. 283–294, 2009.

[99] Z. Wang, Y. Nagai, E. Kim, N. Zou, J. Hou, X. Liu, and D. Zhu, “Lighting Style and Color Temperature to Emotion Response in Architecture Illumination of the Historic Buildings in Dalian,” in *2020 International Conference on Computer Engineering and Application (ICCEA)*. IEEE, 2020, pp. 613–617.

[100] S. Erol, “Coloring Support for Process Diagrams: A Review of Color Theory and a Prototypical Implementation,” *Vienna University of Economics and Business*, 2015.

[101] R. Zuhra, “Warm and Cold Colors and Their Application in Painting,” *Spanish Journal of Innovation and Integrity*, vol. 3, pp. 39–41, 2022.

[102] T. Lindeberg, “Feature Detection With Automatic Scale Selection,” *International journal of computer vision*, vol. 30, no. 2, pp. 79–116, 1998.

[103] C. Harris, M. Stephens *et al.*, “A Combined Corner and Edge Detector,” in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.

[104] S. Ullman, “The Interpretation of Structure From Motion,” *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 203, no. 1153, pp. 405–426, 1979.

[105] Y. Furukawa and C. Hernández, “Multi-View Stereo: A Tutorial,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 1-2, pp. 1–148, 2015.

[106] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman, “Learning the Depths of Moving People by Watching Frozen People,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4521–4530.

[107] Z. Li and N. Snavely, “Megadepth: Learning Single-View Depth Prediction From Internet Photos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2041–2050.

[108] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer,” *arXiv preprint arXiv:1907.01341*, 2019.

[109] C. Liu, J. Gu, K. Kim, S. G. Narasimhan, and J. Kautz, “Neural RGB->D Sensing: Depth and Uncertainty from a Video Camera,” *CoRR*, vol. abs/1901.02571, 2019. [Online]. Available: http://arxiv.org/abs/1901.02571

[110] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, and J. Kopf, "Consistent Video Depth Estimation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 71–1, 2020.

[111] Z. Teed and J. Deng, "Deepv2d: Video to Depth With Differentiable Structure From Motion," *arXiv preprint arXiv:1812.04605*, 2018.

[112] J. S. Yoon, K. Kim, O. Gallo, H. S. Park, and J. Kautz, "Novel View Synthesis of Dynamic Scenes With Globally Coherent Depths From a Monocular Camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5336–5345.

[113] J. Kopf, X. Rong, and J.-B. Huang, "Robust Consistent Video Depth Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1611–1621.