映像の意味論的文脈に基づく演出自動生成と分散制御を用いた家庭用 4D システム

松田彩成[1]　マルチェンコダニール[1]　久米蒼輝[1]　並川天夢[1]　中沢実[1,2]

概要：映像視聴の個人化が進み利便性が飛躍的に向上した一方で，風や振動等を伴う身体的な映像体験は，依然として映画館やテーマパーク等の専用施設が中心である．家庭環境においては，このような提示技術やコンテンツ生成の手間が障壁となり，未だ普及に至っていない．そこで本研究では，動画フレームのサンプリング列を入力として 4D 演出メタデータ（JSON）を生成し，家庭内デバイスを駆動するシステムを提案する（現状はバッチ解析であり，完全なリアルタイム同期は今後課題である）．手法として，マルチモーダル AI により映像の文脈を解析し，演出データを生成する．ハードウェアはRaspberry Pi と ESP モジュールを用いた分散構成とし，柔軟なデバイス制御を行う．これにより，既存の Web 動画等に対し，制作コストをかけずに高度な没入体験の付与を実現した．

1. はじめに

映像メディアの歴史は,19 世紀末の映画館における集団的な鑑賞体験に端を発する.当初,暗闇の中で巨大なスクリーンを見上げる行為は,人々にとって特別な「非日常」であった.その後,テレビやスマートフォンの普及により,映像は「いつでも,どこでも」消費可能な日常的情報となり,視聴の利便性は飛躍的に向上した.これに対し,映画館などの専用施設は,家庭用メディアとの差別化を図るべく,風や振動,水しぶきなどの物理的刺激を伴う「4D 映画」のような,身体的で没入度の高い体験価値の提供へと進化を遂げてきた.

しかし,この「体験の進化」は依然として専用施設の中に閉じているのが現状である.家庭環境において 4D 体験が普及しない主な要因は二点ある.第一に,既存システムが大掛かりであり,設置が困難というハードウェアの制約である.第二に,より本質的な課題として,物理効果のデータ制作に多大な人的コストを要する点が挙げられる.従来の手法では専門家による手作業が必須であるため,現代人が日常的に接する YouTube 等の膨大な Web 動画コンテンツに対応することは事実上不可能であった.

そこで本研究では,マルチモーダル AI による自動生成と分散型ハードウェア制御を統合した,汎用的な家庭用 4D システムを提案する.図 1 に本システムの利用風景を示す.

図 1：提案システムの利用風景

本システムは,AIが映像の意味論的文脈を解析して演出を生成し,無線分散ネットワークがデバイス群を同期駆動するものである.図 1 に示すように,ユーザーは既存の視聴環境にデバイスを追加するだけで,風や振動といった物理効果を体感できる.このアプローチにより,コンテンツ生成のボトルネックを低減し,多様な映像コンテンツに対する身体的体験の付与を支援する.一方で,現状はフレームサンプリングとクラウド推論に依存するため,タイミング同期や静寂区間の扱いには課題が残る.

2. 関連研究

2.1 映画館および家庭における 4D 視聴体験

商業施設においては,韓国 CJ 4DPLEX 社の「4DX」[1]や,米国 MediaMation 社の「MX4D」[2]などが普及している.これらのシステムは,座席の可動,風,水,香りなどの多様な物理刺激を映像と同期させることで,高い没入感を提供している.

一方,家庭環境向けの研究としては,振動素子を埋め込んだベストや椅子[3][4],あるいはディスプレイ周辺にファンを配置する手法などが多数提案されている.しかし,これら既存研究の多くはハードウェアの提示機構に主眼を置いており,それらを駆動させるためのメタデータをいかに低コストで用意するかという,コンテンツ制作の課題については十分に解決されていない.

2.2 物理効果の自動生成手法

コンテンツ制作のコストを削減するため,映像や音声信号から物理効果を自動生成する試みも行われている.初期の研究では,音声信号の振幅（音量）を振動強度に変換する手法[5]や,映像のオプティカルフロー（動きベクトル）を解析して風の強さを決定する手法[6]などが提案された.これらはリアルタイム処理が可能である反面,「意味論的な文脈」を理解できないという課題がある.

1 金沢工業大学
2 Email ： nakazawa@infor.kanazawa-it.ac.jp

例えば,音声ベースの手法では「静かなシーンでの強風」を表現できず,オプティカルフローベースの手法では「カメラパンによる動き」と「被写体の疾走」を区別することが困難である.すなわち,従来の信号処理的なアプローチでは,製作者が意図する演出と自動生成結果の間に大きな乖離が生じるという問題があった.

2.3 マルチモーダル AI の活用と本研究の独自性

近年,画像と言語を統合的に扱うマルチモーダル AI (Vision-Language Models)が急速に発展し,映像要約やキャプション生成において高い性能を示している.しかし,これを物理的なアクチュエータの制御に応用した事例は未だ少ない.

本研究は,従来の「信号レベルの解析(音量や画素の動き)」ではなく,LLM を用いた「意味レベルの解析(爆発,水中,空中などの状況理解)」に基づいて物理効果を生成する点に新規性がある.これにより,従来手法では実現困難であった「文脈に即した演出」を,構造化制御データ(JSON)として自動生成するプロトタイプを示す.

3. 提案手法

本章では,開発した「4DX@HOME」の具体的なシステム構成について述べる.本システムは,物理的な効果を提示するハードウェア,それらを統括する分散制御ネットワーク,および映像の意味論的解析を担う AI エンジンの 3 要素により構成される.

3.1 ハードウェア構成と筐体設計

本システムでは,視覚・聴覚以外の体験を提供するために,役割の異なる 2種類の デバイス「 EffectStation 」と 「ActionDrive」を開発した.全ての筐体は 3D CAD によって設計され,FDM方式の 3Dプリンタを用いて製作されている.

図 2 : EffectStation の実機外観

図 2 に示す「EffectStation」は環境効果を提示する複合デバイスであり,卓上への設置を想定して設計されている.内部には,サーボモータによる物理トリガー式の 水噴射機構,PWM 制御された DC ファンによる風生成機構,および高輝度 LED による光,RGB 対応 LED テープによる色提示機能を備える.

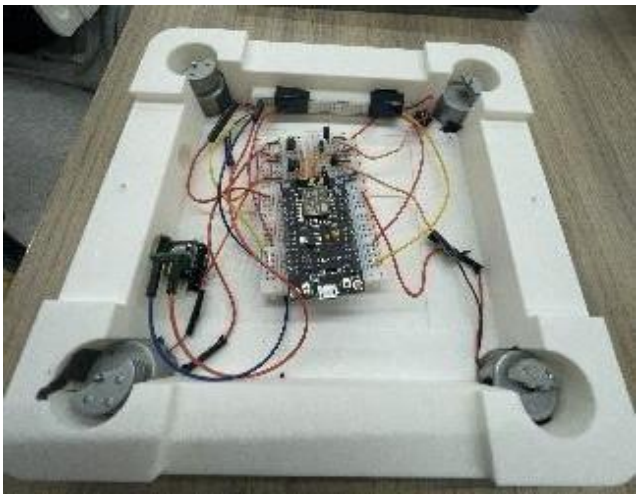


図 3 : ActionDrive 実機外観(上)および 3D プリントされたケースと振動モジュール(下)

一方,図 3 に示す「ActionDrive」は,ソファや椅子に設置して使用する振動提示クッションである.本デバイスの特徴は,振動機構を単に埋め込むのではなく,コンポーネント単位で最適化・モジュール化している点にある.

図 3(下)に示すように,振動源となる偏心モータ(ERM)と,無線制御を担う ESP マイコンは,専用ケース内に一体化して格納されている.この独立した振動モジュールを座面と背面に計 8 基配置することで,映像内のエンジンの微細な鼓動から激しい衝突衝撃までを,部位ごとに使い分けて表現することを可能にした.

3.2 分散制御ネットワークと組み込み実装

複数のデバイスを配線の制約なく配置するため,Raspberry Pi と ESP マイコンを用いた無線分散制御アーキテクチャを採用した.

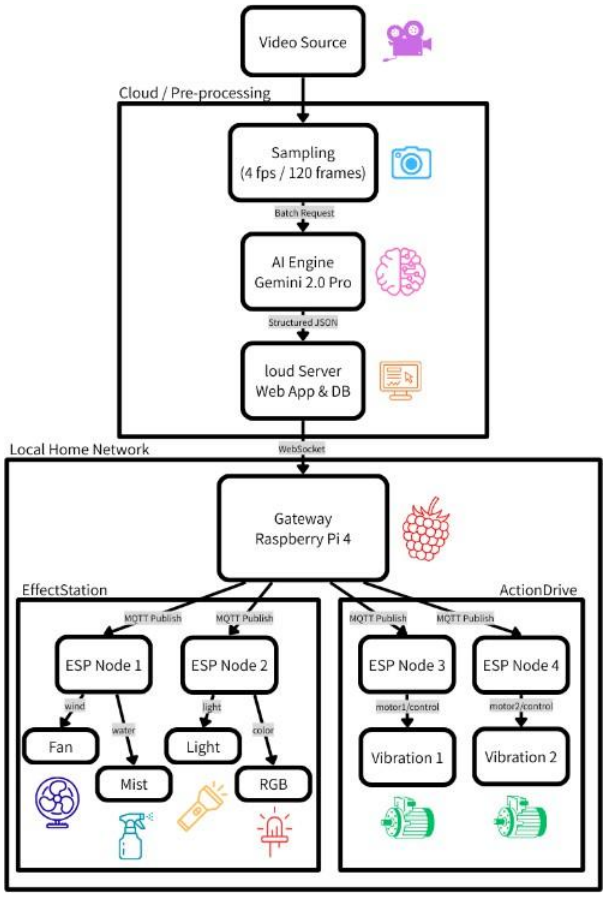


図 4：システム全体のデータフロー図

図 4 にシステムの全体構成を示す. 本システムでは, AI によって解析された内容に基づいて各デバイスを制御する. そ☆中間に位置するゲートウェイには Raspberry Pi[8]を使用し, クラウド上の Web アプリケーションとは WebSocket を用いて接続する. これにより, 再生中の映像のタイムスタンプやシーク操作などのステータス情報がストリーミングされ, 同期制御が行われる.

ゲートウェイから各フィードバックデバイスへの指令には, 軽量かつ 1 対多の通信に適したMQTT プロトコル[10]を採用した. Raspberry Pi 内部に構築された MQTT Broker に対し, 風や振動などの機能ごとに定義されたトピック（例： 4dx/control/wind）へ制御コマンドを Publish する. 各デバイスの制御ノードには Wi-Fi 機能を持つ ESP-12E (ESP8266)[9] を使用しており, 自身に必要なトピックのみを Subscribe することで, 無線環境下での制御を実現している.

3.3 マルチモーダル AI による映像解析

映像解析には, Google Gemini API (Gemini 2.0 Pro)[7]を使用した. 映像データをトークンコストや通信帯域を考慮して効率的に処理するため, フレームサンプリングを用いたバッチ処理手法を採用している.

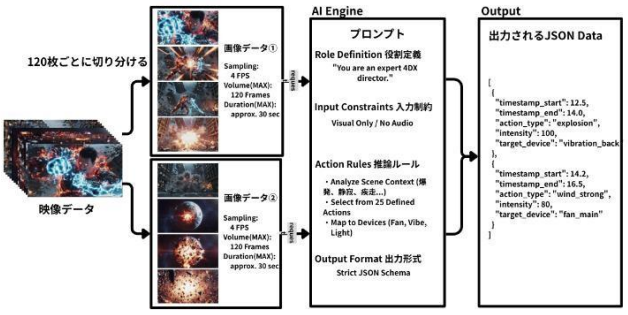


図 5：プロンプト構造と解析プロセス[1]

具体的には, 対象動画を 4 FPS（0.25 秒に 1 枚）の頻度で静止画として切り出し, 連続する 120 フレーム（約 30 秒分の映像）を 1 つのリクエストとしてまとめて AI に入力する. 単一画像ではなく時系列順に並んだフレーム群を提示することで, 前後の文脈（Temporal Context）を含めた解析を行っている. なお, 本実装においては映像フレームのみを入力とし, 音声データの解析は行っていない.

AI の出力制御においては, 図 5 に示すプロンプトエンジニアリングを導入した. AI に対し「4D 効果のパラメータ設定を行う」という役割（Role）を与え, 出力形式を独自の JSON スキーマに固定している. これにより, 映像内の事象を定義済みのアクション（爆発, 疾走, 水辺など）に分類し, 構造化データとして生成させている.


```
{
 "events": [
  {
   "t": 12.00,
   "action": "caption",
   "text": "静かな敵基地の全景。緊張感が漂っている。"
  },
  {
   "t": 12.25,
   "action": "caption",
   "text": "画面中央の火薬庫から、突然白い閃光が走り始める。"
  },
  {
   "t": 12.25,
   "action": "start",
   "effect": "flash",
   "mode": "fast_blink"
  },
  {
   "t": 12.50,
   "action": "caption",
   "text": "爆発が拡大し、激しい炎と衝撃波が画面全体に広がる。"
  },
  {
   "t": 12.50,
   "action": "start",
   "effect": "color",
   "mode": "red"
  },
  {
   "t": 12.50,
   "action": "start",
   "effect": "vibration",
   "mode": "up_down_strong"
  },
  {
   "t": 12.75,
   "action": "caption",
   "text": "爆風により瓦礫が手前側へ吹き飛んでくる。"
  },
  {
   "t": 12.75,
   "action": "start",
   "effect": "wind",
   "mode": "burst"
  }
 ]
}
```


図６：生成された JSON 形式のデータ

1 使用した画像は生成 AI を用いて出力したものである

生成されるデータは図 6 のような JSON 形式であり,0.25 秒ごとのサンプリングフレーム全てに対して caption（状況説明）を生成している点が特徴である.この高密度な文脈記述に基づいて,タイムスタンプをキーとした効果種別（effect）および強度や動作パターン（mode）の開始・停止が決定される.このように,断続的な信号検出ではなく連続的な意味理解を行うことで,複雑なマルチモーダル演出を標準化されたフォーマットで記述し,後段の制御システムでの定型処理を可能にした.

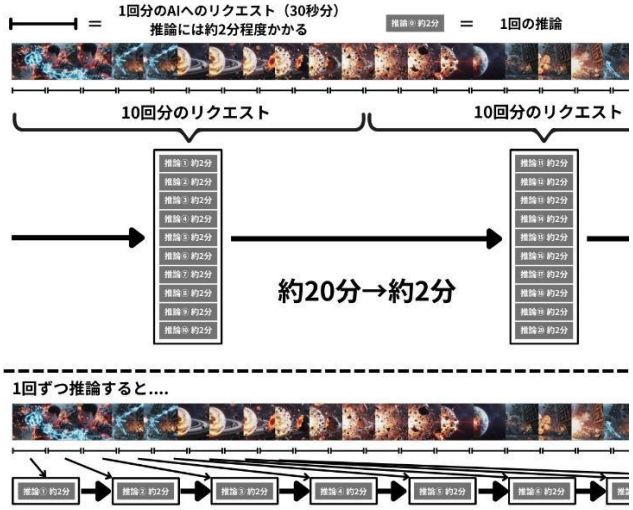


図 7：非同期処理による解析を表した図[1]

AI による映像解析処理は計算コストが高く,Gemini 2.0 Proを用いた場合でも,30 秒の映像セグメントに対し約 2 分の推論時間を要する.そのため,単純な順次処理（Sequential Processing）では映像の再生速度に解析が追いつかず,視聴体験を損なう要因となる.

そこで本システムでは,Python の AsyncIO[11]を用いた非同期並列処理機構を実装した.図 7 にそのタイムライン比較を示す.本手法では,API のレートリミットを考慮しつつ最大 10 の並列リクエストを同時に実行することで,実質的な処理時間を約 1/10 に短縮することに成功した.これにより,理論上「30 秒の映像を約 12 秒で解析」することが可能となり,システムの実用性を著しく向上させている.

このような並列推論をローカル環境の LLM で実装する場合,高価な GPU リソースを複数台用意する必要があり,一般家庭への導入は困難である.対して,クラウド LLM（API）を採用することで,安価かつ容易に大規模な並列リソースを活用できる点は,本提案手法のハードウェア制約における大きな利点である.

なお,レイテンシを極限まで排除した完全なリアルタイム推論の実現については,本成果を基盤とした今後の展望とする.

4. 実験と評価

4.1 検証条件と倫理的配慮

本システムの入力には MP4 形式等の動画ファイルを用いる.なお,近年ではストリーミング配信の無断保存は多くのプラットフォームで規約により制限されている.そのため本検証は,ユーザーが正規の手順で取得した映像,商用利用および改変が許可されているクリエイティブ・コモンズ（CC）ライセンスの映像,および画像生成 AI を用いて独自に生成した映像素材を対象として行った.

4.2 システム稼働実験

構築したシステムを用いて, 実際の家庭環境を模したセットアップを行い, 動作検証を実施した.

図 8： Web アプリケーション上の操作画面[2]

図 8 は,ユーザーが操作する Web アプリケーションのインターフェースである.ユーザーはこの画面から動画を選択し,リクエストを送るだけで,生成されたメタデータに基づく 4D 視聴を開始できる.

映像が再生されると,画面内のアクションシーン（爆発や強風など）に合わせて,卓上の EffectStation から風が噴出し,同時に ActionDrive が振動することで,視聴者に物理的な衝撃を伝達していることが確認された.

4.3 実験結果と考察

4.3.1 定量的評価

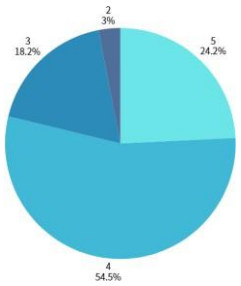
システムの有効性を定量的に評価するため,被験者による視聴実験およびアンケート調査を行った.実験協力者として,20 代から 50 代の男女 33 名を選定した.被験者はシステム稼働下で約 2 分間の映像コンテンツを視聴し,その後 5段階のリッカート尺度（1:否定的 ～ 5:肯定的）を用いたアンケートに回答した.質問項目は以下の通りである.

Q1. 臨場感: 映像の中に入り込んだような感覚（実在感）があったか
Q2. 迫力: 通常の視聴と比較して迫力が増したか
Q3. 整合性: 風や振動の演出は映像の内容（文脈）と合

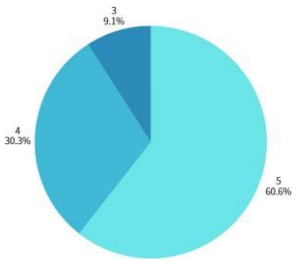
1 使用した画像は生成 AI を用いて出力したものである

2 使用した画像は生成 AI を用いて出力したものである

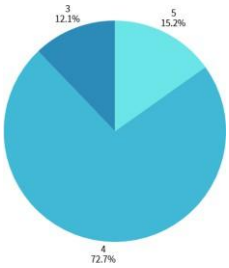
っていたか
Q4. 同期性: 映像と効果のズレ（遅延）は気にならなかったか
Q5. 総合評価: 体験は楽しかったか
.



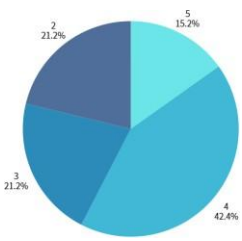
(a) Q1 臨場感（実在感）に対する回答分布



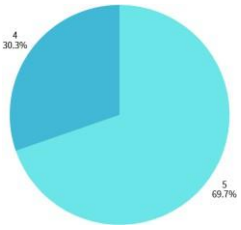
(b) Q2 迫力に対する回答分布



(c) Q3 演出の整合性に対する回答分布



(d) Q4 同期性に対する回答分布



(e) Q5 総合評価に対する回答分布
図 9: 各質問の評価の回答分布

表 1: 各質問の評価に関する平均値および標準偏差

指標	平均値	標準偏差
臨場感	4	0.75
迫力	4.52	0.67
演出の整合性	4.03	0.53
遅延	3.52	1
総合評価	4.7	0.47

図 9 に各項目の回答分布を示すまた, 全回答の平均値および標準偏差は表 1 の通りである.

集計の結果,「Q5. 総合評価」は平均 4.70 と極めて高い評価を得た. 特に 33 名中 23 名が最高評価の「5」を選択しており, 本システムが提供する体験が視聴者にとって非常に魅力的であることが示唆された. また,「Q2. 迫力」においても平均 4.52 を記録し, 通常の視聴環境と比較して顕著な体験の拡張が確認された.「Q1. 臨場感」(平均 4.00) および「Q3. 整合性」(平均 4.03) についても概ね肯定的な評価が得られた.

一方で,「Q4. 同期性」の平均値は 3.52 と他の項目に比べて低く, 標準偏差も 1.00 とばらつきが大きかった. これは, 画像分析と時間同期の精度の不足により, 一部のシーンで映像と物理効果のタイミングにズレが生じたことが原因と考えられる.

4.3.2 定性的評価

自由記述によるフィードバックの分析結果を述べる. 肯定的な意見として最も多かったのは「振動」に関するものであり, 全回答の約 3 割にあたる 11 件で言及された.「振動のインパクトが凄かった」「背面の振動により没入感が増した」といった声が多数寄せられ,「ActionDrive」による身体的フィードバックが体験の核となっていることが確認できる.

代表的な失敗例として, (F1) ガラスが粉砕される瞬間に水が噴射されるなどの効果種別の取り違え, (F2) アクションシーンとアクションシーンの間の静寂を認識できず効果が持続/再発火する状態遷移の誤り, (F3) 衝撃の瞬間がどのフレームかの認識が甘く早かったり遅かったりする時間境界の誤差が確認された. これらは映像と物理効果のズレ

として知覚されやすく,前述の同期性スコア（Q4）の低さに寄与したと考えられる.また,機能要望として「光（ライティング）の強化」（5 件）や「風の演出強化」(4 件),さらには「熱」や「匂い」の追加を求める声もあり,より多感覚な演出への期待が高いことが明らかになった.

以上の結果より,本システムは「迫力」や「楽しさ」の面では十分に実用水準に達しているものの,没入感をさらに高めるためには,推論・制御の低遅延化による「同期性」の向上が最重要課題であると結論付けられる.

結言

4.4 まとめ

本研究では,マルチモーダル AI の文脈理解能力と,安価な分散ハードウェアを組み合わせることで,映像フレーム列から 4D 演出メタデータ（JSON）を生成して提示する家庭用システム「4DX@HOME」を提案した.評価実験の結果,本システムは特別なコンテンツ制作スキルを持たないユーザーであっても,既存の膨大な映像リソースを「体験」として再発見できる可能性を示した.特に振動演出は高い評価を得た一方で,映像と演出の同期性や,視覚以外の環境演出に関しては改善の余地が確認された.

一方で,フレームサンプリングとバッチ推論に依存するため,効果種別の取り違え,静寂区間の扱い,および衝撃タイミング推定に起因する同期誤差が生じうる.したがって,推論・制御の低遅延化と状態遷移設計の改善が,没入感向上のための主要課題である.

4.5 今後の展望

実験で得られた知見に基づき,今後の展望として以下の 3 点に取り組む予定である.

第一に,推論のリアルタイム化とマルチモーダル化である.実験におけるアンケート結果（Q4: 同期性）では,通信や処理遅延に起因する評価のばらつきが見られた.そこで,現在はバッチ処理となっている推論フローを最適化し,低遅延化を図ることでライブ配信等への即応を目指す.また,現在の視覚情報に加え,音声信号を解析対象に統合することで,画面外の環境音や BGM の雰囲気も考慮した,より精度の高い演出生成を可能にし,演出の整合性を向上させる.

第二に,提示デバイスのダイナミクス向上である.定性評価において最も支持を集めた振動演出をさらに強化するため,現状の振動クッション（ActionDrive）を拡張する.具体的には,映像の動きに合わせて椅子全体が物理的に揺動する機構の開発を行うことで,全身での没入感を高める.

第三に,家庭環境ならではの演出手法の開拓である.実験では「光演出の強化」や「部屋全体の雰囲気作り」を求める声が多く寄せられた.これに応えるため,プロジェクタ等を用いて壁面に映像内の環境色や模様を投影し,部屋全体をコンテンツの雰囲気で満たすような空間演出[12]の実装を検討している.

これにより,ディスプレイの枠を超えて居住空間そのものを没入環境へと変容させる,家庭ならではの新しい視聴体験の創出を目指す.

謝辞　実験に参加し,アンケート調査にご協力頂いた皆様に，謹んで感謝の意を表する.

参考文献

[1] CJ 4DPLEX: 4DX, https://www.cj4dx.com/（閲覧日 2024/12/14）.
[2] MediaMation: MX4D, https://mediamation.com/（閲覧日 2024/12/14）.
[3] 南澤 孝太, 筧 康明, 仲谷 正史, 三原 聡一郎, 舘 暲：TECHTILE toolkit: 触感コンテンツ作成・共有のためのプロトタイピングツール，日本バーチャルリアリティ学会論文誌，Vol.17，No.3，pp.267-276（2012）.
[4] 渡邊 恵太，安村 通晃：振動呈示機能付き椅子による映像コンテンツの臨場感向上，情報処理学会インタラクション 2007 論文集，pp.147-148（2007）.
[5] 秋田 純一，伊藤 清宏，末長 隆司：音楽信号の振幅情報を用いた振動提示による音楽鑑賞支援システム，情報処理学会論文誌，Vol.48，No.12，pp.3838-3846（2007）.
[6] S. Kim, Y. Jo, K. Yoon and I. S. Jang: Automatic generation of sensory effects using content analysis for 4D broadcasting, ETRI Journal, Vol.36, No.5, pp.819-829 (2014).
[7] Google DeepMind: Gemini - Google DeepMind, https://deepmind.google/technologies/gemini/（閲覧日 2024/12/21）.
[8] Raspberry Pi Foundation: Raspberry Pi 4 Model B, https://www.raspberrypi.com/products/raspberry-pi-4-model-b/（閲覧日 2024/12/20）.
[9] Espressif Systems: ESP8266EX Datasheet, https://www.espressif.com/sites/default/files/documentation/0a- esp8266ex_datasheet_en.pdf（閲覧日 2024/12/20）.
[10] OASIS: MQTT Version 5.0, https://docs.oasis-open.org/mqtt/mqtt/v5.0/mqtt-v5.0.html（閲覧日 2024/12/20）.
[11] Python Software Foundation: asyncio – Asynchronous I/O, https://docs.python.org/3/library/asyncio.html（閲覧日 2024/12/21）.
[12] B. Jones, H. Benko, E. Ofek and A. D. Wilson: IllumiRoom: peripheral projected illusions for interactive experiences, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13), pp.869-878 (2013).